Analysis of the Formality of Text and its Impact on Pharmacovigilance Systems

Néstor ÁLVARO GRADILLAS

Doctor of Philosophy

Department of Informatics
School of Multidisciplinary Sciences
SOKENDAI (The Graduate University for Advanced Studies)

SOKENDAI
(The Graduate University for Advanced Studies)

Doctoral Thesis

Analysis of the Formality of Text and its Impact on Pharmacovigilance Systems

Author:
Néstor Álvaro Gradillas

Supervisor:
Dr. Yusuke Miyao
Dr. Nigel H. Collier

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy

in the

Department of Informatics
School of Multidisciplinary Sciences

September 2016

Declaration of Authorship

I, Néstor Álvaro Gradillas, declare that this thesis titled, 'Analysis of the Formality of Text and its Impact on Pharmacovigilance Systems' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

"There are no small problems. Problems that appear small are large problems that are not understood."

Santiago Ramón y Cajal

SOKENDAI
(The Graduate University for Advanced Studies)

Abstract

Digital Content and Media Sciences Research Division
School of Multidisciplinary Sciences

Doctor of Philosophy

Analysis of the Formality of Text and its Impact on Pharmacovigilance Systems

by Néstor Álvaro Gradillas

This thesis aims to answer the question of whether drug use reports obtained from formal and informal sources have noticeable differences in their formality, and also assess whether these differences in the formality can provide gains to pharmacovigilance systems. Working towards these goals we made three clear contributions that are the development of new resources, i.e. corpora, to perform our studies, a linguistic analysis of the differences between formal and informal pharmacovigilance reports, and an assessment on the impact of the register-related features in pharmacovigilance systems.

The first contribution is motivated after finding that there is no repository meeting our requirements, this is, a data set composed of sentences retrieved from academic texts and from social media messages that contain reports on the use of a closed set of drugs that could be used in our linguistic study to provide our second contribution.

Our second contribution focuses on the linguistic register where we explore it by using a well established method (Multidimensional analysis) from the area of linguistics that is known for being used to evaluate differences in the formality of texts. By using the multidimensional analysis (MD) proposed by Biber we are able to study the register from a more inclusive point of view as this method does not only account for a set of traits, but it also characterizes the texts using different combinations of the studied features so that other elements such as the "abstractness" or the "narrativeness" of the texts are assessed. The assessment on the differences between generic tweets and drug-related tweets shows that even if both of them belong in the same type of register Biber's schema is able to capture the differences and helps in telling apart the drug use reports due to their higher level of informativeness. Similarly, an analysis comparing the set of drug-related tweets and the corpus of PubMed sentences shows that the academic

texts have more traits of "Information Productions", proving that the MD analysis can capture those characteristics.

For our third contribution we use a set of features able to capture differences in the register and explore the power of those features in drug safety systems. For that we prepare four different binary classifiers for the tasks of detecting sentences containing first-hand experience reports on the drug use, sentences containing beneficial outcomes, sentences containing negative outcomes and sentences containing any type of outcome (positive or negative) related to the drug use. We also build a named entity recognition (NER) system to detect the mentions of drugs and diseases and symptoms. Those systems also explore additional set of features that MD analysis did not assess, but which are known to carry important formality information. Those experiments show that for the set of assessed classifiers using PubMed texts the use of Biber features, the use of custom expansion to those features, nor the use of our set of Politeness features can beat the other configurations of the classifiers that do not include such register-related information. We can see, though, that the two of the classifiers for Twitter data do benefit from the use of our custom set of Biber features and from the use of our set of Politeness features as those two sets of features provide significant gains to the baseline system. On the other hand, when evaluating NER systems targeting at the identification of drugs and symptoms and diseases we saw that the set of Politeness features combined with the baseline and word2vec features scored the best result in both PubMed and Twitter data sets, showing that these features can contribute to pharmacovigilance systems.

Acknowledgements

I want to start by acknowledging the Japanese Ministry of Education, Culture, Sports, Science and Technology, MEXT, as I had the luck of obtaining one of the scholarship they provide to Spaniards. That provided the backing, both economically and logistically, I needed to pursue my research in Japan.

To my family I owe my deepest gratitude as they always backed me at each step I took, no matter how hard any decision would be for them, always having in mind what would be best for me. Víctor and Laura have been a source of unlimited strength and inspiration, same as Jose Andrés and Teresa, although the role of my parents was much harder as they had to encourage me without showing any sign of sadness because of the distance between us.

There are a number of people who should be thanked here besides my closest relatives, cousins, aunts and uncles. First of all I would like to name Gonzalo Domínguez as he helped since minute zero on my endeavour to complete my PhD. Besides the funding I receive from MEXT, I feel that none of this would have been possible without Gonzalo's support and advice. Another key person during my PhD. has been Nahúm: he has guided me at many different levels providing me with invaluable knowledge he acquired after facing numerous challenges. I am very glad he led the way before me so that I did not have to face as many challenges as he did, and I am happy because we had a six months overlap during both of our PhD. studies when we created very good memories.

I also want to acknowledge friends from my childhood: Jaime Mayor, Óscar, Jaime López, Cristina, Elena, José de Diego, Roberto, Enrique, Guille and David; friends from my home university: Pablo, Rubén, Luis, Manuel Moraleda, Javier Nares, Javier Fuentetaja, Iván and Rosa; and friends whom I met while at work: Alex, Alberto, María, Javier, Sebas, Gonzalo García, José A.M.P. and Quique Canorea as all of them showed me their support in many different ways.

When I started my research period I was not sure how to address many different issues, but I was lucky enough to be guided by Prof. Collier. He provided me with all the support I could need. I can not put in words how much I appreciate all what Prof. Collier has done for me.

While at the middle of my PhD studies Prof. Collier had to leave my university and I continued under the guidance of Prof. Miyao, who happily accepted to guide me during the rest of my thesis. I want to thank Prof. Miyao for his time and help, specially taking into consideration that he was already assisting many other students when I joined his laboratory.

All the members of the committee who reviewed my thesis provided very useful feedback, and even if Prof Collier and Prof. Miyao closely followed up on my research I can not forget Prof. Hideaki, Prof. Kitamoto and Prof. Tsuruoka as they also helped me with their insightful comments and pointed me to the right direction with their advices. All the input and advices I received from the whole committee was very much appreciated.

Prof. Miyao's laboratory had a number of members that provided very useful input, and I can easily recall Yin, Bevan, Christian, Phang, Hoshino, Noji, Tam and Tateisi among others, but one name stands out: Pascual. He guided me as if he was another adviser from the committee and supported me even before I joined Prof. Miyao's laboratory.

Ramiro also deserves a special mention as he helped me at a moment when I needed him the most. He provided me with the tools I required to curate my data sets. For this task I also received the invaluable help of Ana and Marta, who annotated the data set I needed for my research, and whom I want to thank for providing high quality annotations in a timely manner.

I also met very amazing people while pursuing my PhD, and all of them enriched me in one way or another. I am sure I will forget some names, but I hope they understand my gratitude does not only go to the people listed here. Some of them are Oussama, Ossa, Vanessa, Viktors, Tokuda-san, Amano-san, Yamaguchi-san, Kobayashi-san, Nut, Juan M. Banda, Jin-Dong, Carmen and Andrés.

To my wife, Lucía, I owe my greatest gratitude as she pushed towards the completion of this research as hard as I did, helping me as much as she could on everything I needed.

Finally, I want to thank everyone reading my thesis for taking the time to go through it. I hope it helps you in some way.

Thank you!

Contents

List of Figures

List of Tables

Abbreviations

ADR	**A**dverse **D**rug **R**eaction
ADE	**A**dverse **D**rug **E**vent
AE	**A**dverse **D**rug **E**vent
API	**A**pplication **P**rogramming **I**nterface
ChEBI	**Che**mical **E**ntities of **B**iological **I**nterest
CUI	**C**oncept **U**nique **I**dentifier
EMA	**E**uropean **M**edicines **A**gency
FDA	**F**ood & **D**rug **A**dministration
FTA	**F**ace-**T**hreatening **A**cts
ISO	**I**nternational **O**rganization for **S**tandardization
MD	**M**ulti **D**imensional
ML	**M**achine **L**earning
MHRA	**M**edicines and **H**ealthcare products **R**egulatory **A**gency
NCBO	**N**ational **C**enter for **B**iomedical **O**ntology
NER	**N**amed **E**entity **R**ecognition
NLP	**N**atural **L**anguage **P**rocessing
PATO	**P**henotypic **Q**uality **O**ntology
PCA	**P**rincipal **C**omponent **A**nalysis
PFA	**P**rincipal **F**actor **A**nalysis
POS	**P**arts **O**f **S**peech
SSRI	**S**elective **S**erotonin **R**euptake **I**nhibitors
WHO	**W**orld **H**ealth **O**rganization
YCS	**Y**ellow **C**ard **S**cheme

Formulae

Cohen's kappa	κ	$=$	$\frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$,
Covariance	$\mathrm{cov}(X, Y)$	$=$	$\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{2}(x_i - x_j) \cdot (y_i - y_j)$
F-Score	F-Score	$=$	$2 * \frac{precision * recall}{precision + recall}$
Fleiss' kappa	κ	$=$	$\frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$
Information Gain	IG	$=$	$H(Class) + H(Attribute) - H(Class, Attribute)$
Informedness	Informedness	$=$	$recall + invRecall - 1$
Inverse recall	invRecall	$=$	$\frac{true\ negatives}{true\ negatives\ +\ false\ positives}$
Jaccard similarity coefficient	J(A,B)	$=$	$\frac{\lvert A \cap B \rvert}{\lvert A \cup B \rvert}$
Kendall's Tau	τ	$=$	$\frac{\text{(number of concordant pairs)} - \text{(number of discordant pairs)}}{n(n-1)/2}$
Precision	P	$=$	$\frac{true\ positives}{true\ positives\ +\ false\ positives}$
Recall	R	$=$	$\frac{true\ positives}{true\ positives\ +\ false\ negatives}$
Spearman's Rho	ρ	$=$	$\rho_{\mathrm{rg}_X, \mathrm{rg}_Y} = \frac{\mathrm{cov}(\mathrm{rg}_X, \mathrm{rg}_Y)}{\sigma_{\mathrm{rg}_X} \sigma_{\mathrm{rg}_Y}}$
Standard deviation	σ	$=$	$\sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})^2}$

Dedicated to my Family

Chapter 1

Introduction

This first chapter introduces the main ideas that are discussed in the thesis, beginning with a description of the motivation and an overview of the linguistic register and pharmacovigilance as the major areas where the efforts are being put. We then proceed to focus on the current area of interest of this thesis and present the problems by introducing the three main contributions as well as the hypothesis driving this work. We conclude by giving an overview on the rest of the chapters included in this thesis.

1.1 Motivation

The interest on drug safety has been present in the society for many years [1], and even if the methods to monitor the outcomes related to the intake of medicinal products have changed the need for early detection and prevention for adverse drug reactions is a constant. Moreover, the use of the Internet has allowed anyone with access to a computer to relate the reactions linked to drugs use [2]. Same as there has been an increase in these non-technical reports, the number of academic reports (i.e. reports issued by pharmacists or doctors) has also increased at a steady rate [3] providing researchers with vast amounts of information from which useful ADR can be extracted.

Nowadays, a number of researchers use Natural Language Processing, NLP, methods [4–6] to extract the information from the available reports, and even if it is clear that the reports written in scientific journals use different textual constructions than a report contained in an Internet forum those differences have not been explored yet, which is totally understandable as drug safety is a new emerging field in the area of NLP that is capturing the interest of many researchers.

Different pharmacovigilance studies, i.e. drug safety studies, have showed that the performance achieved in the detection of named entities is linked to the type of texts that are being used, and while systems using academic texts show scores close to 85%[1], the performance of systems using informal reports obtained from internet forums and social networks show a constant lower score.

Nikfarjam [6] also showed that the same NER system obtained a 10% difference in the F-score result when using texts from a medical forum (F-score=0.82) and when using drug use reports obtained from a social network (F-score=0.72).

To date, researchers in the area of pharmacovigilance have not explored the differences caused by the use of different linguistic registers, or more formally, the variations in the language due to a particular purpose or to fit a particular social setting.

Our goal is to understand whether we can capture the ***differences in the use of different registers*** *in drug use reports and to test if these differences in the register can provide gains to pharmacovigilance systems.*

As noted above, the main areas of this research are the linguistic register and pharmacovigilance systems, and given these are two different research areas we will start presenting the register and introduce the area of pharmacovigilance afterwards.

1.2 The register

To understand the concept of "linguistic register" we borrow the following examples from Isham [7] to see how the formality of the texts changes while the main idea (i.e. the information being conveyed) remains. In the first case we see the use of a formal register, while the second sentence makes use of an informal register:

- **Formal register**: *"Excuse me, ladies. My mother not only taught me to stand up for my convictions, she also counseled politeness towards those whose beliefs differed from my own."*
- **Informal register**: *"Hey... Hey. Ya know, my mother taught me that it's okay not to agree but the least I could do is be nice about it."*

The variations in the surface of the words used to express the concepts appear in almost any form of communication, and the differences due to the linguistic register remain across domains when the ideas are expressed after an adaptation process so the words

[1] http://banner.sourceforge.net/

that are used take into account a number of elements as could be the formality of the communication, the education level of the speaker, the closeness to the intended audience and other elements such as the race, age, culture, or ethnicity of both the speaker and the audience. Moreover, those listed elements take into consideration other factors, and in the case of formality its level would be affected by the kind of occasion, the social class, and other differences between participants.

For our work we use drug use reports obtained from formal and informal sources, namely PubMed and Twitter, and assess the differences that can be attributed to the register in both groups of reports. The formality in the reports from those sources of information is expected to differ because Twitter is a social network, where each individual message is known as a "tweet", while PubMed is a repository of scientific documents, also known as scientific papers.

To assess the differences in the register we use the multidimensional (MD) analysis proposed in Biber's 1991 study [8], which is a widely used framework in register studies[2] [9–12] and has been of key importance in helping us to assess the factors affecting the *"linguistic register"*.

Biber's study [8] used twenty-three spoken and written registers such as political or financial press reports, editorials, academic documents, radio broadcasts, and university lectures among others. However in the twenty-five years that have elapsed the landscape has drastically changed, and new forms of communication have blossomed and new research fields have emerged.

Although there is not a common agreement between linguistics in terms defining what is "genre", "register" and "style" we will follow Biber's disambiguation [13] to clarify these terms here as well as in the rest of the thesis:

- The **register**: can be understood as the combination of the linguistic characteristics that are common in a text variety with the situation of use of the variety.
- The **genre**: can be understood as the conventional structures used to construct a complete text within the variety.
- The **style**: can be understood as the aesthetic preferences appearing in the text, usually related to particular authors or historical periods.

It is important to note that the register covers very different aspects also including the adaptation in the vocabulary used in a given context [14], which would be a perfect

[2]Having more than 4,600 citations in Google Scholar as of June 2016 https://scholar.google.com/scholar?cites=10294423621661754O8

match for studying the examples we presented above. Besides that very focused study on the register, as Biber described it, it has been widely studied by different researchers under other names such as "style" [15–17], "genre" [18] or "tenor" [19], which illustrates that the register has been actively studied for many years.

Similar discrepancies arise when categorizing the different types of registers: Joos [20] studied the register modelling it into five different groups, being "Formal" and "Casual" two of these groups. The International Organization for Standardization (ISO), has also defined standard ISO 12620 on Data Category Registry [3] covering eleven different types of registers, where also "Formal" is present, and the corresponding term to Joos' "Casual" would be ISO's "Slang".

To keep a clear focus, in this thesis we are going to understand the register as Biber did and study how it is used in informal texts, or colloquial texts as Biber referred them, as well as its use in formal texts from scientific publications from a NLP (Natural Language Processing)[4] perspective.

In the NLP area the study of the linguistic register is not new, and even if not all NLP areas have explored the use of different linguistic registers one example is the area of machine translation where a number of studies took that linguistic perspective into account [21, 22].

As for the data used to perform the study of the register we take into consideration that in a different study Biber mentions that most English grammar studies have used a collection of texts, or corpus, that was readily available to the researcher, and one problem that there has often been is the lack of control for register[23] as most studies were either based on a single register or based on discourse examples with disregard to register.

Even if we put the required measures in place to control for the use of a certain types of register there are other elements that can bring in variability. To reduce that external variability we decided to control for the main two external elements:

- The **domain**: can be understood as the subject field [24]. It is the area of knowledge upon which the text orbits. Examples of domains are the "biomedical" domain or the "legal" domain. It is important to notice that a single domain can have other subdomains, and in the case of the "legal" domain possible subdomains would be "treaties", "regulations", "laws", "ordinances".

[3] http://www.iso.org/iso/catalogue_detail.htm?csnumber=37243
[4] See Abbreviations section (6)

- The **topic**: can be understood as the lexical aspect of internal analysis of a text [24]. It is the main theme where the text would be categorized. Examples of topic are "movies", "music", "games" and "restaurants".

Within corpus linguistics, the study of the register in is not new and in NLP it has been hard to leverage the notion of register. The register is often thought to be bound up with topicality and domain, and even if there are a number of studies on domain adaptation [25, 26] few of them are explicitly studies on the register. For these reasons it is important to see if we can, besides controlling these elements to reduce variability, utilize these notions and understand the effect of the "register" since it does in fact seem to be real and important.

1.3 Pharmacovigilance

To give a better view of the chosen domain we should say that pharmacovigilance, also known as "drug safety", is defined by the World Health Organization (WHO)[5] as the science relating to the collection, detection, assessment, monitoring, and prevention of adverse effects with pharmaceutical products [1]. Pharmacovigilance heavily focuses on adverse drug reactions (ADRs)[6] which are any response to a drug which is noxious and unintended.

In the area of drug safety there are two main trends developing in parallel having a key difference in the corpora they use as it comes from very different sources of information where the main difference is the linguistic register. In one case, we have the systems that use formal scientific texts [27–29], mainly obtained from published papers. On the other hand we can find the systems that are fed with texts collected from social networks or forums [30–32].

Here we provide two examples of drug use reports from formal and informal texts:

- *"I need to come up on an **addy** prescription asap, my **concentration skills** are non existent"* (text from Twitter[7])
- *"Drugs like methylphenidate (Ritalin, Concerta), dextroamphetamine (Dexedrine), and **dextroamphetamine-amphetamine (Adderall)** help people with ADHD feel **more focused**"* (text from PubMed[8])

[5] http://www.who.int/en/
[6] This and other acronyms can be found in Abbreviations section (6)
[7] Tweet extracted from https://twitter.com/JaslynDiaz01/status/691462610512908288
[8] Excerpt extracted from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3489818/

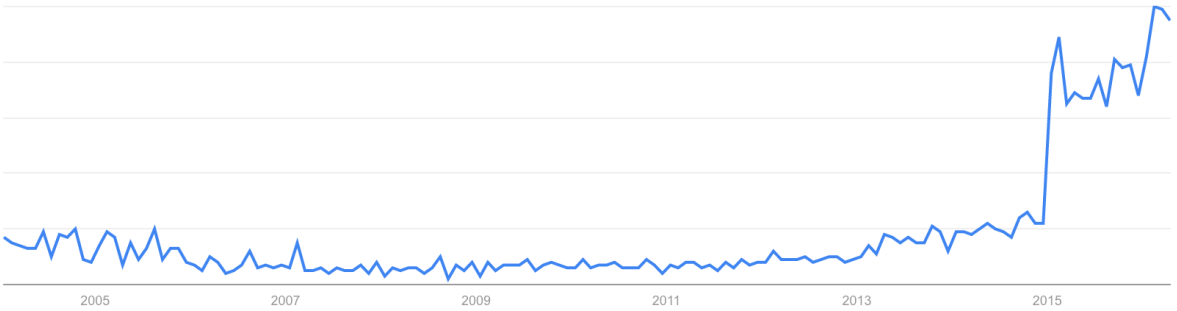

FIGURE 1.1: Interest over time on the term "precision medicine" as reported by Google trends https://www.google.com/trends/explore#q=precision%20medicine.

It is clear that even if the drug, introduced as ***addy*** in one case and as ***dextroamphetamine-amphetamine (Adderall)*** in the other, and the concept related to the increase in attention, presented as ***concentration skills*** and ***more focused***, do not have a resembling surface both sentences convey the same message about the drug and its contribution to increasing attention.

The exploration of Pharmacovigilance from a NLP perspective is a new area of research and as such a number of approaches are still to be assessed. To date, researchers working on this area of knowledge have not fully explored most linguistic features, and even if some studies have taken into account the use of linguistic negations [33, 34], a number of approaches only focus on the use of part of speech features or n-grams, and there is still room for improvements. Moreover, we believe pharmacovigilance is a promising field as interest on the area of precision medicine, also known as personalized medicine [35], and other disciplines related to pharmacovigilance are getting more and more attention (see Figure 1.1).

Drug safety, despite its popularity is vital by itself given that early detection of ADRs can help in knowledge acquisition [36–38], that can be used to impact positively on patient's health and even save patients' lives.

For these reasons, besides the study of the register perspective from a linguistic point of view we aim to study if register-related features can contribute to NLP (Natural Language Processing) systems in the area of drug safety. By doing so we aim to provide a two ways contribution: On the one hand we are going to study the linguistic differences between the same kind of drug use reports in two different linguistic registers differing in their level of formality. On the other hand, we are going to explore whether a linguistic approach using the register perspective can provide gains to a pharmacovigilance system.

1.4 Problem Statement

The main problem driving this thesis is the understanding of whether the differences in the formality of drug reports obtained from different sources of information can provide gains to pharmacovigilance systems. While addressing that main question we produced three different contributions. The first contribution overcame the first problem we faced when we did not find appropriate data sets where we could develop our study. Our second contribution is the linguistic study in which we assessed the differences in the formality in texts from Twitter and Pubmed. The third contribution produced the study of the gains provided by register-related features in pharmacovigilance systems. These contributions were the result of studying the hypothesis that we introduce in this section.

1.4.1 The need for formal and informal data for linguistic studies on pharmacovigilance

To face our problem we used two very different types of registers within the pharmacological domain. In particular, we studied different drug use reports coming from formal and informal texts. On one hand we took the formal, scientific, drug use reports found in PubMed texts. On the other hand we focused on drug use reports obtained from Twitter.

Our understanding, backed by the number of published papers using corpora from either domain [27–32] is that both sources of information are very useful for drug surveillance systems as drug use reports found in PubMed and Twitter will contain valuable information in terms of drugs and symptoms or diseases related to the intake of those compounds, the rationale behind this is that formal drug use reports can be used to fed a system to be aware of new scientific findings as well as to recognize adverse drug reactions (ADRs) in an automated way. On the other hand, social media users' reports, i.e. informal reports, could be contributing to new paths of research in case these drug-symptom reports are not in the data bases or those reports help in detecting a potential health problem. In brief, those data can provide an interesting source of knowledge that can be used to improve patients' condition.

1.4.2 The need for understanding the differences in formal and informal drug use reports

We were interested on assessing how similar were the drug use reports coming from Twitter and the drug use reports from PubMed as we noticed that both sources of

information were used very actively in the area of pharmacovigilance, and we believed that a reason for that was that both formal and informal reports were similar in their contents even if the words used to present the information were of very different nature and the way the reports make use of different elements related to the linguistic register in which the reports are found differ. If that was the case, the information in the messages would be similar in terms of the drugs, symptoms and diseases being mentioned and the relations between them.

Hypothesis 1: *Formal drug use reports in scientific texts and informal drug use reports in Twitter should report similar relations between drugs and symptoms, although they should however be expressed using different registers as shown by Biber's[8] set of features.*

Even if the information would be the same it has been already demonstrated that texts in Twitter are known for the use of very different linguistic resources making it a noisy and informal source of information [39], and although we share the idea that tweets often use very specific formality settings (i.e. informal settings) such as orthographic variations and taboo words the traits that are used when sharing drugs use reports, being it a very specific type of messages, do not have to contain all the elements that are usually found in generic tweets and tweets reporting drug use only utilize a subset of those linguistic elements.

This observation aims to provide useful information in two aspects. First, if we discover drug use reports in Twitter do not include all informal features seen in tweets that can prove that not all tweets share all common traits observed by other researchers, showing that at least tweets discussing medical conditions use higher formality settings, which could provide useful information when using tweets in future studies. Secondly, identifying the linguistic features that are characteristic in these kind of tweets has potential for helping in noticing where further efforts should be put to improve drug surveillance systems fed with tweets.

Hypothesis 2: *The set of register related linguistic features seen in informal drug use reports in Twitter is not the same set of linguistic features that we can observe in generic tweets.*

For testing **Hypothesis 2** we will use the methodology proposed by Biber [8] because it is a well known tool in the area of linguistics for evaluating different features as well as the different aspects of the texts using different registers.

Although **Hypothesis 1** can prove to be useful in characterizing the contents being discussed in formal and informal drug use reports and **Hypothesis 2** will help in discovering if drug use messages are not expressed using traits commonly seen in generic

tweets there is still one unknown regarding the linguistic constructions used in formal and informal sources of information that we should assess.

Our understanding is that if we can accept **Hypothesis 2** and prove that the constructions used in social media reports are not the usual constructions we see in social media texts, and we also accept **Hypothesis 1** and show that the contents in formal and informal drug use reports are similar, then Biber's approach [8] may not be able to clearly detect that the drug use reports from Twitter and PubMed are in fact coming from different registers.

Hypothesis 3: *Biber's approach fails to completely describe register differences in formal and informal sources between formal and informal drug use reports.*

1.4.3 The need for testing the contribution of the register in pharmacovigilance system

Once these hypotheses have been explored we would have a better knowledge of the differences and similarities between drug use reports in different registers, and also have a new set of features able to capture those differences. Knowing that pharmacovigilance systems have not made use of register-related features yet, our goals is to assess if recognizing those differences in the formality have potential for contribution in pharmacovigilance systems, and given that Biber's MD analysis is useful in comparing registers but does not capture all the variations in the formality of the texts we expand our study to account for different formality settings [40] and use that additional information to test its contribution in pharmacovigilance systems as can be binary classifiers and named entity recognition (NER) systems.

Hypothesis 4: *Linguistic features used in register studies can be implemented into pharmacovigilance systems and contribute with gains in accuracy.*

1.5 Contributions

This section gives an overview on the three contributions that we produced as a result of our work. These are:

- The development of new resources, i.e. corpora, to perform our studies.
- A linguistic analysis of the differences between formal and informal pharmacovigilance reports.

- An assessment on the impact of the register-related features in pharmacovigilance systems.

1.5.1 First contribution

To begin with our research we explore Twitter and curate a corpus of first-hand experience drug use reports. The reasoning behind that is that to evaluate drug use reports from both formal and informal sources one key element, also mentioned by Biber [23], should be the correct choice of the corpus that would be used in the study. We found that in most cases corpora composed of Tweets were not directly available due to Twitter Terms and Conditions[9] that disallow the direct share of tweets as they clearly state: *"If you provide Content to third parties, including downloadable datasets of Content or an API that returns Content, you will only distribute or allow download of Tweet IDs and/or User IDs."*. That allows researchers to share the annotated data by providing the Tweet ID, and although that can be of some use that could pose a problem as the existing list of shared tweets can be outdated and some tweets could be off-line by the time of the download.

Another key element to keep in mind is the set of drugs used in pharmacovigilance studies, as it typically varies from one study to another as also does the set of annotated entities and tokens. Annotations tend to target different elements as can be the chemical entity itself, outcomes, symptoms, diseases, or even drug-symptom relations, and that implies that even in the case of having data sets studying the same set of drugs the annotations could vary at great extent, being one example of these the freely available data sets that only provide binary annotations on the ADR mentions.

Bearing those ideas in mind we decided to first agree on the set of drugs that would be part of the study and then look for existing data sets that could help us in our research, finding that none of the available data sets met all our requirements which motivated the curation of our own resources. To curate our corpora we explored two different approaches in terms of data annotation as we used expert annotators and also laymen to helped us in labelling the data.

For our first study we decided to focus on the personal use of popular drugs, i.e. first-hand experience, finding that no other study was targeting at the same set of drugs that we wanted to include in our research, for which we curated a corpus to be used in our study, which is the first by-product we produced, presenting it in 3.1. While exploring the potential of Twitter as a source of data for a register study we try to answer the question of whether we can use Twitter as a reliable source for building a system to

[9] https://dev.twitter.com/overview/terms/policy

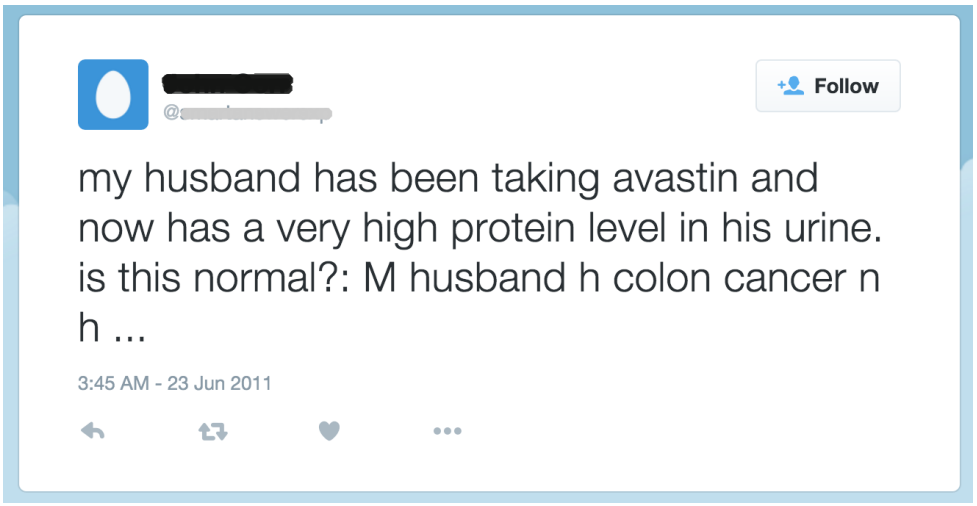

FIGURE 1.2: Non-first hand experience tweet on the drug use (avastin).

extract first-hand experience reports on drug use, and the question of whether we can rely on laymen to help us in curating the corpus for such a system.

To create a similar corpus of drugs we used the texts from PubMed and PubMed Central, PMC, and curate a corpus using the same set of drugs from the previous study and also constraint the list of extracted sentences to those mentioning some keyword related to patients under the assumption that those sentences would contain drug use reports. This corpus, known as "Neuroses", was also produced as part of this study and it is another by-product freely available online as described in 3.2.

Having those data sets ready we realized that there were three key points that we could improve to produce a corpus of much higher quality and use it in our study:

- In the case of Twitter some drugs appeared much more frequently than others, thus biasing the sample.
- The list of drugs was very focused on two types of drugs, and a more diverse set of medicines could capture more insights on the data.
- We were missing important information by only using first-hand experience reports from Twitter, as some reports from relatives or doctors were left out (See Figures 1.2 and 1.3).

For the first and second points we decided to expand the list of drugs to be used in our study to include drugs studied by other researchers. The third point was also addressed by including in the study any drug use report containing drug mentions appearing in a sentence reporting symptoms or diseases, which would include in the study tweets as

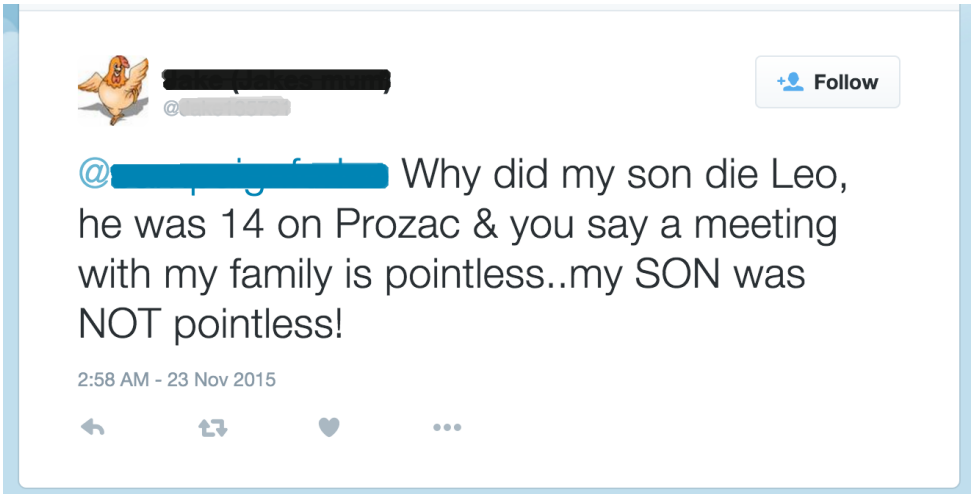

FIGURE 1.3: Non-first hand experience tweet on the drug use (prozac).

the ones presented in Figures 1.2 and 1.3. This improved version of the drug use corpus is explained in detail in 3.3.

Additionally, our work to produce these data sets allowed us to detect some entities and relations that caused disagreements in the annotation between pharmacists. For these findings we explain the problematic elements, the causes for those differences in the annotation, and present our strategy for reducing those disagreements.

1.5.2 Second contribution

While curating this third corpora we studied if, as stated in **Hypothesis 1**, the contents found in formal and informal drug use reports were similar. To measure the similarity of the contents we inspected the different information that we found in each data set. Our understanding was that the information to be compared should be the one that we would record in a database, this is, the relations between drug, symptoms and diseases found in the sentences. By studying which were the drug-related reports mentioned in each source of information we got an idea of the similarity of those drug use reports, and addressed **Hypothesis 1** discovering that there is very little overlap in the drug use reports in each source of information, although in the case of "Outcome-negative" relations the similarity between the drug use reports in PubMed and Twitter texts was strikingly low.

The **Hypothesis 2** was motivated by the assumption that drug use reports in informal media are probably not sharing some of the traits commonly seen on generic social media messages due to the fact that these reports are providing important content, and elements typically seen in social media messages such as contractions or slang appear at a much lower extent in drug use tweets. To address **Hypothesis 2** we gathered generic

tweets, i.e. tweets retrieved from the API without applying any strong constraint nor filter, and also drug-related tweets, i.e. the tweets distributed in our TwiMed corpus and presented in 3.3. We then compared our data sets using Biber's approach [8] and found that the tweets containing drug use reports had some features that characterised them as more informative using Biber's schema.

To test **Hypothesis 3** we used the set of sentences from PubMed and Twitter included in TwiMed corpus 3.3 and applied the method proposed by Biber [8] in the same way as we applied it to test **Hypothesis 2**. In this case too, we discovered that the most salient differences were the features related to the informativeness of the texts, and confirmed that PubMed texts were more informative in general. We also observed that some of the features that were different between generic tweets and drug related tweets also appeared when comparing drug related tweets and PubMed texts. In this case we saw that Biber's schema reported that the set of drug related tweets were not so different from the set of PubMed sentences.

1.5.3 Third contribution

Our last contribution was aimed at the area of pharmacovigilance to study positive and negative drug use reports, to understand which are the set of features that help in detecting either report, and also to assess which features vary depending on the type of register in which the reports are written. Addressing **Hypothesis 4** showed that there are some features that can provide gains in NER systems for pharmacovigilance and in classifiers targeted at detecting sentences containing drug use reports describing both beneficial as well as negative outcomes, and the gains provided by these features have different impact in systems using Twitter and PubMed corpora.

1.6 Outline

- **Chapter 2**:
  In this chapter we present the background and diverse researches performed by different groups to give a grounding on the area of linguistics and pharmacovigilance, which are the main topics treated in the rest of the thesis.

- **Chapter 3**:
  This chapter presents the corpus selection strategy and annotation details, and explains the decisions we made, the problems we encountered, and the findings we discovered. We conclude the chapter by explaining the details of the data we shared with the community.

- **Chapter 4**:
  In this chapter we study the variation in the information contained in drug use reports obtained from Twitter and PubMed. We also assess the register used in different data sets composed of generic tweets, tweets including drug-use reports, and a corpus composed of PubMed sentences. In this chapter we answer **Hypothesis 1**, **Hypothesis 2**, and **Hypothesis 3**.

- **Chapter 5**:
  In this chapter we assess the performance of NLP systems (in particular, a set of binary classifiers and a NER system) and use the set of features assessed in the previous chapter to enhance these NLP systems. This assessment is also complemented with the study of additional register-related features to cover different aspects related to the formality of the texts. In this chapter we answer **Hypothesis 4**.

- **Chapter 6**:
  In this final chapter we present the conclusions from our study.

Chapter 2

Background

This thesis builds on two areas: linguistics, as we focus on the study of the register, and also on the area of drug safety or pharmacovigilance, as we use that domain for our register studies. Bearing these ideas in mind we are going to present here the background in these two fields, beginning with the linguistic area.

2.1 Register studies

Conversation is the most common type of spoken language that people produce. It can be seen in television shows, commercials, news reports, and political speeches to name a few. Similarly to spoken language, the texts we all read are of different kinds: newspaper, magazines, e-mails, blog posts or history books.

Each of those kind of texts has its own characteristic linguistic features and, as Biber shows [13], even if the following conversation is often heard it would be inconceivable that this sentence would end in a textbook:

ok, see ya later.

Biber explains that it is much more common to see a sentence such as the following one in a textbook:

Processes of producing and understanding discourse are matters of human feeling and human interaction. An understanding of these processes in registers, genres, and styles language will contribute to a rational as well as ethical and humane basis for understanding what it means to be human.[1]

[1] These are, in fact, the concluding two sentences from a book studying conversational styles [41].

the **Situational Context** of use (including communicative purposes)	< – – – **Function** – – – >	**Linguistic Analysis** of the words and structures that commonly occur

TABLE 2.1: Components in a register analysis as described by Biber

Biber also clarifies that *"a register is a variety associated with a particular situation of use (including particular communicative purposes)"* [13], and explains the three major components that are covered by the register: the situational context, the linguistic features, and the functional relationships between the first two components. He illustrated these elements using the Table 2.1 where we can see that the registers are described for their typical lexical and grammatical characteristics, i.e. their linguistic features, and also for their situational contexts.

One of the central arguments of his book is that when the linguistic features are considered from a register perspective they are always functional. Biber clarifies his point by stating that *"linguistic features tend to occur in a register because they are particularly well suited to the purposes and situational context of the register"* [13], which is a way to express that the third component of any register description has to be the functional analysis.

When talking about previous art on the *register* we have to stress that there is no general consensus concerning the use of *register* and related terms such as *genre* and *style* among linguists.

One of the reasons for this to happen is that *register* and *genre* have both used to refer to varieties associated with particular situations of use and particular communicative purposes, and that caused that many studies [8, 42–48] simply adopted the term *genre* to cover these concepts and disregard the term *register*. Conversely there is also a number of studies where only the term *register* was used [49–57].

Regardless the term being used the key idea is the linguistic aspect that is under evaluation, and even if the used keyword was *genre* or *register* in each case different areas were at the center of the research. In this study we use the distinction stated on Biber's book [13], and focus on the register perspective:

- The **genre perspective**: focuses on the linguistic characteristics that are used to structure complete texts. The genre perspective usually focuses on language characteristics that occur only once in a text
- The **register perspective**: characterizes the typical linguistic features of text varieties, and connects those features functionally to the situational context of the

variety. The focus is on words and grammatical features that are frequent and pervasive.

The study of the register has attracted the interest of many researchers including the spoken registers used in corporate meetings [58] or the spoken registers characterizing a classroom discourse [59, 60] to name two different types of spoken registers. The features of interest in each study were different in most cases, and taking back the previous register study on the K-12[2] classroom the subject of interest were the discourse practices in one case [59] and the genres and macrogenres in the other [60] evidencing the vast area of research that is covered by the register.

When the focus is put on written registers we can find that researchers also assessed different elements in scientific articles and academic papers as are the lexico-gramatical moves and features [61, 62], moves and reporting verbs [63], the use of hedges[3] [64], the textual and interpersonal metadiscourse [65], *that*-clauses [66], the use of concrete nouns [67], the frequency of rethorical structures and modal verbs [68], the politeness strategies [69], the modality expressions [70], and the types of references, e.g. quotation, used in the research articles [71] to name a few. The authors pointed out common features found in academic texts characterizing it as highly informative, non-narrative and using a personal style [8, 62, 63]. Scientific texts were also found to make extensive use of hedgings [64, 65] and modality expressions [68, 70].

Modern types of texts have been also studied from a register perspective. Crystal [72] studied the common characteristics of internet registers such as e-mail and chatgroups to find some distinctive features as can be the use of lower case, spelling conventions and messages length, concluding that the features he found were typical of face-to-face conversations.

Following that study Thurlow [73] gathered a corpora composed of mobile phone messages, or text messages, and studied different features such as shortening, contractions and the use of letter and number homophones (e.g. “U” instead of “you”), finding that those messages were remarkably short and made extensive use of non-standard features.

In those researches we can see the study of the register in a similar was as we aim to address it, but one missing element is the study using texts on the same topic with different formality features. Such a study has not been fully explored in the area of pharmacovigilance, and the only study having some similarities is the one from Grabowski [74] where he studied the variation of the recurrent linguistic patterns in two different pharmacological texts: patient information leaflets and summaries of product characteristics.

[2] K-12 is a term for the sum of primary and secondary education ranging from kindergarten (K) to twelfth grade (12).

[3] Hedges refer to the use of a cautious language (or “vague language”): “seem”, “may”, “usually”...

Grabowski found that the patterns of language use were different and the differences were linked with the situational and functional characteristics of the studied types of register.

Grabowski continued his line of research and expanded on the previous study adding two different registers [75], namely clinical trial protocols and chapters from academic textbooks on pharmacology, to the registers he studied in his previous work. Showing that patterns of language use differ considerably due to topic and function-related differences between the text types, despite dealing with a similar theme: medicinal products (medicines).

In the studies from Grabowski it is clear that his efforts were only put in formal registers, which is an important difference with our study as we will also include texts using an informal register. One more key difference is the area of interest as he focused on the use and functions of keywords and also identified the top-4 lexical bundles, which are the occurrences of 4-consecutive words[4], in each type of register.

For our study we are going to use the multidimensional analysis as proposed by Biber [8], which is a method aimed at assessing different aspects of the texts.

2.1.1 Biber's multidimensional analysis

As a way to perform his register studies Biber opted for performing a multidimensional (MD) analysis as these dimensions "provide comprehensive descriptions of the patterns of register variation" [55]. The way in which MD studies act is by:

- Identifying underlying linguistic parameters of variation. These parameter are also known as "dimensions".
- The information for each one of those "dimensions" is then used to specify similarities and differences among registers.

To clarify what Biber understood as a dimension it is important to note that the dimensions were used to cover a range of linguistic features. That was due to the fact that a single feature alone was not enough to determine a register, and for that reason features were grouped in "dimensions". Moreover, the dimensions allow the researchers to analyse whole texts, and not individual constructions. In a way, Biber's MD study could be presented as a comparison of co-occurring features among different texts.

[4] In the area of NLP these lexical bundles are known by the name of word n-gram. In this case 4-word n-gram.

The set of linguistic features that Biber used in his multi-dimensional analysis contained a total of sixty-seven linguistic features [8] to capture different linguistic aspects of the texts. Among other, those features covered:

- **Semantic features**: Such as Hedges[5], or the use of "speech act verbs"[6].
- **Grammatical features**: Such as the nouns, or predicative adjectives.
- **Syntactic features**: Such as relative clauses, or the use of passive constructions.

To study those features using Biber's method the main steps are:

- Tag texts with features (e.g. via an automatic tagger).
- Compute frequency co-occurrence patterns of linguistic features using factor analysis.
- Sum the features on each dimension.
- Use the mean dimension scores for each register to analyse similarities and differences.

Applying this method provides a common framework where frequently co-occurring elements are grouped together, and the resulting groups can be compared as if they were a "dimension" of the text.

In particular, the way in which the MD analysis works is by:

- Building a correlation matrix of all features.
- Use the correlation matrix to determine the loading, or weight, of each linguistic feature[7].

These weights are used to indicate the strength of a feature in the corresponding dimension. In his analysis a positive weight value characterized a positive correlation, while a negative weight indicated a negative correlation, and the higher the absolute value would be the more representative the feature would be to characterize that dimension.

In his analysis Biber first computed sixty-seven different features, which he grouped using Principal Factor Analysis (PFA)[8] obtaining seven linguistic dimensions[9]. After finding those seven dimensions he interpreted them as explained below:

[5] Constructions used to lessen the impact of an utterance such as "almost" or "maybe".
[6] E.g. "acknowledge", "affirm", "agree"...
[7] All the weights are in the range from -1.0 to 1.0.
[8] Biber used PFA over Principal Component Analysis (PCA) because PFA accounted for the shared variance instead of all of the variance.
[9] Although Biber computed 67 features the linguistic dimensions only make use of 59 of those features.

- (1) **Involved *versus* Information Productions**: Marks affective, interactional and generalized content versus high informational density and exact informational content.
- (2) **Narrative *versus* Non-Narrative Concerns**: Distinguishes narrative discourse from other types of discourse.
- (3) **Explicit *versus* Situation-Dependent Reference**: Distinguishes between highly explicit, context-independent reference and non-specific, situation-dependent reference.
- (4) **Overt Expressions of Persuasion**: Marks persuasion, including the speaker's own persuasion or argumentative discourse designed to persuade the addressee.
- (5) **Abstract *versus* Non-Abstract Information**: Indicates abstract, technical and formal informational discourse.
- (6) **On-Line Informational Elaboration**: Marks informational discourse but produced under real-time conditions.
- (7) **Academic qualification**: Marks academic qualification or hedging.

This MD analysis was first used by Biber to compare twenty-three different written and spoken registers [8]. In a different study Biber used it to compare different written and spoken registers in four different languages [55], and also used it to assess the differences in lexical and grammatical features [56].

Besides being used by the creator of the study, Doublas Biber, the MD analysis has been widely used to evaluate differences between registers for many years, and new sources of information that did not exist when it was first published have also been assessed: A study published in 2015 [76] compared Internet and pre-Internet text varieties using the MD approach on a corpus of webpages, blogs, emails, Facebook messages and Twitter messages, or tweets. The results from that study show that the used Internet registers are not so different from the pre-Internet registers, and even if the new-born registers have particular characteristics that set them apart there also are considerable linguistic similarities between pre and post-Internet registers.

Computing Biber's dimensions

Biber MD analysis studies the differences in the seven dimensions we presented before, and in order to obtain the values for those dimensions a number of features are taken

#.Dimension	(#) Features
1.**Involved *versus* Information Productions**	(29) private verbs, THAT deletion, contraction, present tense verbs, second person pronouns, DO as pro-verb, nouns, word length, prepositions, type/token ratio, attributive adjective
2.**Narrative *versus* Non-Narrative Concerns**	(6) past tense verbs, third person pronouns, perfect aspect verbs, public verbs, synthetic negation, present participal clauses
3.**Explicit *versus* Situation-Dependent Reference**	(8) WH relative clauses on object positions, pied piping constructions, WH-relative clauses on subject positions, phrasal coordination, nominalizations, time adverbials, place adverbials, adverbs
4.**Overt Expressions of Persuasion**	(6) infinitives, prediction modals, suasive verbs, conditional subordinations, necessityModals, split auxiliaries
5.**Abstract *versus* Non-Abstract Information**	(6) conjunctions, agentless passives, past participial clauses, by-passives, past participal WHIZ deletions, other adverbial subordinators
6.**On-Line Informational Elaboration**	(4) that clauses as verb complements, demostrative, that relative clauses on object positions, that clauses as adjective complements
7.**Academic qualification**	(1) seem/appear verbs

TABLE 2.2: Biber's Dimensions and features used to compute the value for those dimensions. Features with positive loadings are shown in green. Features with negative loadings are shown in red

into account. In this section we are going to present which are the features involved in each dimension, and the way in which Biber computed the values for the dimensions.

The first thing that has to be clarified are the features involved in each dimension, for which we present Table 2.2. The table also shows that not all features contribute positively, and even if most features contribute with positive weights (shown in green), there are some features (in red) that have a negative weight and decrease the score for the dimension.

As introduced before, we showed that some features have positive weights (indicated in green in Table 2.2) while other features have negative weights (shown in red in Table 2.2), but besides the fact that not all features contribute with weights having the same sign (i.e. positive or negative weights), the features used to compute each dimension have different magnitudes in their weights to denote the particular importance that a feature has when computing the value of a dimension.

The full description for each weights is presented in Biber's work [8] although we present here an example on how to compute Dimension 3 ("Explicit versus Situation-Dependent Reference") using a tweet and a sentence from PubMed (Figure 2.1 and Figure 2.2).

The sentences we are going to use are: "I need to come up on an addy prescription asap, my concentration skills are non existent", obtained from the Tweet shown in Figure 2.1, and the sentence "Drugs like methylphenidate (Ritalin, Concerta), dextroamphetamine (Dexedrine), and dextroamphetamine-amphetamine (Adderall) help people with ADHD feel more focused.", which is a sentence from a PubMed article, as shown in Figure 2.2

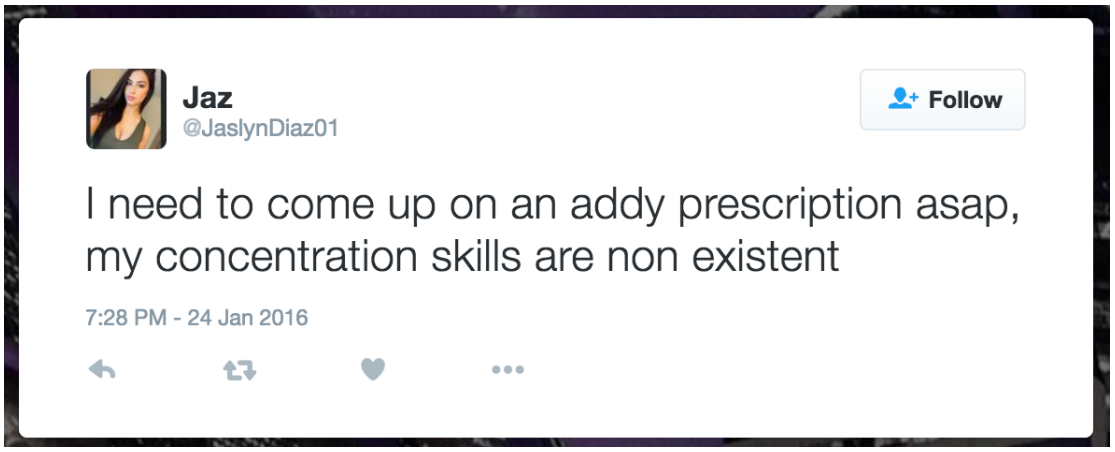

FIGURE 2.1: Sample of a tweet.

Abstract Go to:

Prescription stimulants are often used to treat attention deficit hyperactivity disorder (ADHD). Drugs like methylphenidate (Ritalin, Concerta), dextroamphetamine (Dexedrine), and dextroamphetamine-amphetamine (Adderall) help people with ADHD feel more focused. However, misuse of stimulants by ADHD and nonaffected individuals has dramatically increased over recent years based on students' misconceptions or simple lack of knowledge of associated risks. In this review, we discuss recent advances in the use and increasing misuse of prescription stimulants among high school and college students and athletes.

FIGURE 2.2: Sample of a PubMed excerpt.

As these texts were obtained from different sources of information, we used Charniak-Johnson parser [77] for both tagging and tokenizing the sentence obtained from PubMed, while in the case of Twitter we used ARK tagger [78].

The resulting set of tags after tokenizing the sentences are:

- **Twitter**: [['I', 'PRP'], ['need', 'VBP'], ['to', 'TO'], ['come', 'VB'], ['up', 'RP'], ['on', 'IN'], ['an', 'DT'], ['addy', 'NN'], ['prescription', 'NN'], ['asap', 'NN'], [',', ','], ['my', 'PRP$'], ['concentration', 'NN'], ['skills', 'NNS'], ['are', 'VBP'], ['non', 'JJ'], ['existent', 'NN']]
- **PubMed**: [['Drugs', 'NNS'], ['like', 'IN'], ['methylphenidate', 'NN'], ['-LRB-', '-LRB-'], ['Ritalin', 'NN'], [',', ','], ['Concerta', 'NN'], ['-RRB-', '-RRB-'], [',', ','],

['dextroamphetamine', 'NN'], ['-LRB-', '-LRB-'], ['Dexedrine', 'NN'], ['-RRB-', '-RRB-'], [',', ','], ['and', 'CC'], ['dextroamphetamine-amphetamine', 'NN'], ['-LRB-', '-LRB-'], ['Adderall', 'NN'], ['-RRB-', '-RRB-'], ['help', 'VB'], ['people', 'NN'], ['with', 'IN'], ['ADHD', 'NN'], ['feel', 'VBP'], ['more', 'RBR'], ['focused', 'VBN'], ['.', '.']]

Dimension 3 takes into acount the number of occurrences of WH relative clauses on object positions, pied piping constructions, WH-relative clauses on subject positions, phrasal coordination, and nominalizations, and as features with negative weights it takes into account the occurrence of time adverbials, place adverbials, and adverbs.

In the case of the tweet only the count for nominalizations is greater than zero. The nouns that are included in this count are "prescription" and "concentration" (other nouns such as "addy", "asap", "skills" and "existent" are taken into account in a different category). After finding these two nouns in the sentence we normalize that result by the total number of tokens appearing in the sentence (17), so the resulting value for this feature is 0.117 (2/17). Once having the normalized result for the count of nouns we then multiply it by the corresponding load for that feature (a weight of 0.36, as stated in Biber's MD analysis description), resulting in the final value of **0.042** for Dimension 3 in this tweet.

In the case of the sentence from PubMed we only find one adverb ("more"), and divide that count by the length of the sentence (27), obtaining the normalized score of 0.037 (1/27) for the adverbs. That score is then weighted by the corresponding loads (-0.49 for the adverbs), showing that the final score for Dimension 3 in this PubMed sentence is **0.018**, obtained by using the corresponding score and weight: $0.037 * (-0.46)$

In this example, these results for Dimension 3 tell us that the sentence from Twitter is more explicit (i.e. context independent) than the sentence from PubMed, and conversely, the sentence from PubMed is more situation-dependent (non-specific) than our tweet.

2.1.2 Other supporting studies

Although this thesis orbits around the MD analysis method proposed by Biber there are different elements that can be used when assessing the use of different registers. We make use of some of these additional features in the classifier presented in Chapter 5 and for that reason we take a moment to introduce some of these elements.

Another key factor in the differences in register that can be seen in Twitter and PubMed texts is the politeness, and although it involves many domains: pragmatics, conversational analysis, stylistics, sociolinguistics and ethnography of communication; we follow

the approach proposed by Spencer-Oatey and Jiang [79] and approach it from the domain of pragmatics as sociopragmatic interactional principles.

We study "politeness" using the original theories proposed by Brown and Levinson [40] where the authors presented a universal model underlying the use of polite utterances including both polite friendliness and polite formality. Their research characterizes politeness as the desire to please the interlocutor through a positive manner of addressing, and characterized those interactions as acts that threaten their addressees' face, "Face-threatening acts" or "FTA", to indicate that some actions would threaten the speaker's face, e.g. in the case of using an expression of gratitude as that would indicate the speaker is in debt towards the addressee, as well as face threatening actions towards the addressee, e.g. not caring about the addressee's feelings or wants. By studying those interactions the authors identified both positive and negative forms of politeness.

In our work we only focus on positive forms of politeness and prepare our analysis based on the work presented by Abdul-Majeed [80] to capture the realization of positive politeness strategies in language using some of the strategies he presented as can be the identification of "exaggeration", the use of *in-group* identity markers −also known as address forms−, the use of pseudo-agreement, or the use of jargon among others.

Informal social network messages are known for the use of some of these features [81, 82] and adding them as a way to enhance register features seemed a natural step forward in our study.

As our work is on the register, and considering we use the topic of pharmacovigilance to control for register we will also present the field in the next section.

2.2 Pharmacovigilance

Pharmacovigilance, as defined by the World Health Organization (WHO)[10] is the science relating to the collection, detection, assessment, monitoring, and prevention of adverse effects with pharmaceutical products [1]. Pharmacovigilance is also known by some other names such as "drug safety", and it heavily focuses on adverse drug reactions (ADRs) which are any response to a drug which is noxious and unintended. These responses include the lack of efficacy or medication errors.

Pharmacovigilance, aiming at identifying the hazards related to the use of pharmaceutical products and minimize the risk of any harm that may come to patients, has been a key focus of concern for public health systems, especially in the United States, since at

[10] http://www.who.int/en/

least the turn of the century [83] with an estimated 100,000 deaths attributed to adverse drug reactions every year in US hospitals [84].

In the United Kingdom, the Yellow Card Scheme (YCS)[11], introduced in 1964 in response to the thalidomide disaster[12], is a very well known system for collecting information on suspected adverse drug reactions (ADRs) to medicines that are on the market to be monitored. During 2014 alone, fifty years since its inception, the YCS received 1920 reports on fatal adverse drug reactions [85] showing that it is a successful system.

Even if a number of ADR reports are filed to YCS, on average, 94 per cent of ADRs are not reported [86], which demonstrates the need for higher levels of reporting. A key element for understanding the source of the low levels of reporting is that patients and carers only produce six per cent of reports to the YCS. Both the reports from the patients and the health professionals are valuable as both present two sides of the same coin and patients report different drug reaction types while health professionals report the symptoms and impact of an adverse drug reaction.

New directions point to the use of social media as a source for capturing reports from patients, and in fact some researchers have already explored these possibilities [87]. To date, a number of systems have been built to work on ADRs detection [31], relation extraction [88, 89] and ADR classification [90, 91] using the information obtained from social media messages.

The types of social networks that have been used in those studies are very varied and range from mainstream social networks as can be Yahoo [92, 93] or Twitter [32, 94], to more specific medical support groups and communities like DailyStrength[13] [95, 96] and MedHelp[14] [97, 98].

Even if there is a larger base of social networks such as SteadyHealth[15], PatientsLikeMe[16], DrugRatingZ[17] and ForumClinic[18] that have been already explored by researchers [99–102] the area of pharmacovigilance is not limited to the use of data extracted from social media.

The counterpart to those informal messages are the texts from academic texts. One of these repositories of academic data is PubMed, providing access to references and

[11] https://yellowcard.mhra.gov.uk/the-yellow-card-scheme/
[12] https://en.wikipedia.org/wiki/Thalidomide#Birth_defects_crisis
[13] http://www.dailystrength.org
[14] http://www.medhelp.org
[15] http://www.steadyhealth.com/
[16] http://www.patientslikeme.com
[17] http://www.drugratingz.com
[18] https://www.forumclinic.org

abstracts on life sciences and biomedical topics, and a very used resource in pharmacovigilance [103–105].

PubMed is one of the academic resources that researchers have used to curate a number of corpora for pharmacovigilance studies. Some recently curated corpora are the drug-drug interaction corpus (DDI) [106], the corpus of adverse drug event annotations [107], a corpus of chemicals, diseases and their interactions [32], and the EU-ADR corpus: annotated drugs, diseases, targets, and their relationships [108].

Even if the number of corpora that can be used in pharmacovigilance is growing, the techniques have evolved drastically. By the turn of the century most pharmacovigilance systems were based on manual methods were hospital pharmacists and doctors submitted most of the reports. Healthcare professionals were found to report serious and rare ADRs and ADRs associated with newly marketed drugs more likely than other ADRs, showing that the pharmacists should be properly trained in order to improve ADR reporting [109], and even if that study took place in the United Kingdom similar findings have been discovered in different countries such as Turkey [110], Australia [111], and the United States [112], showing that this is a global issue and the specialists need to be properly trained.

Additionally, a report on patient safety [83] evidenced the need for new safe practices, and a few years after that report a follow-up study showed that developments in this area were not as numerous as one could expect [113, 114]. Similar studies have appeared along the years, showing that better reporting systems that improve the recording and analysis of patient safety incidents aiming at preventing the repetition of incident events are not yet of common use [115].

More recently, researchers have started exploring the area of machine learning and its application to pharmacovigilance using clinical texts [116, 117], academic articles [103, 104], and social media messages [94, 95] showing that even if there is always room for improvement current systems are able to produce satisfactory results at certain tasks [4, 6, 118].

The trend of applying NLP techniques to pharmacovigilance is being fostered by recent venues such as the Informatics for Integrating Biology and the Bedside (I2B2) challenge[19] or BioCreative challenge[20] that host contests aimed at the detection of ADRs.

Another important point is the detection of "off-label" drug use, which is the use of the drugs in a manner that is not approved by regulatory agencies. For this goal, social media has been shown to be a promising data source for pharmacovigilance data due to

[19] https://www.i2b2.org/
[20] www.biocreative.org/

its real-time nature and utility in providing insights into those off-label consumer habits [95, 119].

2.2.1 Drug use reports from different registers

Interest in social media as a signal source seems to be growing as can be seen by recent official announcements: On June 2014, the FDA presented its guidelines on how to use social media [120], and the Medicines and Healthcare products Regulatory Agency (MHRA) announced an application intended to report suspected ADRs, called WEB-RADR [121], on September 2014. EMA (European Medicines Agency) also published guidelines on good pharmacovigilance practices during 2013 [122] indicating that *"marketing authorisation holders should regularly screen internet or digital media"*, clarifying that web sites, web pages, blogs, vlogs, social networks, internet forums, chat rooms, and health portals should be considered [123]. Those announcements show that there is an increasing awareness of the potential for social media as a source of evidence.

Scientific publications would be the counterpart to social media contents as scientific texts do not show some of the problems appearing frequently in social media, being the main differences that there are less ungrammatical constructions, abbreviations and metaphors. However, formal texts pose different challenges being the lack of normalization one of the well known ones appearing when the authors refer to the same relevant entity in many ways, and also when the abbreviations vary with the context [124]. Regulatory and binding events pose another challenge as those events usually have multiple arguments and such complexity makes it hard for NLP tools to extract those events [125]. Similarly, scientific literature is known for the number of new findings it usually includes, and those new discoveries are one of the reasons why extracting core information using text processing approaches is an open problem [126].

It is clear that both formal scientific texts, e.g. PubMed, and also informal sources of information, e.g. Twitter, bring different challenges and opportunities to drug surveillance, and even if the texts' surface in those reports are quite different the underlying information could be equally important. Supported by those findings we can see that understanding the information contained in those reports is an important task, but given the outstanding differences between those sources of information a correct understanding of the discrepancies between those types of texts should be a primary goal. For that, we consider that it is crucial to analyse the formality, i.e. the linguistic register, used in those texts.

2.2.2 NLP methods used in pharmacovigilance

Within the area of pharmacovigilance most NLP systems use very basic linguistic features such as n-grams, bag of words, drug-related lexicons [87], and since more recently word vectors [6, 127] but few of these systems explore the use of other linguistic features to help in the task although in some cases the researchers mention the use of additional linguistic features as a new approach to improve their systems. This is not an observation that can be only seen in the area of pharmacovigilance, and other areas of research such as finance tend to leave out linguistic features related to the use of different registers [128] and we can see that the trends were to explore the use of sentiment features [129] or the use of deep neural networks [130], same as in the area of pharmacovigilance [131, 132].

Although there are a number of different areas within linguistics such as "morphology" −studying the structure of morphemes and other linguistic units−, "orthography" −studying how to write−, or "syntax" −studying the rules involved in the structure of sentences−, we will explore the area of "pragmatics" to study the register used in different drug use reports.

Besides the linguistic features, studies in the area of BioNLP have mostly focused on using lexicons [6, 133], ontologies such as CheBI [134] or Phenominer [135], and adding word embeddings models such as word2vec [136] to classifiers and NER systems [6, 137], and even if BioNLP is also concerned with language as linguistics is, the area of BioNLP does not have many studies exploring other techniques from the area of linguistics.

It could be said that the fields of BioNLP and linguistics have evolved in parallel since BioNLP systems have not included other features studied within the area of linguistics. This opens a door for exploring the potential gains that linguistic approaches could contribute to the area of pharmacovigilance. It is important to note that although not all of those features are expected to contribute equally some of them could be telling an important part of the story that may be missed in current systems.

To fill that gap, and aiming at better understanding the linguistic differences due to the register in an environment where we aim to reduce the variability due to external factors, as can be the domain or the topic, we are going to use texts from the area of drug safety and study the differences in the register found in two sources of information differing in their formality, i.e. formal and informal texts. Additionally, we will also implement classification and NER systems including register-related features as well as other features used in pharmacovigilance systems, and explore if the contribution from the register-related features have potential for providing gains.

Chapter 3

Sourcing the data

This chapter presents the different data sets we produced during the development of the thesis. After clarifying the goals of this study we looked for data where we could test our hypothesis discovering that no resource had the information we required. That finding evidenced the need for such data and led us to the creation of the three data sets that we present in this section. These data sets were produced in an iterative manner by improving the data collection strategies and filtering steps as ways to improve the quality of the data.

The first data set we prepared was composed of messages from Twitter in which the author relates the use of the drug or, as we refer to it, first-hand drug-use reports. When preparing this data set we explored different sentence annotation approaches by hiring laymen and experts. While describing this first data set we also present the different filtering approaches we used to gather the tweets of interest for the study.

Our second data set was exclusively composed of PubMed sentences covering the same set of drugs we used when building the first-hand experiences tweets data set. This PubMed sentences data set was annotated in an automatic way at token level for which we used different APIs available on-line and custom dictionaries. The annotated elements targeted in this process were the drugs and also the phenotypes appearing in the sentences.

The third data set that we present in this section is the final one and is composed of sentences extracted from PubMed and Twitter. Those sentences contain annotations at token level for the set of drugs included in the two previous data sets also covering a larger number of compounds. In this data set we included annotations for the tokens corresponding to the symptoms and diseases appearing in the sentences as well as the relations between them and the drugs. This data set also includes annotations for a

number of attributes for the annotated tokens. The annotations were produced by two pharmacists and this is the data set that we use more extensively in following chapters of the thesis.

These three data sets were produced at different points of the research, but in an iterative manner. To produce the second data set we reused the set of drugs that we used to filter the tweets in our first-hand experiences data set, although the target set of documents changed to retrieve sentences from PubMed, and for that we had to use a drastically new technical approach to first filter the documents, and then extract the sentences containing the drug mentions. For our third data set we improved the coverage as a mean to produce a more balanced sample (in terms of the included drug names) and also to cover more conditions instead of the two conditions of interest targeted when preparing the previous data sets (i.e. "depression symptoms" and "attention deficit hyperactivity disorder", or ADHD). Besides building this final data set by using the previously acquired knowledge we were also able to reuse most of the techniques and tools we prepared for extracting and filtering our two firsts data sets.

For the experimental set up, presented in chapter 5, we start by performing a number of experiments on the the first data set we prepared, this is, the data set composed of first-hand experience tweets. These experiments are shown in section 5.1 . On this same chapter we also present the remaining classification and named entity recognition experiments, in section 5.2, where we use our third data set, and although that chapter does not include experiments where we use the second data set, composed of PubMed sentences, we introduce that second data set in this chapter because it was of great help for preparing our third data set.

3.1 Data set containing first-hand experience drug use tweets

During the last years the number of scientific studies has experienced a remarkable increase passing the 25 million citations mark in PubMed[1]. Those numbers include different areas of research, also covering pharmacovigilance, and provide an excellent source of pharmacological texts using formal register.

Even if the scientific literature is one of the biggest contributors to pharmacological reports there are other sources of information hosting pharmacological reports that can be exploited to complement the information available in scientific texts. A new source of pharmacological reports are the social networks, where Twitter is among the most outstanding contributors as it reportedly has 500 million messages sent per day[2], and

[1] As of April 2016 http://www.ncbi.nlm.nih.gov/pubmed?cmd=search&term=1600%3A2100%5Bdp%5D
[2] http://www.internetlivestats.com/twitter-statistics/

even if most messages are not related to pharmacovigilance, a number of these messages contain publicly available drug reports allowing researchers to carry out drug safety studies [31, 32, 138, 139].

Twitter offers several potential benefits as a source for pharmacovigilance surveillance data. First, a significant fraction of the content is freely available via a public application programming interface (API). Second, the volume of data available is huge, and unmediated by gatekeepers, with approximately 500 million tweets sent per day in 2013 [140]. Third, Twitter content is *"real-time"*, allowing health researchers to potentially investigate and identify new ADE types faster than traditional methods such as physician reports. As such, we regard Twitter as an excellent testbed for our goal of identifying reports of ADRs among potential off-label drug users that may go under-reported by general practitioner visits [96] or undetected in clinical trials [141].

At least one potential unknown is the influence of population bias. Since Twitter users tend to have a particular demographic [142] this may influence the ability of the media to provide useful evidence for some classes of drugs, e.g. those drugs used primarily by paediatric and geriatric patients.

A key difficulty in working with Twitter data, and social media data more generally, is distinguishing between first-hand experiences (*"I feel real groggy after taking <DRUG>"*), second-hand experiences (*"I've heard <DRUG> makes you real tired"*), and other kinds of information related to the drug, like news (*"Court found <DRUG> company liable"*) or advertising (*"Buy <DRUG> now!"*).

As a first stage in gathering data on ADRs, it is vital to identify first-hand drug usage experience. This is a challenging area for natural language processing (NLP) as social media messages contain a high proportion of ungrammatical constructions, out of vocabulary words, abbreviations and metaphoric usage. First-hand experience is defined as being where the person making the report has actually taken the drug. For example, *"<DRUG> is no joke have you up forever took it at 8 haven't been sleepy since #<HASHTAG> #<HASHTAG> #<HASHTAG>"*. On the other hand, a tweet like *"Think I'll just take some <DRUG> and get stuff done instead of sitting here like a worthless piece of shit."*, or *"New Years resolution. Be less boring by staying up past 8pm. #<HASHTAG> or <DRUG>"* would not be classified as first person as there is doubt as to whether the authors have taken the drug.

Previous studies [94] used a reduced set of drugs to compare the adverse events reported on social networks with the adverse events registered in official databases such as FAERS[143], but to the best of our knowledge no studies have explored the genre, i.e. the type of tweet, in which the users refer to the drugs.

Drug Name	Synonyms	# tweets
Adderall	Amphetamine mixed salts; amphetamine salt	300
Ritalin	Concerta; Daytrana; Phenida; Attenta; Hynidate; Focalin;	300
Modafinil	Modafinilum; Provigil; Sparlon; Alertec; Modavigil;	59
Adrafinil	Olmifon;	0
Armodafinil	Nuvigil	3

TABLE 3.1: Cognitive enhancers by drug name along with each synonym and number of tweets.

Drug Name	Synonyms	# tweets
Citalopram	Celexa	65
Escitalopram	Lexapro; Cipralex	145
Paroxetine	Paxil; Seroxat	123
Fluoxetine	Prozac	300
Fluvoxamine	Luvox	14
Sertraline	Zoloft; Lustral	239

TABLE 3.2: SSRIs by drug name along with each synonym and number of tweets.

3.1.1 Data sampling

The drugs selected for our study were either cognitive enhancers, i.e. drugs that enhance some mental function like attention and memory (See Table 3.1), or Selective Serotonin Reuptake Inhibitors antidepressants -SSRIs- (See Table 3.2). SSRIs were selected due to public concerns regarding the risk of suicidal ideation in children and adolescents [144]. The cognitive enhancer drug category was chosen due to the wide spread off-label use of prescription drugs such as *"Ritalin"* and *"Adderall"* as study aids by university students [145]. In terms of specific drugs, for cognitive enhancers we took into account some of the drugs that are anecdotally reported as being popular among the student population [146], while in the case of the SSRIs we analysed widely prescribed drugs identified by previous studies [147]. In both cases we read the existing articles available at Wikipedia on each of the target drugs and obtained a list of synonyms for these drug names as shown in Table 3.1 and Table 3.2.

3.1.2 Annotation

Our annotation efforts were divided into a first annotation step, carried out by expert annotators, and a second phase where laymen were asked to perform the annotations using a superset of documents including the annotations obtained from the experts.

Additionally, we requested the experts to annotate a new list of documents to further increase the size of the corpus.

Those annotation phases are described in detail in the next subsections.

First stage annotation

We used the Twitter streaming API [148] to obtain a random sample from all public tweets for a 12 month period (8th May 2012 - 20th April 2013). This gave us 420,983,674 messages. These data allowed us to understand how Twitter users mention the drugs of interest against a standard background.

Once the full random sample was gathered we used our list of synonyms to identify tweets mentioning any of the drugs of interest (see Table 1.1 and Table 1.2). We then applied a further filter where we would only keep a maximum of 300 matching tweets (selected at random among the matched tweets) for each one of the 11 drugs, aiming at a maximum of 3300 tweets. This was done after we noticed that some drugs such as *"Adderall"* and *"Prozac"* had a far higher number of mentions than the other drugs. In order to obtain a balanced sample we set that upper bound of 300 samples for each drug. Moreover, in the case of *"Adrafinil"* we did not get a single mention on any of the synonyms we used. This can be considered an important finding on the sensitivity of the data source. The final data set used for our study consisted of 1548 tweets (see Table 1.1 and Table 1.2). Since the distribution of drug mentions is not evenly balanced we will investigate a targeted approach in the future in order to increase the volume of rare drug name mentions. With the data in hand we constructed our gold standard annotation set by selecting 496 tweets to be annotated by 2 PhD students with training in computational linguistics.

In order to check for influences on reporting bias we looked for popular stories that appeared during the time frame when we collected the tweets to check possible environmental influences from the media. The stories we found were *"FDA warns of counterfeit Adderall"* [149], *"John Moffitt on Adderall: 'It was a total mistake' "* [150], and *"Aurobindo Pharma gets USFDA nod for Modafinil tablets"* [151]. But on the whole there was no major evidence showing that these would have an impact on the data set we collected during the sample period.

The annotation categories we used were:

- **Tweet written in English language?** This question reported which tweets were written in English language.

- **Tweet about the drugs of interest?** Some drug names appeared as strings within the tweet, providing texts that were not of interest to us.
- **First-hand experience**: Used to identify personal use of the drug.
- **Other's Experience**: Used to identify someone else's use of the drug.
- **Activism**: Used to identify an alarm or call for change in the drug policy.
- **Cultural reference**: Used to identify when the annotator found the tweet referring to a song lyric, movie title, etc.
- **Humor**: Used to indicate that a tweet contained a formulaic joke, bumper sticker, etc.
- **News**: Used to identify news items.
- **Info/resource**: Used to identify factoids or informational resources.
- **Marketing**: Used to identify sales of the drug product/accessory.
- **Opinion**: Used when the writer was reporting a personal opinion related to the drug.
- **Sentiment**: Used to describe whether the author was positive, negative or neutral in terms of sentiment about the drug.
- **Pleasure**: Used to indicate that the writer reports the drug usage as a pleasurable activity.
- **Craving**: Used to indicate that the writer reports stress relief related to the usage of the drug.
- **Disgust**: Used to indicate that the writer sees the studied drug usage or the drug users as repulsive.

Our annotation guidelines for laymen and experienced annotators have been included in *"A. Expert annotator guidelines for annotating first-hand experience tweets."* where we elaborate on the annotation categories we used in the project.

For the initial annotation effort, we obtained the Cohen's Kappa [152] and Fleiss' Kappa [153] values comparing the inter-annotator agreement between experienced annotators as shown in Table 3.3 (columns 2 and 3) by using R's irr package [154]. We studied the Kappa values and identified possible causes of disagreement. These were loosely classified as follows:

- **Lack of context:** Some tweets were written using only proper and common nouns making it hard for the annotator to understand the tweet and whether the tweet was written in English. For example, *"@<PERSON> Ronaldo"*, *"@<PERSON> @<PERSON> @<PERSON> <DRUG> #rx" or "@<PERSON> <DRUG> FTW."* Major causes of disagreement were identified specifically in short tweets, the use of acronyms, emoticons, popular names and multilingual keywords.

- **Meaningless mention:** As the tweets were extracted based on keywords that matched the drug of interest's name it was very important to read the tweet carefully to confirm that the drug itself was mentioned, especially given that some user names in Twitter can resemble the drug name, e.g. *"@Adderall_RB I'm on it"*, *"RT @Adderall_XR: SO excited for the #entouragemovie"*. Here we can see how drug names do not appear in the tweets once we remove the user names (*"@<PERSON> I'm on it!"* and *"RT @<PERSON>: SO excited for the #entouragemovie"*, respectively).

- **Identifying first-hand reports:** We found that in some cases it was not straightforward to distinguish a first-hand experience from rhetorical thought: *"I wish I could prescribe <DRUG> myself for all these depressing ass tweets cheer tf up"*, and also how to annotate the tweet in the case of forwarding a tweet from someone else (doing a Retweet): *"RT @<PERSON>: @<PERSON> @<PERSON> - Fear not! I've got a couple of bottles of #<DRUG> right here. Pass me a doughnut, plea ..."*. In other cases it was not easy to tell for sure whether the writer was actually taking the drug: *"Popular antidepressants <DRUG>, <DRUG> and <DRUG> can lower libido and prevent orgasms #fact"*. In the same way it is not straightforward to realize whether the user took the drug and stopped taking it or whether she still takes it as in the following example: *"@<PERSON> yep. i honestly think the <DRUG> has messed up my memory and concentration or something because they suck now"*, *"Hello, <DRUG>. Miss me?"*.

- **Ambiguous genre:** Another area of disagreement was when annotating *"Opinions"* and *"Other's experience"*, as in some examples it could be understood in either way as in: *"@<PERSON> go to sleep already Joe and put down the <DRUG> really shit!"*,
"@<PERSON> @<PERSON> I just found it funny that people used <DRUG> against him.", *"Jesse needs to lay off the <DRUG> lmao"*.

Crowdsourcing annotation

Although the two PhD annotators could have annotated all the tweets within the data set, and given that experienced annotators are a scarce resource, we decided to study other possibilities and rely on a crowdsourcing engine, also taking into account that the annotations obtained from the experienced annotators could be used as the gold standard when collecting laymen annotations.

We opted for CrowdFlower[3] as the service allowed us to use a subset of the tweets previously tagged by our experienced annotators, enabling us to provide a set of data items with correct responses, which in turn were used to discard tainted contributions. We also configured the settings to target contributors from several English speaking countries (Australia, Canada, New Zealand, the United Kingdom, and the United States) on the assumption that annotators from these countries were more likely to be native English speakers.

We decided that the gold standard to be used in the crowdsourcing platform would be composed of 100 tweets where both expert annotators agreed on all fields. After that selection, the annotations provided by the expert annotators were then analysed to understand the cause of disagreements observing the points presented in the previous section. These 100 gold questions became the testing questions for laymen in CrowdFlower, acting as a filter to discard all the annotations coming from any annotator scoring lower than 70% on those test questions.

The experienced annotators used the extended version of the guidelines -expert annotator guidelines- prior to annotation. These guidelines were based on those created for a study into usage of electronic tobacco products reported on social media [155]. All the categories in our study except three were also used in the electronic tobacco product study. We added two categories in order to refine the results by annotating whether the tweet was written in English, and also to focus on the drug reporting tweets. The third category we added was used to understand if the tweet was reporting a first hand experience.

Laymen annotators were presented with a simplified set of the annotation guidelines -laymen annotator guidelines- in the form of a questionnaire. A file showing a sample of annotation requested to the laymen can be found on *"B. Laymen annotator guidelines for annotating first-hand experience tweets."*.

[3] https://www.crowdflower.com/

Once we obtained the aggregated results from CrowdFlower[4] we extracted the tweets that were written in English language and mentioned drugs of interest. This yielded 899 tweets that became our gold standard[5].

Second stage annotation

Between September 26th 2014 and December 9th 2014 we collected a new data set from Twitter by filtering the tweets containing any of the drug names or drug synonyms listed in Table 3.1 and Table 3.2. We gathered 159,007 tweets and chose 4000 tweets at random to be annotated by two experts using the same version of the "Expert guidelines".

In this case, and given the experimental set-up where this data set was to be used, we only focused on the annotation of one single field, i.e. the "genre", out of all the fields that were annotated at previous stages of the annotation process, obtaining 3211 tweets where both expert annotators agreed on the annotation for the genre and which were written in English language and about the drugs of interest.

3.1.3 Resulting data

We used the "Full report" CrowdFlower provided, which contains all the annotations obtained from the contributors, to calculate the inter-annotator agreement showed in Table 3.3 (column 4, *"Fleiss' kappa for 5 raters (CrowdFlower)"*). The results were of comparable quality to the experienced annotators, although in general the crowdsourced results scored slightly lower than those obtained from experienced annotators. In the case of *"Activism"*, *"News"*, *"Marketing"* and *"Disgust"* the Kappa scores were higher than the values obtained from PhD raters. Once we obtained Fleiss' kappa results we ranked these values to calculate Spearman's Rho [156] (ρ=0.471) and Kendall's tau [157] (τ=0.352), where we observed moderate agreement [158]. This confirmed that the data we obtained from the crowdsourced annotations were of comparable quality to those obtained by expert annotators, a result consistent with previous work in the domain [159]. We observed several categories of question such as *"Cultural reference"* where the correlation values were markedly low. This is not surprising since Twitter contains many culture-specific references.

We further analysed the annotations from expert annotators to obtain Wilson score interval as suggested by [160]. Apart from calculating Wilson score interval between

[4] A modified version of this file complying with Twitter's TOS can be found on github https://github.com/nestoralvaro/JBI_Pharmacovigilance/tree/master/1548_CrowdFlower.

[5] A modified version of this file complying with Twitter's TOS can be found on github https://github.com/nestoralvaro/JBI_Pharmacovigilance/tree/master/899_CrowdFlower.

Question	Cohen's kappa for experienced raters	Fleiss' kappa for experienced raters	Fleiss' kappa for 5 raters (CrowdFlower)
Tweet written in English language?	0.962	0.962	0.943
Tweet about the drugs of interest?	0.888	0.888	0.845
First-hand experience	0.674	0.673	0.556
Other's Experience	0.391	0.390	0.231
Activism	-0.002	-0.005	0.075
Cultural reference	0.427	0.424	0.112
Humor	0.392	0.390	0.377
News	0.338	0.336	0.352
Info/resource	0.382	0.381	0.294
Marketing	0.361	0.357	0.409
Opinion	0.282	0.266	0.244
Sentiment	0.395	0.385	0.314
Pleasure	0.076	0.075	0.057
Craving	0.362	0.360	0.239
Disgust	0.045	0.044	0.129

TABLE 3.3: Inter annotator agreement between raters using Cohen's and Fleiss' Kappas.

expert annotators we also computed the percentage agreement. The results are presented in Table 3.4.

CrowdFlower provided us with the *"aggregated"* results file, which only contains the most trustworthy annotation based on individual contributors' trust ratings for every question independent of the number of judgements that were requested per question (we requested 5 judgements per tweet). The confidence score describes the level of agreement between multiple contributors (weighted by the contributors' trust scores), and indicates CrowdFlower's *"confidence"* in the validity of the result [161]. Once a job is complete, all of the judgements on a row of data are aggregated with a confidence score, and in order to provide the aggregated result CrowdFlower chooses the response with the greatest confidence [162].

In order to control quality we had to apply some validation mechanism, and we used expert annotators to gauge laymen annotators quality. Apart from the validation mechanism we also believe it is important to mention the following points:

Question	Wilson conf. Interval (min)	Wilson conf. Interval (max)	Percentage agreement
Tweet written in English language?	0.968	0.990	0.982
Tweet about the drugs of interest?	0.925	0.960	0.945
First-hand experience	0.876	0.922	0.902
Other's Experience	0.920	0.956	0.941
Activism	0.978	0.995	0.989
Cultural reference	0.945	0.974	0.962
Humor	0.876	0.922	0.902
News	0.963	0.986	0.977
Info/resource	0.907	0.946	0.929
Marketing	0.959	0.984	0.974
Opinion	0.847	0.897	0.874
Sentiment	0.850	0.900	0.877
Pleasure	0.954	0.980	0.970
Craving	0.948	0.977	0.965
Disgust	0.963	0.986	0.977

TABLE 3.4: Wilson confidence interval (minimum and maximum), and percentage agreement between 2 expert annotators.

- **Resource scarcity**: Finding expert annotators was much harder than we initially expected. This, in the end, delayed the start of the experiments.
- **Costs**: Expert annotators were much more expensive to hire than laymen annotators. Given the experimental set up this point in particular did not affect us, but we realized it could have been an issue to consider in case we would have had to annotate a large amount of tweets.
- **Time constraints**: Expert annotators can only devote a limited number of hours per day to the annotation task. On the other hand, crowdsource annotators are a potentially unlimited work force and once the task was launched in CrowdFlower platform laymen annotators worked on it at a constant rate.

After these observations we consider that both laymen and expert annotators contributed to our annotations very positively. We believe that the combination of laymen annotators, who can work on large volumes of data, and expert annotators, who can validate the annotations produced by laymen, provided a very good data set suited to our needs.

As a by-product, when curating this corpus we showed that the inter-annotator agreement from CrowdFlower is of comparable quality to the inter-annotator agreement obtained from experienced annotators, confirming that we can rely on crowdsourced annotations to identify personal drug reports, although there are still difficulties such as some notable disagreements (e.g. cultural references, disgust) that need to be recognised. To overcome this we have to analyse how human agreement might be improved as there are some open areas of work such as better guideline development and better interface selection.

The corpus we curated has been released following Twitter's TOS and it can be found at https://github.com/nestoralvaro/JBI_Pharmacovigilance/.

3.2 Data set containing PubMed sentences mentioning drugs and their related phenotypes

Besides the informal texts found in Twitter we decided to explore the contents of PubMed articles, and during the first edition of the Biomedical Linked Annotation Hackathon (BLAH) [6] we worked towards curating a subset of PubMed abstracts composed of the excerpts containing patient symptoms. Our goal was to develop an automatic annotation pipeline to curate our corpus.

The main target of the event was to:

- Curate annotations that are comparable to each other.
- Produce annotations that are searcheable across multiple data sets.
- Produce annotations that are referenceable through the dereferenceable URIs[7].

3.2.1 State of the art in linked corpora

As explained in BLAH's website[8], all important scientific discoveries have been published in the scientific literature, thus making this source of information the most important repository of scientific knowledge, and putting it at the center of data and text mining. However, the annotation of scientific texts pose an important challenge as most annotations have been done in a manual way. Automated annotation tools are starting

[6] http://2015.linkedannotation.org/

[7] A dereferenceable URI is a resource that allows the retrieval of a copy or representation of the resource it identifies, and in this context that allows a natural integration to other data mining efforts.

[8] http://1.linkedannotation.org/background

Drug Name	Geographic annotations	Literature annotations
The target is unstructured data	map image	text
The annotations identify where those entities in the data	restaurants and shops	drugs and diseases
These show how the entities are connected to each other	streets	dependency paths

TABLE 3.5: Comparison of common characteristics between geographic annotations and literature annotations.

to emerge, and even if also some researchers are sharing their annotated datasets these data sets are mostly independent of each other.

These days, crowdsourcing and other technologies for sharing data are gaining much attention. An example of this is Google Maps[9] that allows:

- Sharing data using the same coordinate system.
- The use of dereferenceable URIs for any position.
- The use of dereferenceable URIs for annotation data.
- The use of APIs and tools.

Some traits seen in geographic data are also present in literature annotations as can seen in Table 3.5

Taking these ideas into account new venues and innovative proposals for text mining are appearing. The main goal is to produce new data sets in a richer way while combining these data with heterogeneous annotations at many different levels, as can be syntactic and semantic, and for multiple data sets, such as genomic or clinical annotations to name a few.

3.2.2 Curating a linked corpus

The purpose of our annotation effort was to create a collection of sentences from the literature containing evidence about side effects (also known as adverse drug reactions or ADRs) in two classes of drugs, i.e. cognitive enhancers and antidepressant, for which we reused the list of drugs presented in [139]. Our efforts were in line with the main objective of the hackathon as we also aimed at:

- Use normalized texts from the scientific literature (i.e. PubMed and PMC articles).

[9] https://www.google.com/maps

- Annotate the data set composed of scientific texts.
- Produce dereferenceable URIs to the annotated corpora.

To reach our goal we automatically consumed Europe PubMed Central API[10] selecting out all the articles mentioning any of the drug names included in our dataset. Those articles were further filtered out by using a custom-built list that included common patient descriptors such as "baby", "women", "student" and "cohort" among others.

Once we had all the articles in place we annotated them in an automatic way using two different approaches due to the targeted entities:

- **Phenotypes**: Phenotypes are characterised by the modifiers that mark out abnormalities in anatomy, physiology and behaviour (e.g. mental states). Our phenotypes lexicon was obtained using Phenominer Database [135], and the annotations we produced were obtained by using PhenoMiner ontology via a dictionary-based tagger. Examples of phenotypes include: impaired vision, increased dyspnea, tightness of chest, high blood pressure as well as single word symptoms such as nausea, confusion and anxiety.
- **Chemicals and drugs**: These entities were identified directly using the National Center for Biomedical Ontology (NCBO) Annotator using the Chemical Entities of Biological Interest (ChEBI) ontology. Examples of chemicals and drugs include: stimulant drug, prescription amphetamines, modafinil, 800mg ibuprofin and antidementives.

Having the pipeline in place we decided to enrich the phenotype annotations by requesting NCBO annotator to identify the entities included in Phenotypic Quality Ontology (PATO).

As a validation method we selected a sample of 150 sentences and inspected them manually with the aim to verify the quality of our annotations, finding no conflict nor incorrect offset. We then released the corpus so that it can be publicly used by the research community.

Our efforts resulted in the curated corpus named NEUROSES (beNchmark litEratURe cOrpus Side Effect aSsociations). This corpus contains 8,605 articles from PubMed which include patient descriptors and mentions of drugs known to be used off label by the student community as cognitive enhancers: Adderall, Ritalin, Modafinil, Adrafinil,

[10] https://europepmc.org/RestfulWebService

Arnodafinil, and drugs which are commonly used as anti-depressants: Citalopram, Escitalopram, Paroxetine, Fluoxetine, Fluvoxamine and Sertraline, also including the known synonyms of all the mentioned drugs. All the documents in this data set contain annotations for all the phenotypes included in PATO ontology and Phenominer database, along with annotations for all the chemicals and drugs included in CheBI ontology.

The corpus is freely available online: http://pubannotation.org/projects/NEUROSES

3.3 Data set containing tweets and PubMed sentences mentioning symptoms and diseases related to the drug use

The rapid growth of social media provides an excellent resource to gather vast amounts of data to be used in experimental studies for pharmacovigilance. The number of texts from scientific publications is growing at a steady rate [163] providing valuable information produced by experts on the field. On the other hand, researchers in pharmacovigilance have recently started exploring Twitter and other non-scientific texts where patients describe diseases and symptoms [31, 87, 94].

The research using social media information is gaining traction although there are some gaps that have to be filled as recent studies have shown that the same Named Entities Recognition (NER) task suffers from a 10% loss in terms of F-Score when using Twitter data [6] against the results that are obtained when utilizing user posts from a social media forum[11]. From these results it is clear that there is a need to understand the key differences causing the loss in performance.

Comparing scientific texts and social media texts is not a task that can be directly done as the corpora that we have at this date are very different in their nature. There are a number of characteristics, as could be the underlying traits in those texts: On one hand we have social media texts, which are known for being informal and noisy [39], containing a high proportion of ungrammatical constructions, out of vocabulary words, abbreviations and metaphoric usage. On the other hand, scientific texts are known for the use of specialized vocabulary, and well formed sentences. Secondary key factors involved in a direct comparison are the topicality and the data selection methods. As pointed by other researchers text variations are due to the sub-domain itself even in close areas such as genetics and molecular biology [164], and this is why such variability should be reduced to the minimum when comparing texts from different sources.

In this section we present our efforts to develop a corpus for analysing differences in performance of NLP technologies, like NER, when tested on informal and formal media

[11] http://www.dailystrength.org/

sources. Texts in Twitter and PubMed are different in many aspects (e.g. the length of the reports, or the intended audience) and to take those aspects into account we devised a sentences sampling criteria targeting at texts from both Twitter and PubMed containing the same information (i.e. reports on the same set of drugs) understanding that those reports would differ in the linguistic register (i.e. formal and informal registers).

To date, most of the curated corpora for pharmacovigilance come from scientific formal texts obtained from PubMed [27, 106, 165], although data sets curated from other scientific resources such as Khresmoi project[12] are also available [166].

Since a few years ago, corpora obtained from social media texts started emerging. At first, researchers focused on blogs and forums [30, 88], moving then to the consumption of Twitter's data [31, 139] due to the high volume of the information it provides, with approximately 500 million tweets sent since 2013 [140][13], and also motivated by its "realtime" information, allowing health researchers to potentially investigate and identify new Adverse Drug Events (ADE) types faster than traditional methods such as physician reports.

To curate those corpora the first step was the selection of the drugs of interest, often focused on few compound and trade names. In the case of social media texts some trade names could be discarded to filter messages by countries where the drugs were sold, and even expanded by obtaining a list of possible misspellings of drug names [138] to include potential mentions.

It can be noticed that the strategies for data collection have not changed much in recent years, but the source of the information has. The nature of social media messages is very different from the one in scientific articles, and it is hard to perform a direct comparison using the data sets obtained by other researchers mainly because of the selected set of drugs included in those studies -varying from one study to another-, and also because of the length of those texts as it is clear that the 140-characters texts from Twitter can not be directly compared with the texts spanning a number of paragraphs in PubMed articles.

It was our aim to fill that gap, and create a corpus, known as TwiMed, composed of texts from two sources of information using very different writing styles (I.e. Twitter and PubMed) also keeping in mind and aiming at reducing the possible variability due to external factors.

[12] http://www.khresmoi.eu/

[13] The number seems to be stable as this is still the same number of tweets sent per day http://www.internetlivestats.com/twitter-statistics/

3.3.1 Data sampling

For our study we selected a set of 30 different drugs used in other pharmacovigilance −drug safety− studies [6, 31, 95, 99, 139].

We employed Twitter's API to download messages mentioning any of those drug names or their synonyms by running our script from September 7th 2015 to October 10th 2015, obtaining 165,489 tweets. In the case of PubMed we obtained the list of articles about those drugs by using EuropePMC RESTful Web Services[14], issuing our query on October 21st 2015 to search for texts containing the same keywords that we used when collecting tweets. Once we had the list of PubMed articles we processed them to extract the sentences containing the drug mentions, obtaining 29,435 sentences[15].

In our data filtering step we removed all non-ascii characters (e.g. emojis), replaced all user name mentions with *"__username__"*, all e-mail addresses with *"__email__"*, and all numbers with *"__number__"*. We also reduced characters elongation by removing the repetition of a character after the second occurrence (E.g. *"greeeeeeat"* would become *"greeat"*), and lowercased all sentences from PubMed and Twitter. We also aimed at maximizing the informativeness and variability of the texts by discarding the sentences shorter than 20 characters in length, the retweets, the tweets not written in English, the sentences containing keywords related to marketing campaigns[16], and sentences that included URLs. Lastly, we applied some heuristics to discard possibly duplicated sentences, and limited the number of tweets any user could contribute to five.

The data filtering step presented above was also useful to discard possibly duplicated information as from each selected sentence having more than 40 characters we extracted the substring starting on character 20 and having a maximum length of 40 characters (less in case a sentence would not have at least 60 characters). Those sub-strings were searched for in the candidate sentences, keeping only the messages not containing them. By doing so we aimed at further increase the variability of the texts filtering out similar messages.

Out of the resulting sentences we selected 6000 sentences each for both Twitter and PubMed, aiming at a balanced sample of the drug mentions for which we extracted the sentences in a round-robin fashion. However, given the differing sample frequencies of each drug, the final numbers varied showing that in both Twitter and PubMed some drugs attract more attention than others. We found that the frequency of the drugs in

[14] http://europepmc.org/RestfulWebService

[15] We extracted the sentences using the document sequencer scripts from PubAnnotation project: https://github.com/pubannotation/pubannotation/tree/master/lib

[16] Our list was built heuristically using five words commonly related to marketing campaigns: "buy", "cheap", "online", "pharmacy", "price".

Drug name	# sentences in Twitter	# sentences in PubMed
bevacizumab	69	239
buprenorphine	363	244
carbamazepine	74	239
ciprofloxacin	81	250
citalopram	331	251
cortisone	344	231
destroamphetamine sulphate	373	19
docetaxel	34	246
duloxetine	242	241
fluoxetine	344	238
fluvoxamine maleate	13	204
lamotrigine	168	242
lisdexamfetamine	348	84
lisinopril	56	147
melphalan	2	234
methylphenidate hydrochloride	349	112
modafinil	287	10
montelukast	71	239
olanzapine	190	248
paroxetine	365	249
prednisone	350	249
quetiapine	339	247
rupatadine	1	45
sertraline	343	236
tamoxifen	122	238
topiramate	133	231
trazodone	206	70
triamcinolone acetonide	14	253
venlafaxine	326	238
ziprasidone	62	226

TABLE 3.6: Total number of sentences for each drug name in Twitter and PubMed.

both sources of information had no correlation (Spearman's Rho, ρ= 0.03), as shown in Table 3.5.

Figure 3.1 shows the distribution in number of tokens in the sentences. Tweets are characterized by containing less tokens than PubMed sentences. There is no tweet

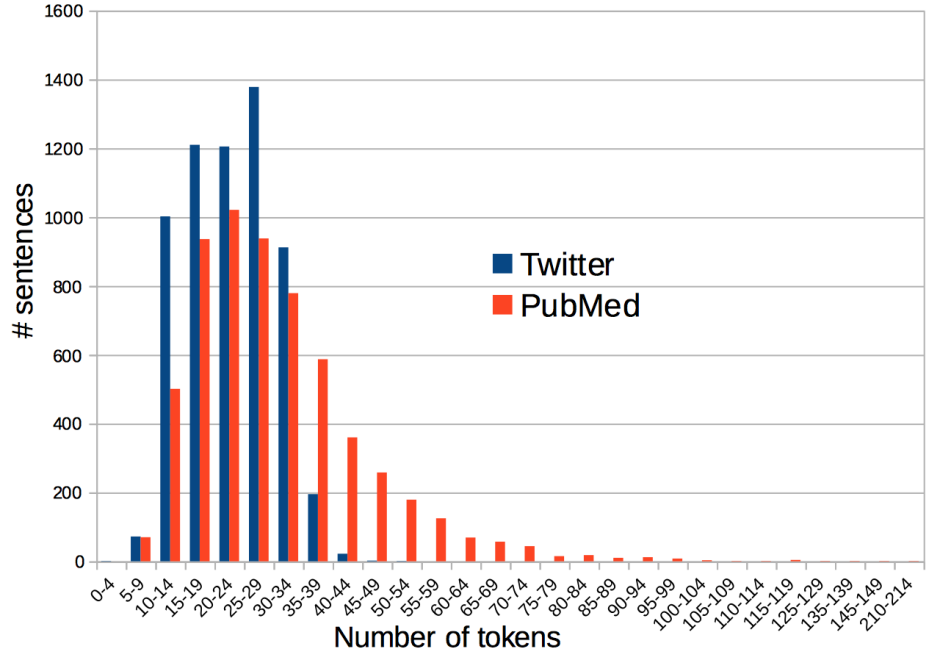

FIGURE 3.1: Number of sentences from Twitter and PubMed grouped by the number of tokens in each sentence.

having more than 54 tokens, and 1378 sentences in Twitter have 25-29 tokens. On the other hand, PubMed data set included sentences containing more than 200 tokens, leading to a 33% difference in the number of tokens between PubMed (178,892) and Twitter (134,142).

3.3.2 Selecting the annotators

We aimed at two annotators and since we found six annotators who were willing to contribute to the task we prepared a test phase to decide who would be hired for the annotation. The annotators were requested to indicate which sentences contained a drug and a disease or symptom related to the drug effects in humans, and used a list of 40 sentences (20 sentences from Twitter and 20 sentences from PubMed) out of the 6000 sentences in our data set.

The six initial annotators had different backgrounds: one of them was a native English speaker, three of them were expert pharmacists, and the last two of them were active social media users[17]. We were interested on understanding whether the background differences could play a role in the test phase as we believed that pharmacists would probably do better in PubMed texts while the rest of the annotators would probably do better on Twitter texts. To evaluate the results we used a gold set of labels that we

[17]Except for the native English speaker, the rest of the annotators were native Spanish speakers able to read English texts.

Data set	Twitter (min)	PubMed (min)	Total (min)
Social1	0.70 (9)	0.80 (10)	0.75 (19)
Social2	**1.00** (8)	0.70 (7)	0.85 (15)
Native speaker	0.85 (6)	0.50 (6)	0.67 (12)
Pharmacist1	0.90 (8)	**0.85** (7)	**0.87** (15)
Pharmacist2	0.70 (11)	0.80 (9)	0.75 (20)
Pharmacist3	0.50 (15)	0.70 (15)	0.60 (30)

TABLE 3.7: Agreement with gold data during the annotator selection phase. We compared the results from two very active social media users, one native English speaker and three pharmacists. We indicate between brackets the time it took to complete the anotation for that data set (time in minutes).

generated by using the annotations received from the six annotators and the annotations produce by the author of this thesis.

We discovered that one pharmacist scored the best result 87.5% agreement with the gold data (See Table 3.6), followed by one social media user.

Those results were in line with our expectations as social media users got the best scores in social media texts (obtaining perfect agreement in one case), and the best scores in PubMed texts were obtained by the pharmacists. However, we were very surprised by the low scores obtained by Pharmacist3 and the native English speaker. We followed up with them discovering that Pharmacists3 had some trouble understanding the samples because of those being written in English language (it was also evidenced in the time it took her to complete the task). In the case of the native English speaker he told us that he was not an active social media user, and in fact he complaint that the sentences were not well formed and hard to understand. Overall, we discovered the native English speaker was too cautious when indicating which sentences were positive cases as he annotated 7 out of the 40 sentences[18] while the rest of the annotators indicated 13-18 sentences were positive[19].

We decided to hire Pharmacists1 as she scored the best results, and out of Social1, Social2 and Pharmacist2 we decided to hire Pharmacist2 taking into account that the resulting corpus would require annotation at entity level for which Pharmacist2's skills would be decisive.

[18]The gold standard had 16 sentences tagged as positive sentences.

[19]Pharmacist3, who obtained the lowest score, was above that range as she annotated 24 sentences as positive.

3.3.3 Annotation

To curate the corpus we followed a different approach from the ones we presented in previous sections. In this case the annotation was also a manual one, same as when we annotated our "first-hand experience corpus", but we also divided the annotation task in two very different phases in order to optimize the annotation effort:

- **First annotation phase**: In this phase we focused on identifying the sentences that were mentioning a drug and a symptom or disease related to the drug intake.
- **Second annotation phase**: In this phase we requested the annotators to identify the drug, symptoms and diseases annotating their spans, and also include in the annotation a number of attributes for the tokens along with the relations between the drugs and the symptoms and diseases.

The reason for performing the annotation using the two exposed phases was to provide the annotators with the sentences containing the entities of interest, i.e. the sentences filtered during the first annotation phase, as we noticed that in a number of sentences no symptom nor disease were mentioned.

By taking this approach filtering out the sentences of interest first, and then requesting the annotators to annotate the selected subset of sentences we expected to obtain the desired number of sentences with entity level annotations in a more efficient way.

Annotation guidelines for annotating sentences

Following, we present the annotation guidelines we provided to the annotators for the first phase. Although the task was to identify the sentences containing drug and symptoms or diseases we had to clarify the following points:

- The annotation has to be "1" (the number one) in the sentences containing a drug name and also the name of some symptom or disease related to the drug mentioned in the sentence.
- In case the sentence does not mention any medicine, symptoms nor diseases the annotation for that sentence will be left blank.
- We are using a closed list of medicines that will be provided and can be used to check whether the drug appearing in the sentence is one of the drugs of interest.

- All the sentences provided contain the name of some medicine, although it could be case that the name of the medicine matches the name of some other element not related to such medicine resulting in a sentence not talking about the medicine. In case the sentence is not referring to the medicine the annotation for the sentence is left blank.

- The symptoms can appear before taking the medicine (e.g. causing the intake of the medicine), or after taking the medicine.

- The symptoms appearing after taking the medicine can be indications (the patient's condition improves), or other effects worsening the patient's condition (e.g. allergic reactions).

- When the annotator is not sure about whether a concept is a symptom he or she has to check if it appears on this website http://purl.bioontology.org/ontology/MEDDRA (search for it using the field "Jump to"). If the concept appears listed it means that it is a valid symptom.

- Some concepts can appear in an explanatory way. One example of this would be "I could not sleep during the whole night", which would correspond to "insomnia".

- It is important not to rely on guesses/conjectures nor assumptions.

- The symptoms should not be inferred, as these should be clearly identifiable in the given sentence. This means that in case we are annotating the sentence "*I just took an antidepressant*", the keyword *antidepressant* should not be considered as a symptom given that such keyword refers to a medicine and not to the symptom itself. Such keyword does not indicate whether the patient is suffering from *depression*. On the other hand, in the sentence "*I took my medicine for the depression*" does contain a symptom that is clearly identifiable (the word *depression*). Lastly, the sentence "*I took an antidepressant and I am feeling better now*" does contain the mention to a symptom *feeling better*. It is very important to understand the differences between these three examples.

- The study is focused on studying the effect of the medicines in humans. In case the sentence is not referring to the intake of the medicine by humans, or the symptoms are not reported for humans that sentence should not be annotated with a "1" (the annotation for that sentences will be left blank). It is of particular importance to be careful in sentences that refer to murines such as rats or mice.

- In case the annotator is not clear on how to annotate the sentence he or she can use an "X" to indicate his or her uncertainty in the annotation.

Annotation guidelines for annotating entities

To create the guidelines we followed an iterative approach as we read guidelines developed by other researchers working on the pharmacovigilance area. Namely, we took into account the guidelines used to curate the ADE-CORPUS [28], the guidelines for the meta-knowledge annotation of bio-events [167], the Arizona disease corpus annotation guidelines [20], the guidelines for the annotation of disorders in clinical notes used in ShARe/CLEF eHealth 2013 shared task [21], and the annotation guidelines for DDI corpus [106].

By using those documents as a starting point we prepared the first draft of our guidelines and had three external annotators with a background in computer science annotating PubMed and Twitter sentences using our guidelines. During that time we had daily meetings after each annotation session and refined the guidelines upon the discrepancies we found and the questions raised by the annotators. We used that feedback to provide the annotators with an updated version of the guidelines for the next annotation session. After two weeks and six annotation sessions the number of discrepancies was reduced to a minimum and no more questions were raised, leading us to agree on freezing the guidelines so that these would be used as they were.

Another supporting documents we also reviewed were CRAFT concept annotation guidelines [168], the guidelines used to annotate the EU-ADR corpus [108], the guidelines used to curate GENIA corpus [169], the annotation guideline manual for extracting adverse drug event information from clinical narratives in electrical medical records [170], the ADR expert annotation guidelines for Twitter messages [139], and THYME Annotation Guidelines [171]. Those were of particular interest to understand the approaches some researchers took in certain matters, such as deciding whether to allow relations crossing sentences (which we allowed following the approach of [170] and [171]).

After reading those guidelines, we decided to include three different entities in our study:

1. **Drug**: Any of the marketed medicines that appears in the SIDER database[22], which is also listed in the closed set of drugs we provided to the annotators.
2. **Symptom**: Any sign or symptom contained in MedDRA[23] ontology.
3. **Disease**: Any disease contained in MedDRA ontology.

The complete list of the attributes we allowed for those entities is the following:

[20] http://diego.asu.edu/downloads/AZDCAnnotationGuidelines_v013.pdf
[21] http://blulab.chpc.utah.edu/sites/default/files/ShARe_Guidelines_CLEF_2013.pdf
[22] http://sideeffects.embl.de/
[23] http://bioportal.bioontology.org/ontologies/MEDDRA

1. **Polarity**: Used to indicate whether the entity was negated or not. The negation had to be a linguistic negation ("not", "don't"...).

2. **Person**: Used to indicate whether the entity was affecting the "1st", "2nd", "3rd" person, or whether there was no information. This attribute was based on the original sender.

3. **Modality**: Used to indicate whether the entity was stated in an "actual", "hedged", "hypothetical" or "generic" way.

4. **Exemplification**: Used to indicate whether the entity was presented using an example or a description. This attribute was only to be used when the entity was presented through an exemplification.

5. **Duration**: Used to indicate whether the entity's lasting span was "Intermittent", "Regular", "Irregular". In the case of *drugs* this attribute referred to the time span when the *drug* had been taken.

6. **Severity**: Used to indicate whether the seriousness of an entity was "Mild", "Severe" or not indicated. This was the only attribute that did not apply to *drugs*.

7. **Status**: Used to indicate whether the duration of the entity was "Complete", "Continuing" or not indicated. In the case of *drugs* this attribute referred to the time span when the *drug* was perceived as having effect.

8. **Sentiment**: Used to indicate whether the entity was perceived as "positive", "negative" or "neutral".

9. **Entity identifier**: Used to indicate the concept unique identifier (CUI) for that entity. This was the only attribute that had to be filled for all annotated entities. For this attribute we provided a list of allowed values, and used the value "-1" (Not found) for entities whose CUI would not be present in the list.

Our list of attributes was decided based on elements commonly used when curating other pharmacovigilance corpora from formal texts (e.g. "duration" or "modality"), and also when preparing corpora composed of informal texts (e.g. "polarity" or "sentiment").

We also requested the annotators to mark three possible relations between the existing entities:

1. **Reason-to-use**: Used to represent the relation appearing when a *symptom* or *disease* lead to the use of some *drug*.

2. **Outcome-positive**: Used to represent the relation between a *drug*, and an expected or unexpected *symptom* or *disease* appearing after the *drug* consumption. The outcome had to be positive.

3. **Outcome-negative**: Used to represent the relation between a *drug*, and an expected or unexpected *symptom* or *disease* appearing after the *drug* consumption. The outcome had to be negative.

The file containing the full set of guidelines we provided to the annotators can be found on *"C. Guidelines for Drug-Disease-Symptom annotation in Twitter and PubMed texts."*.

Preprocessing annotation data

The pre-processing strategy we applied to the sentences prior to showing them to the annotators slightly differ with the one used in the data filtering step. First of all, we replaced the emojis with a string describing each character, and discarded any other non-ascii character. We also did not lower case all sentences as we thought that would ease the annotator's task to detect some sentiments and disambiguate acronyms. Apart from those two points we applied the same pre-processing strategy presented before.

Entity level annotation

We were interested in annotating sentences at entity level, and in order to identify which sentences would be of interest we divided the main task in two phases. During the first phase both annotators were requested to identify which were the *positive sentences* (sentences including drug mentions, and symptoms and diseases related to the drug effects in humans). During the second phase the annotators would annotate the entities and the relations between entities present in the sentences identified during the first phase.

We fixed in 1000 the target number of sentences to obtain for both Twitter and PubMed, and we requested the annotators to annotate the first 3000 sentences from Twitter, in the 6000-sentences data set, aiming at obtaining 1000 positive tweets that both annotators agreed to label as *positive*, but that number did not provide 1000 sentences so we asked them to annotate another 200 sentences in Twitter, obtaining 1004 *positive sentences*, 31.37%, where both annotators agreed. As the number of *positive* tweets was very close to the target number of 1000 sentences, in the case of PubMed we asked them to annotate 3300 sentences in total −from the 6000 PubMed sentences we obtained before−

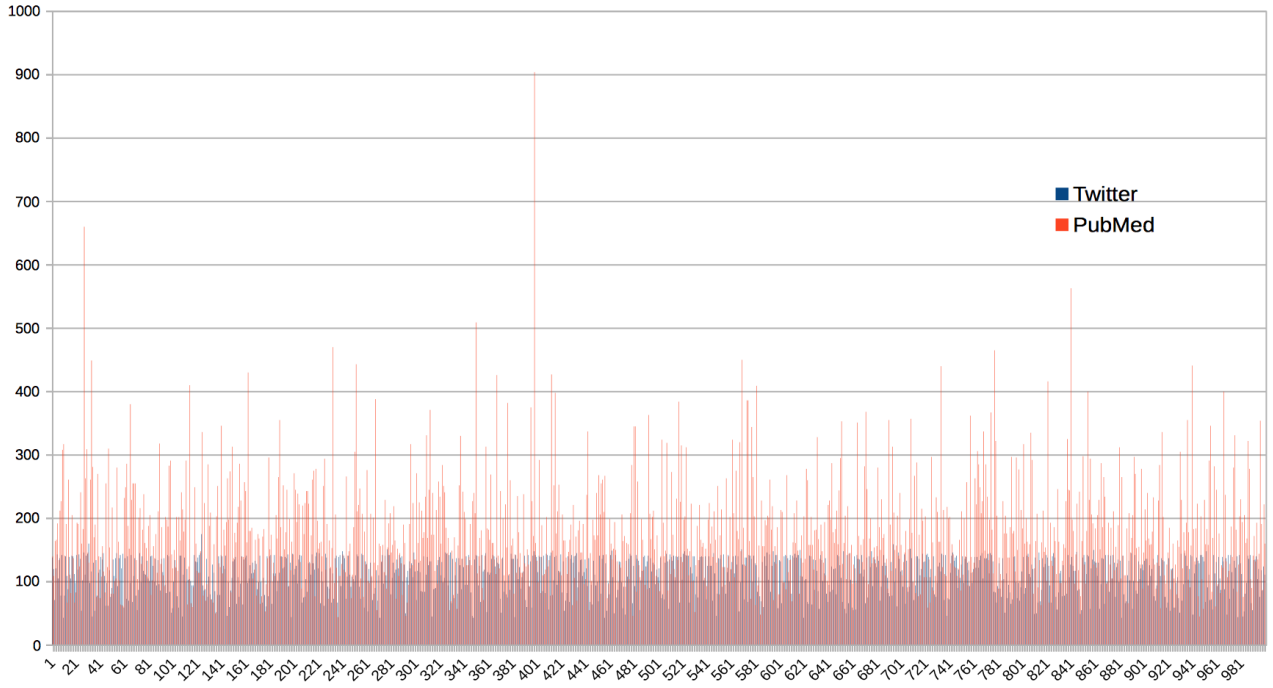

FIGURE 3.2: Number of sentences from Twitter and PubMed in our 1000-sentences sample showing the number of characters per sentence.

and got 1038 *positive sentences*, 31.45%, from PubMed where both annotators agreed. Observing that in both cases the same ratio of messages, 31%, were of interest.

Once we had the 1000 positive sentences from PubMed and Twitter we provided the annotators with the guidelines to be used during the annotation process. Those guidelines are described in detail in *"Annotation guidelines for annotating entities"* section.

After having the data we visualized the distribution of the length of the sentences in number of characters in Figure 3.2, and in number of tokens (see Figure 3.3).

3.3.4 Results

To compare the annotations produced by the experts we focused on both the *type* assigned to the entity (i.e. *disease*, *drug* or *symptom*), and also on the offsets for that entity. Taking that into account we decided to compute the results when using ***relaxed constraints*** and ***strict constraints***. In the case of using ***relaxed constraints*** we say that the entity annotated by both annotators is a *match* if the *type* for the entity matches between annotations and the spans of those annotations have some overlap. In the case of using ***strict constraints*** the match would happen if the *type* in both annotations matches and the spans for the annotated entities have the same offsets. Discontinuous annotations were allowed and taken into account when computing the matches, which

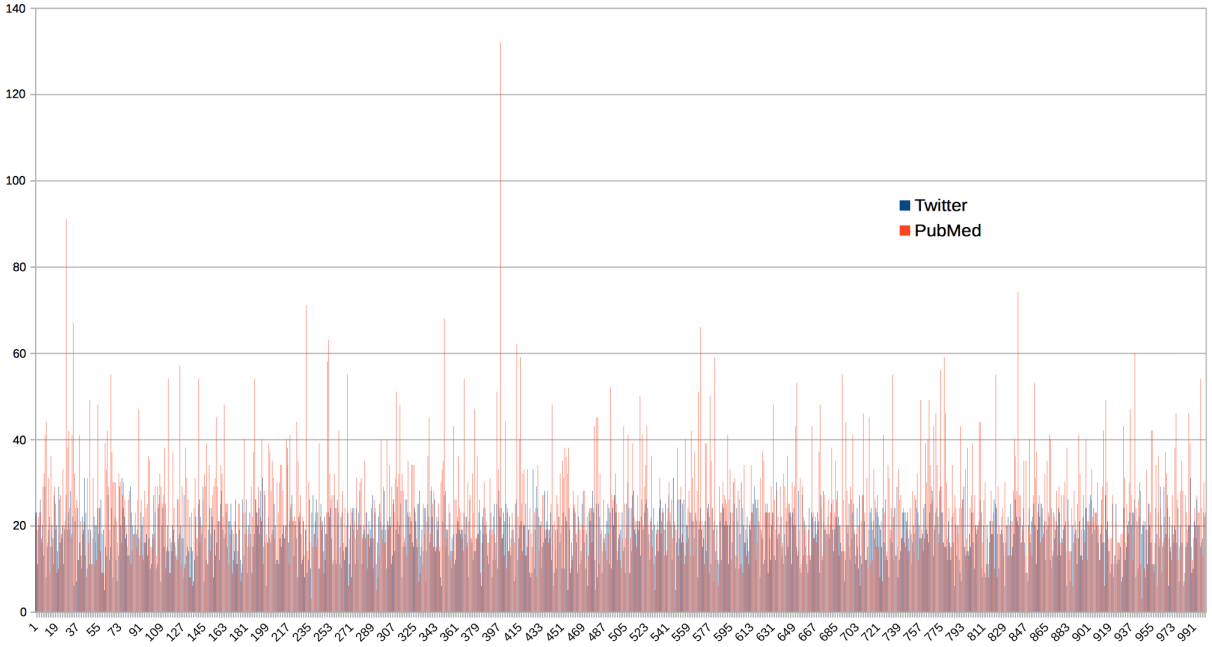

FIGURE 3.3: Number of sentences from Twitter and PubMed in our 1000-sentences sample showing the number of tokens per sentence.

means that in case of using ***strict constraints*** all the spans taking part on the entity's annotation should be the same.

We measure the level of agreement between the annotations produced by our experts using the inter annotator agreement (IAA) metric used by [172]:

$$IAA = matches/(matches + non\_matches)$$

In our case ***matches*** accounts for the total number of token matches for which both annotators agreed, and ***matches + non_matches*** counts all annotations performed by the annotator being evaluated. The results for Twitter and PubMed are shown in Table 3.7 and Table 3.8.

By focusing on the average results shown in Table 3.7 and Table 3.8 we see that the agreement for the *drugs* in both sources of information is **almost perfect**, according to [173][24]. The average results for *diseases* and *symptoms* show **substantial** agreement in both PubMed and Twitter.

When comparing the relations we can see **moderate** agreement for *outcome-negative* in both Twitter and PubMed, and also for *reason-to-use* in Twitter with low **substantial** agreement in PubMed. The agreement for *outcome-positive* relations was lower scoring **fair** agreement in Twitter, and **moderate** in PubMed. When analysing the number

[24]This classification categorizes 0-20% as *slight agreement*, 20-40% as *fair agreement*, 40-60% as *moderate agreement*, 60-80% as *substantial agreement*, and 80-100% as *perfect agreement*.

	Ph1 (Rlx)	Ph2 (Rlx)	Ph1 (Strc)	Ph2 (Strc)	Avg	#Ph1	#Ph2	#Matches (Rlx)	#Matches (Strc)
Drug	97.39	98.72	93.52	94.80	96.11	1111	1096	1082	1039
Disease	50.86	91.47	46.12	82.95	67.85	464	258	236	214
Symptom	77.23	76.71	54.21	53.84	65.50	1164	1172	899	631
Outcome-negative	63.27	75.19	43.02	51.12	58.15	795	669	503	342
Outcome-positive	11.01	40.00	8.26	30.00	22.32	109	30	12	9
Reason-to-use	55.82	60.18	44.66	48.14	52.20	842	781	470	376
Duration	46.37	8.96	39.11	7.56	25.50	248	1283	115	97
Exemplification	10.11	64.77	3.37	21.59	24.96	564	88	57	19
Modality	56.92	30.58	49.57	26.63	40.93	585	1089	333	290
Person	72.56	58.55	60.21	48.58	59.97	1709	2118	1240	1029
Polarity	76.06	52.43	53.52	36.89	54.72	71	103	54	38
Sentiment	72.48	19.46	60.92	16.36	42.30	476	1773	345	290
Severity	64.18	19.59	44.03	13.44	35.31	134	439	86	59
Status	59.41	22.07	45.94	17.07	36.12	542	1459	322	249

TABLE 3.8: Detail of annotations in Twitter. The first column shows the element being evaluated. Columns 2-5 show the Inter annotator agreement scores of Pharmacist 1 (Ph1) and Pharmacist 2 (Ph2) using relaxed (Rlx) and strict (Strc) constraints, with the average of these results in column 6. Columns 7 and 8 show the number of elements annotated by each pharmacist. Columns 9 and 10 show the number of matching elements between pharmacist's annotations using relaxed (Rlx) and strict (Strc) constraints.

of annotations it is clear that the use of *outcome-positive* relation varied considerably between annotators, and that contributed to the low inter annotator agreement score.

When comparing attributes we realized that *person* is the only one obtaining **substantial** agreement (in PubMed), closely followed by *modality* in both PubMed and Twitter, and *polarity*, and *sentiment* in Twitter (scoring **moderate** agreement). The rest of the attributes scored **fair** agreement and it was evident that the attribute *exemplification* in both PubMed and Twitter, *sentiment* and *Polarity* in PubMed, and *duration* in Twitter were very prone to disagreements as these scores were the lowest in Table 3.7 and Table 3.8.

3.3.5 Discussion

We observed that *drug* mentions had very good inter annotator agreement score, probably because *drugs* were not easily mistaken by *diseases* nor *symptoms*. On the other hand *diseases* and *symptoms* scores were significantly lower, probably due to the fact that our reference does not clearly disambiguate between those terms. One of such examples is the sentence "*Further randomized control trials are required to asses the full benefits of* ***Montelukast*** *therapy in the whole spectrum of eosinophilic* ***gastrointestinal disorders***" where both annotators agreed on annotating ***Montelukast*** as a *drug*

	Ph1 (Rlx)	Ph2 (Rlx)	Ph1 (Strc)	Ph2 (Strc)	Avg	#Ph1	#Ph2	#Matches (Rlx)	#Matches (Strc)
Drug	95.20	97.90	86.23	88.67	92.00	1271	1236	1210	1096
Disease	64.18	95.22	53.41	79.23	73.01	1086	732	697	580
Symptom	85.13	60.59	70.61	50.26	66.64	558	784	475	394
Outcome-negative	60.97	64.86	50.35	53.56	57.44	433	407	264	218
Outcome-positive	56.25	32.73	43.75	25.45	39.55	32	55	18	14
Reason-to-use	62.87	77.39	47.10	57.98	61.33	1535	1247	965	723
Duration	52.17	9.38	48.70	8.75	29.75	115	640	60	56
Exemplification	0.64	50.00	0.32	25.00	18.99	311	4	2	1
Modality	74.23	50.52	64.60	43.96	58.33	1370	2013	1017	885
Person	63.93	77.18	56.08	67.70	66.22	1439	1192	920	807
Polarity	25.00	22.22	25.00	22.22	23.61	16	18	4	4
Sentiment	33.33	1.96	22.22	1.31	14.71	9	153	3	2
Severity	42.22	33.33	37.78	29.82	35.79	45	57	19	17
Status	53.85	2.52	53.85	2.52	28.18	26	555	14	14

TABLE 3.9: Detail of annotations in PubMed. The first column shows the element being evaluated. Columns 2-5 show the Inter annotator agreement scores of Pharmacist 1 (Ph1) and Pharmacist 2 (Ph2) using relaxed (Rlx) and strict (Strc) constraints, with the average of these results in column 6. Columns 7 and 8 show the number of elements annotated by each pharmacist. Columns 9 and 10 show the number of matching elements between pharmacist's annotations using relaxed (Rlx) and strict (Strc) constraints.

and the relation between the *drug* and the concept ***gastrointestinal disorders*** as a *Reason-to-use* relation, but when annotating the concept ***gastrointestinal disorders*** one of them classified it as a *disease* while the other annotator marked it as a *symptom*, although the CUI for the concept was the same in both cases.

In the case of the relation *outcome-positive* discrepancies happened often. We realized that in most cases the relation was correctly marked, and the disagreement happened when one annotator used *reason-to-use* relation, as in "*It's proven* ***MODAFINIL*** *actually improves* ***memory*** *and creativity!!*". We believe the length of those sentences and the lack of context made it hard for the annotators to characterize the existing relation, thus causing the disagreement.

Given the similarities between those concepts and the disagreements that we detected we evaluated the inter annotator agreement score when conflating the concepts *disease* and *symptom* under *Disease/Symptom* concept. We also grouped together *outcome-positive* and *reason-to-use* relations under *Benefit* relation.

These decisions were motivated by the fact that those elements needed further clarification while the annotation process took place, and also because after a manual inspection of those entities and relations we observed that some of them could be conflated as explained.

In particular, the rationale behind the conflation of *Symptoms* and *Diseases* was motivated by the explanation provided in the guidelines, as we stated that these entities were to be annotated in case they were contained in MedDRA ontology[25], and fell back to the annotators' expertise to tell apart symptoms and diseases. Even if they were experienced in the field there were cases where those entities could not be easily distinguished:

Taking for example the sentences: *"Is steroid induced **psychosis** a thing? (Like short term prednisone tx)"*. We observed that one annotator marked ***psychosis*** as a symptom while the other annotated it as a disease, which is in line with studies in psychiatry that clarify that it can be considered to be considered as either a symptom (of a psychiatric disorders such as delirium) or a disease [174].

We also found similar conflicts in PubMed in sentences such as: "*Further randomized control trials are required to asses the full benefits of Montelukast therapy in the whole spectrum of eosinophilic **gastrointestinal disorders***.", and in the following excerpt extracted from another sentence: "*The present case report of topiramate's effect on **comorbid obesity***", among others. The literature [175] clarifies that a ***comorbid obesity*** can be a disease with its own symptoms (e.g. pressure in the chest, or fast heart rate), or be a symptom as patients with ***comorbid obesity*** are at an increased risk of insulin resistance and certain cancers. Similarly, a ***gastrointestinal disorder*** can be a disease by itself or the symptom of another disease (e.g. gastric cancer) .

As for the relations, when we decided to conflate *outcome-positive* and *reason-to-use* we did so to maximize the coverage for the included relations based on two empirical observations. The first observation is that the count for *outcome-positive* relations was much lower than the number of other relations, and in a number of cases such as in the sentences "***fluvoxamine** might improve **fatigue***", or "*How about **trazodone**, so I can just feel a little funny and then knock out and have the best **sleep** of my life*" that type of relation could be easily confounded with *reason-to-use* relation. The second reason was the fact that for both types of relation there was a beneficial link between the drug and the related symptom or disease, and in some cases the link was not clearly identifiable in the text although the type of the (beneficial) relation was clear.

On follow-up meetings with the annotators we also found that the outcomes appearing after the intake of the drug, which were to be annotated as *outcome-positive* relations, were annotated as *reason-to-use* relation because the annotators knew beforehand that those were known and expected effects related to the intake of the medicine. That showed that the descriptions for those relations should be clarified in the future by following this approach:

[25] http://bioportal.bioontology.org/ontologies/MEDDRA

	Ph1 (Rlx)	Ph2 (Rlx)	Ph1 (Strc)	Ph2 (Strc)	Avg	#Ph1	#Ph2	#Matches (Rlx)	#Matches (Strc)
Drug	97.39	98.72	93.52	94.80	96.11	1111	1096	1082	1039
Disease/Symptom	82.25	93.64	61.36	69.86	76.78	1628	1430	1339	999
Outcome-negative	67.30	79.97	46.29	55.01	62.14	795	669	535	368
Benefit	68.14	79.90	52.37	61.41	65.45	951	811	648	498
Duration	50.00	9.66	41.94	8.11	27.43	248	1283	124	104
Exemplification	10.11	64.77	3.37	21.59	24.96	564	88	57	19
Modality	64.44	34.62	54.53	29.29	45.72	585	1089	377	319
Person	77.30	62.37	63.96	51.61	63.81	1709	2118	1321	1093
Polarity	80.28	55.34	57.75	39.81	58.29	71	103	57	41
Sentiment	75.00	20.14	62.61	16.81	43.64	476	1773	357	298
Severity	67.16	20.50	47.01	14.35	37.26	134	439	90	63
Status	61.81	22.96	48.15	17.89	37.70	542	1459	335	261

TABLE 3.10: Detail of annotations in Twitter using conflated categories. The first column shows the element being evaluated. Columns 2-5 show the Inter annotator agreement scores of Pharmacist 1 (Ph1) and Pharmacist 2 (Ph2) using relaxed (Rlx) and strict (Strc) constraints, with the average of these results in column 6. Columns 7 and 8 show the number of elements annotated by each pharmacist. Columns 9 and 10 show the number of matching elements between pharmacist's annotations using relaxed (Rlx) and strict (Strc) constraints.

- *Reason-to-use* relation: If the symptom or disease is contained in the technical description (fact-sheet) of the medicine, or in case the symptom or disease appears before taking the medicine.
- *Outcome-positive* relation: Used to annotate any beneficial effect appearing after taking the medicine.

The use of those conflated categories produced a noticeable improvement in the inter annotator agreement scores. *Disease/Symptom* category obtained 76.78% agreement in Twitter and 86.57% in PubMed, while the newly added *Benefit* relation scored 65.45% agreement in Twitter, and 73.97% agreement in PubMed. This strategy also improved the agreement scores for most of the attributes. The results for Twitter and PubMed are shown in Table 3.9 and Table 3.10 respectively.

One last source of confusion was brought by the use of acronyms in scientific texts. The reason for that was that we extracted the sentences in an automated way, and scientific papers tend to disambiguate acronyms the first time they appear (usually at the beginning of the manuscripts when the term is introduced). In most cases the sentences we extracted did not include such disambiguations, and the annotators inferred what they considered it was the most likely concept for the term being annotated.

TwiMed corpus can be found at https://github.com/nestoralvaro/TwiMed/.

	Ph1 (Rlx)	Ph2 (Rlx)	Ph1 (Strc)	Ph2 (Strc)	Avg	#Ph1	#Ph2	#Matches (Rlx)	#Matches (Strc)
Drug	95.20	97.90	86.23	88.67	92.00	1271	1236	1210	1096
Disease/Symptom	91.91	99.67	74.21	80.47	86.57	1644	1516	1511	1220
Outcome-negative	81.52	86.73	65.82	70.02	76.03	433	407	353	285
Benefit	77.41	93.16	56.86	68.43	73.97	1567	1302	1213	891
Duration	53.91	9.69	50.43	9.06	30.77	115	640	62	58
Exemplification	0.64	50.00	0.32	25.00	18.99	311	4	2	1
Modality	83.43	56.78	71.39	48.58	65.05	1370	2013	1143	978
Person	71.58	86.41	62.13	75.00	73.78	1439	1192	1030	894
Polarity	43.75	38.89	43.75	38.89	41.32	16	18	7	7
Sentiment	33.33	1.96	22.22	1.31	14.71	9	153	3	2
Severity	53.33	42.11	46.67	36.84	44.74	45	57	24	21
Status	53.85	2.52	53.85	2.52	28.18	26	555	14	14

TABLE 3.11: Detail of annotations in PubMed using conflated categories. The first column shows the element being evaluated. Columns 2-5 show the Inter annotator agreement scores of Pharmacist 1 (Ph1) and Pharmacist 2 (Ph2) using relaxed (Rlx) and strict (Strc) constraints, with the average of these results in column 6. Columns 7 and 8 show the number of elements annotated by each pharmacist. Columns 9 and 10 show the number of matching elements between pharmacist's annotations using relaxed (Rlx) and strict (Strc) constraints.

Chapter 4

Comparing the linguistic register and the type of information contained in formal and informal drug use reports

Our motivation in this chapter is to understand the differences in drug reports using formal and informal linguistic registers. For that we explore messages from Twitter and PubMed and answer our first three hypotheses by studying the information conveyed in the drug use reports as well as the differences in their linguistic register.

Hypothesis 1 focuses on the information that is being conveyed in Twitter and PubMed sentences. Our understanding is that even if the reports do not contain the same information on the outcomes related to the drug use, a number of those outcomes are expected to appear in both Twitter and PubMed.

Following with Hypothesis 2 we assess if generic tweets and tweets reporting drug use, expected to be more formal given their sensitive contents, use the same set of register-related features or if the linguistic framework we use to assess the differences in the use of the register shows them as being clearly different.

The last hypothesis we explore in this chapter, Hypothesis 3, assesses whether the linguistic framework that we used to evaluate the differences in drug use reports from Twitter and generic tweets can completely describe the differences in the linguistic register in PubMed and Twitter sentences containing drug use reports.

4.1 Assessing the similarity in the information

As a first step towards understanding differences in drug use reports obtained from Twitter and PubMed we focus on the similarity of the information. Prior to that we also assess the similarity of the contents reported in our set of Twitter messages and in our set of PubMed messages reporting drug use. Doing so allows us to understand how the contents vary across our two data sets. Following, we proceeded to assess the similarity of the information focusing on the reported relations between the drugs and their outcomes.

4.1.1 Topical similarity

To assess the contents being discussed in our data sets we extracted the main topics from either PubMed and Twitter. To explore the topics we used Latent Dirichlet Allocation (LDA) [176] via a python implementation [177] and used the elbow method [178] to discover that the number of clusters, i.e. topics, in our data set was five. To better understand the topics we did not only chose five but a number of topics ranging from one to thirty. To characterize each one of those thirty topics we extracted a number of terms also ranging from one to thirty keywords for each topic, obtaining 900 different topical configurations, i.e. set of keywords, that we used to describe our data sets.

When having each topical configuration we queried the English version of the Wikipedia and extracted the top ten most likely pages matching that set of keywords so that the resulting Wikipedia pages would be used as the labels for that topical configuration following [179]'s approach.

We then grouped the set of labels by the number of topics used to extract the keywords, and proceeded to compare the labels obtained for Twitter and PubMed using Jaccard's similarity coefficient[1] discovering that the obtained set of labels (i.e. Wikipedia pages) became more similar as we increased the number of topics. This finding is probably due to the fact that a better granularity provides more different elements where both sets could be overlapping as can be seen in Figure 4.1.

We also computed the similarity when the number of keywords for each topic was treated in a separated way obtaining Figure 4.2 where we can see that in almost all cases, and independently of the number of topics, using one or two keywords to characterize the topics was not enough as no similarity was found.

[1] $J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A|+|B|-|A \cap B|}$. Where A and B represent the set of labels for Twitter and PubMed

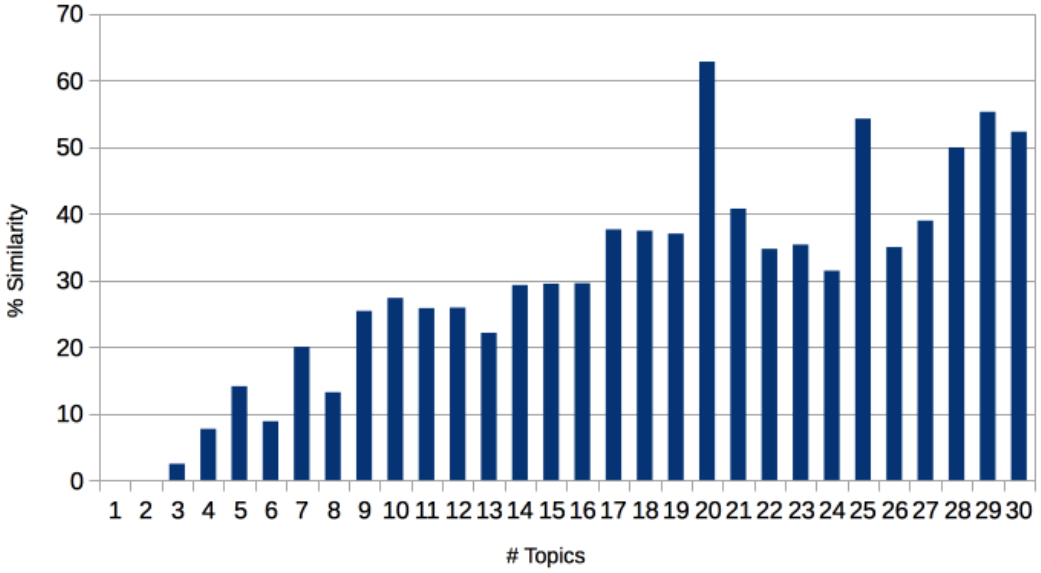

FIGURE 4.1: Results showing the labels' similarity when using 30 topics extracted from PubMed and Twitter. These results show the aggregated similarity results when performing 30 different queries (using 1..30 keywords) to retrieve the labels.

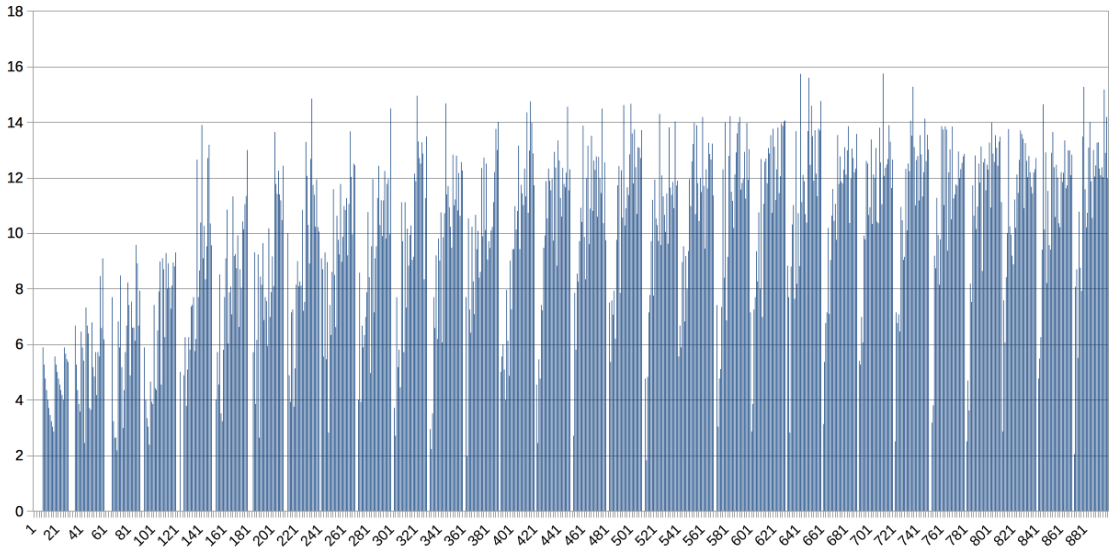

FIGURE 4.2: Results showing the labels' similarity when using 30 topics extracted from PubMed and Twitter. Using one keyword to characterize the topics often got 0% similarity making it easy to visualize the 30 different topics we extracted.

Even if we observed the same trend in which the similarity increased in both Figure 4.1 and Figure 4.2, the results when using five topics were not as high as expected as these were below 15% in both figures.

Given our sampling strategy, the closed set of drugs, and the filtering process performed by the two pharmacists we believed that there should be another key factor playing a role in the difference in the topicality, which in the end could help in understanding how to improve the performance of current NLP systems in case that factor could be controlled.

To nail down the unknown factor we decided to focus in the individual terms obtained when extracting ten keywords for 5 topics (Jaccard similarity=5.08%), 10 topics (Jaccard similarity=7.86%) and 20 topics (Jaccard similarity=13.98%). By using those keywords we aimed at understanding which were the terms appearing in those configurations and identify which terms were frequently appearing as these would be useful in describing the data sets.

We obtained the intersection of PubMed keywords in these three configurations finding 21 different terms that appeared when labelling PubMed topics. In the case of Twitter we obtained 30 terms appearing when looking for the labels for those three topical configurations. Those terms were filtered to discard terms appearing in both sets of elements, finally obtaining 19 terms for PubMed and 28 terms in Twitter after removing the terms *"cancer"*and *"effects"* from both sets.

As a way to study the list of terms we used Wordnet [180] to obtain the first level hypernym of each term using the lemma of the word being evaluated. Using the hypernyms to categorize terms is an standard approach [181, 182], and we decided to use hypernyms to group terms together to obtain a semantics-based hierarchical clustering as shown in Table 4.1.

The first thing we can notice is the category for which there are no hypernyms, including mostly drug names. As this was an automated approach we can see that there are some misses such as the word "acute", probably referring to "sharp" or "severe" although the only hypernyms for "acute" in Wordnet refer to the accent, or "still", which probably refers to "motionless" or "stationary" but the first hypernym provided by Wordnet refers to photography.

Besides those minor mismatches we can see that in general there is no overlap in terms of categories in Table 4.1, and only three of them ("case", "corticosteroid" and "physical condition") appear in both data sets. Another interesting finding is the fact that we have twelve Wordnet categories for PubMed and twenty-two categories for Twitter.

Hypernym	Twitter	PubMed
-	cymbalta, im, really, topamax, vyvanse	clinical, docetaxel, intravitreal, olanzapine, quetiapine, topiramate
accent		acute
act	took	
action	taking	
activity	help	
agent	drug	
antagonist	tamoxifen	
anti-inflammatory	prednisone	
attempt		trial
awareness	feel	
care		treatment
case	time	patient, patients
collection		combination
corticosteroid	cortisone	triamcinolone
digit	2	
effectiveness		efficacy
examination		study
external_body_part	breast	
happening		case
income	take	
interact		treated
kind	like, makes	
knowing	know	
nontricyclic	trazodone	
phenomenon		effect
photograph	still	
physical_condition	sleep	disorder
physical_property	weight	
region	side	
return	get	
symptom	pain	
time_unit	day	

TABLE 4.1: Hypernym categories assigned to the keywords used to produce PubMed and Twitter labels.

These findings evidence that the range of words in Twitter varies to a much greater extent and it can be a possible cause for the systems low performance.

Another important finding is that almost all PubMed's keywords are very closely related to the medical domain (e.g. "treatment", "efficacy" or "disorder"), whereas for Twitter most of the keywords are generic terms that can be used in a number of situations. In the case of Twitter it seems that only the drug names are the only terms that are related to medicine. When studying Twitter keywords we can see a very high number of

	# PubMed	#Unique PubMed	# Twitter	#Unique Twitter	# Inter-section	# Union	% similarity
Disease	562	89	144	41	27	103	26.21
Drug	1140	28	1047	30	28	30	93.33
Symptom	364	160	546	138	46	252	18.25

TABLE 4.2: Similarity between entities in Twitter and PubMed using the non-conflated annotations. Columns 2 and 4 show the total number of annotated tokens. Columns 3 and 5 show the number of unique tokens. Columns 6 and 7 show the intersection and the union, respectively, for the unique tokens (as shown in columns 3 and 5). The last column shows the similarity coefficient.

generic-use verbs ("taking", "like", "know", "get") that have potential for introducing noise when using Wikipedia to label the topics.

We can see that besides the drugs being discussed in the category for which no hypernym was found, i.e. the first category, this category contains an acronym ("im") and an adverb ("really") in Twitter's keywords and two adjectives that are seldom used outside medical texts ("clinical" and "intravitreal") for PubMed.

These results helped us in noticing that there are some generic words extracted when obtaining LDA topics for Twitter that could be introducing some noise in the query contributing to the dissimilarity between the data sets' labels.

4.1.2 Similarity of the information

Regardless the type of the register formality used to express the contents, and having the annotations in place we decided to perform one more comparison to test the similarity of the contents in our sample of Twitter and PubMed sentences. To begin with this comparison we assess the similarity in the annotated entities, for which we used the Concept Unique Identifier, or CUI.

As the annotators were given a list of CUIs all the annotated entities would have any of these elements, meaning that the comparison using these elements can tell which are the CUIs that only appear in either of these sources of information, and also tell us the similarity of the contents in the annotated set of sentences.

As a similarity metric we used Jaccard similarity coefficient, and to compare the similarity we started by assessing the similarity of the contents for the annotated entities.

Table 4.2 and Table 4.3 show the similarity coefficient for the different annotated entities. As we could have expected because of our sentence sampling approach, the similarity for drugs is much higher than the similarity for the diseases and symptoms, and even when using the conflated annotations the resulting similarity values do not increase noticeably.

	# PubMed	#Unique PubMed	# Twitter	#Unique Twitter	# Intersection	# Union	% similarity
Drug	2280	28	2094	30	28	30	93.33
Disease/Symptom	2284	295	1654	217	105	407	25.80

TABLE 4.3: Similarity between entities in Twitter and PubMed using the conflated annotations. Columns 2 and 4 show the total number of annotated tokens. Columns 3 and 5 show the number of unique tokens. Columns 6 and 7 show the intersection and the union, respectively, for the unique tokens (as shown in columns 3 and 5). The last column shows the similarity coefficient.

	# PubMed	#Unique PubMed	# Twitter	#Unique Twitter	# Intersection	# Union	% similarity
Outcome-positive	15	9	6	6	1	14	7.14
Outcome-negative	185	134	294	200	10	324	3.09
Reason-to-use	772	251	306	165	49	367	13.35

TABLE 4.4: Similarity between relations in Twitter and PubMed using the non-conflated annotations. Columns 2 and 4 show the total number of annotated tokens. Columns 3 and 5 show the number of unique tokens. Columns 6 and 7 show the intersection and the union, respectively, for the unique tokens (as shown in columns 3 and 5). The last column shows the similarity coefficient.

	# PubMed	#Unique PubMed	# Twitter	#Unique Twitter	# Intersection	# Union	% similarity
Outcome-negative	470	168	640	224	12	380	3.16
Benefit	1826	316	822	216	66	466	14.16

TABLE 4.5: Similarity between entities in Twitter and PubMed using the conflated annotations. Columns 2 and 4 show the total number of annotated tokens. Columns 3 and 5 show the number of unique tokens. Columns 6 and 7 show the intersection and the union, respectively, for the unique tokens (as shown in columns 3 and 5). The last column shows the similarity coefficient.

As we were interested in observing the similarity of the information, and given our sampling strategy another approach would be to compare the relations being mentioned in either Twitter and PubMed using the same strategy that we used for comparing the entities. Table 4.4 and Table 4.5 show the results when comparing the relations, although we can see that the similarity is really low and no relation gets a similarity score above 15% in neither the conflated nor the non-conflated data sets.

Given that the elements being mentioned are expected to be dissimilar, and some of the diseases and symptoms appearing in Twitter are not expected to appear in PubMed sentences, and conversely, more technical symptoms and diseases appearing in PubMed are probably not seen in tweets, we devised a new scenario where we could perform our comparison. In this case we decided to only take into consideration the set of drugs, symptoms and diseases appearing in both PubMed and Twitter using the relations where these elements appearing in both Twitter and PubMed are involved, as a way to compare the similarity in terms of the information contained in Twitter and PubMed. Those results can be see in Table 4.6 and Table 4.7

	# PubMed	#Unique PubMed	# Twitter	#Unique Twitter	# Inter-section	# Union	% similarity
Outcome-positive	10	5	5	5	1	9	11.11
Outcome-negative	78	51	179	104	10	145	6.90
Reason-to-use	494	112	231	103	48	167	28.74

TABLE 4.6: Similarity between relations in Twitter and PubMed using the non-conflated annotations on the set of elements appearing in Twitter and PubMed. Columns 2 and 4 show the total number of annotated tokens. Columns 3 and 5 show the number of unique tokens. Columns 6 and 7 show the intersection and the union, respectively, for the unique tokens (as shown in columns 3 and 5). The last column shows the similarity coefficient.

	# PubMed	#Unique PubMed	# Twitter	#Unique Twitter	# Inter-section	# Union	% similarity
Outcome-negative	208	71	408	126	12	185	6.49
Benefit	1192	160	660	147	66	241	27.39

TABLE 4.7: Similarity between entities in Twitter and PubMed using the conflated annotations on the set of elements appearing in Twitter and PubMed. Columns 2 and 4 show the total number of annotated tokens. Columns 3 and 5 show the number of unique tokens. Columns 6 and 7 show the intersection and the union, respectively, for the unique tokens (as shown in columns 3 and 5). The last column shows the similarity coefficient.

Relation	Mean	Standard deviation	CI (min)	CI (max)
Outcome-positive	12.00	6.68	11.42	12.59
Outcome-negative	6.55	1.05	6.45	6.64
Reason-to-use	29.04	1.73	28.88	29.19

TABLE 4.8: Similarity between relations in Twitter and PubMed using the non-conflated annotations on the set of elements appearing in Twitter and PubMed, and using Monte Carlo sampling. Mean, standard deviation and minimum and maximum confidence intervals (CI) are shown.

Relation	Mean	Standard deviation	CI (min)	CI (max)
Outcome-negative	6.23	0.94	6.15	6.31
Benefit	27.87	1.59	27.73	28.01

TABLE 4.9: Similarity between entities in Twitter and PubMed using the conflated annotations on the set of elements appearing in Twitter and PubMed, and using Monte Carlo sampling. Mean, standard deviation and minimum and maximum confidence intervals (CI) are shown.

To get a better overview of those comparisons we ran the same experiment 500 times using a Monte Carlo sampling strategy. Each run included 80% of the total number of the annotated sentences, chosen at random, and the corresponding annotations from those sentences.

We can see that the results appearing in Table 4.8 (using Monte Carlo sampling) are very close to the results shown in Table 4.6, and similarly the results shown in Table 4.9 are very close to the ones shown in Table 4.7. We can also notice that the standard deviation values are very small, although for "Outcome-positive" relations the variability is the

Source	Terms
Twitter	"2", **"anxiety"**, "anyone", "back", "bad", **"citalopram"**, **"cortisone"**, **"cymbalta"**, "day", "days", **"dose"**, **"effexor"**, "even", "feel", "fucking", "get", "go", "going", "good", "got", "help", "im", "it", "know", **"lamictal"**, "last", "like", "make", "makes", "making", "need", "night", **"paxil"**, **"prednisone"**, "put", "really", **"seroquel"**, "side", **"singulair"**, "still", "stop", "take", "taking", **"tamoxifen"**, "think", "time", "took", **"topamax"**, **"vyvanse"**, "want", "weight", **"withdrawal"**, "would", "youre", **"zoloft"**
PubMed	**"acetonide"**, **"acute"**, "associated", "background", **"bipolar"**, "case", "children", **"clinical"**, "combination", "compared", "conclusion", **"depressive"**, **"disorder"**, **"docetaxel"**, **"duloxetine"**, "effect", "effective", "efficacy", **"lamotrigine"**, **"lisdexamfetamine"**, "major", "may", "methods", "mg", **"montelukast"**, **"olanzapine"**, **"patient"**, **"patients"**, "randomize", "randomized", "report", "safety", **"serotonin"**, **"sertraline"**, "study", **"symptoms"**, **"therapy"**, **"topiramate"**, **"treated"**, **"treatment"**, "trial", **"triamcinolone"**, "use", "used", "using", **"venlafaxine"**, **"ziprasidone"**

TABLE 4.10: Expanded list of keywords used to characterize Twitter and PubMed topics. Medical related terms appear in bold.

highest. In general, "Outcome-negative" relations do not have much similarity between PubMed sentences and tweets and we can say that only in the case of "Benefit" relations and "Reason-to-use" relations there is some small similarity (always below 30%).

When we assessed the underlying similarities we noticed that some of the keywords were very generic, so we performed one more test extracting 25 keywords used when obtaining 5, 10, 15, 20, 25, and 30 topics, and out of all the extracted keywords we kept the terms that appeared only in PubMed keywords or Twitter keywords, as shown in Table 4.10.

When looking at Table 4.10, we noticed that in the case of Twitter we could manually identify 16 terms, in bold, out of the 55 keywords (29%) related to the medical domain, while in the case of PubMed we observed 24 out of 47 terms (51%) related to the medical domain, confirming that tweets contained more non-medical keywords hampering the identification of the correct labels in Wikipedia, probably contributing negatively to the similarity coefficients we observed.

When taking out the medical related terms from the lists of keywords in Table 4.10 we see that the remaining Twitter terms are very generic and often used in an informal setting when the linguistic register is not using polite constructions; conversely, for PubMed most words belong in the formal register, and even if those can be found in a number of domains, seeing those keywords together is a clear hint that the texts are likely to belong in the medical realm.

Another interesting finding is that the tweets in our data set use very generic keywords and some of them such as "go", "get", "make", "put", or "take" are commonly seen in phrasal verb constructions which is a clear indicator of informal register [183] and proven to be useful in telling genres apart [184]. We believe this feature can provide gains when taken into account in NLP systems in the area of pharmacovigilance.

Lastly, a further analysis of the politeness features would provide more insights on drug reports differences as it is clear that the drug use reports in our data sets make use of politeness features, and while a tweet would talk about "take" a drug, the same text in PubMed would mention the "use" of a medicine; and a "report" found in PubMed would roughly correspond to what a Twitter user "thinks" or "feels", showing that there is a correlation in the set of keywords. Even if formality features are playing an important role the use of taboo words and orthographic variations are elements of potential help in identifying drug use reports in Twitter.

4.2 Comparing the linguistic register used in drug-use reports from Twitter and in generic tweets

In this section we explore Hypothesis 2 to understand if tweets that vary on their contents can be told apart by using the register analysis. We use a set of generic tweets, containing messages that discuss a wide range of topics, and compare that data set against a set of tweets that contain reports on the drugs use. We believe that the topicality of the contents can have an impact on the ways in which the messages are written, and even if both data sets belong in the same linguistic register the framework we use to assess the differences in the use of the register may be able to capture some of these traits.

4.2.1 Data collection

For testing **Hypothesis 2** we prepared a set up that would use the social media messages that we released in TwiMed, as that was the curated data set that we believed would not contain the same type of linguistic constructions seen in generic social media messages.

The data set that we wanted to compare against that data set had to come from Twitter too, and we initially thought on using the data set released by Cheng et al. [185] which is a data set used to study the geolocation of the users that are posting messages in Twitter.

After an exploratory analysis we observed that not all the tweets on that data set could not be used as out of the total of 9,001,669 tweets, the total number of different Twitter users contributing messages was 106,363.

When creating the tweets data set released in TwiMed we made sure that no user contributed more than 5 tweets, and in Cheng's data set there were 104,264 out of the 106,363 different users who had contributed more than 5 tweets, and even if a high

number of users contributed a high number of tweets the base of users was large enough so that we could use this data set.

In general, collecting tweets from other research groups is not very straightforward given a number of limitations imposed by Twitter and most researchers have to come up with their own strategies, but in this case we decided to use this data set because it was collected for a geo-location study and the sample would be on very different topics.

As Cheng's data set was curated during 2009 and 2010 we decided to gather another data set using Twitter's Streaming API[2] to have another set of generic tweets apart from Cheng's data set to get a better overview of the differences between our drug-related tweets sample and two different generic tweets samples.

When curating TwiMed data the used set of keywords we provided to Twitter's Streaming API were the names of the drugs, and in this case we decided to use a much more generic set of keywords. We initially though of using verbs such as "take", "see", "feel" which were generic enough to be part of a number of sentences, but then we realised that using a closed list of verbs could be biasing the sample, so we decided to apply a more generic set of keywords.

We decided to use as our set of keywords the list of English stop words provided by NLTK[186], and out of those keywords we removed the strings that only had one character ("i", "a", "s" and "t") as these keywords could appear in tweets from many different languages, and also because the remaining list of keywords was large enough (123 strings) so that we could get a large number of tweets in almost no time.

By using those keywords we managed to retrieve 91,190 tweets in 1 hour. We then proceeded to remove all tweets not written in English from our data set[3] obtaining 78,081 tweets in total.

The tweets in these data sets were filtered in the same way as we filtered the tweets when curating TwiMed corpus. We discarded the following tweets:

- Messages containing keywords related to marketing campaigns ("buy", "cheap", "online", "pharmacy", "price"):
- Messages containing URLs.
- Messages that are "Retweets" (i.e. forwarded tweets).
- Messages that are addressed to a user in particular.

[2] https://dev.twitter.com/streaming/public

[3] Cheng's data set did not provide that information, although all the tweets in his data set were posted by users living in the U.S.

- Messages that have 20 or fewer characters.

Once we applied that filtering step we had 434,498 tweets from Cheng's data set (406,947 tweets after removing similar tweets), and another 19,206 tweets from our data set (12,841 after removing similar tweets).

Out of all the messages, we decided to extract 6000 tweets from Cheng's data set, and 2000 tweets from our custom data set. Having two very different sets of generic tweets could show whether the trends we observed would be present in other generic tweets, and following this approach we obtained the micro and macro results using Biber approach.

4.2.2 Linguistic similarity

Once we had the data in place we developed our linguistic studies on the set of tweets. In the following sections we present our findings after comparing the set of tweets containing drugs reports against two data sets composed of generic tweets that differ in the dates when the data were gathered, the size of the data and the strategies that were used to filter those generic tweets.

Biber's MD analysis was presented in 2.1.1 using a sample tweet and a sample sentence from PubMed, and before moving on to the actual study we are going to show here some generic and drug-related tweets obtained at random from our data sets. In this section we will use tweets obtained from three different data sets: Chen's data set, our custom data set of generic tweets, and the set of tweets from TwiMed.

A sample of the tweets that can be seen in Chen's data set are the following:

- Recovering from lovely dinner last night with two nice glasses of red wine. My daughter ordered filet mignon, didn't know 24 was price tag.
- Had a great day today :) home to study then back to work to get the store ready for a big visit..
- kelly's gonna watch clover for the weekend and feed her.
- is going to lunch & then shopping w/ my "bestest"! We have a lot of catching up to do!!!
- Ok who is ready for Wales to beat the Aussies?!

A sample of the tweets that can be seen in our custom data set of generic tweets are the following:

- Going for a bike ride
- Ah, that wacky old rogue Gerry Adams is getting himself into a bit of mischief again!
- aisha and megan are the good nice ones be like them
- last night, i came to a realization
- Its all I've been hearing these past 3 days

A first observation tells us that the tweets included in these two data sets have very generic contents mainly around the current activities that the person writing the tweet has recently done or is about to do such as information on generic activities as could be meeting a friend or riding a bike, and thoughts on TV shows.

The counterpart to those messages are the tweets contained in TwiMed data set, which are tweets containing information on the drug and symptoms and diseases related to that drug intake. Sample tweets contained in our set of drug-use tweets (TwiMed) are the following:

- I waited too long to take my Zoloft and now I have been feeling irritated all day. Everything is getting on my nerves.
- The absolute last resort to numb my pain in my back I am getting cortisone jags today at the age of 21
- Paxil withdrawal update: still on 1/2 pill every other day, so sleepy, dizzy and a little sad, Since I know the reason it's ok. over soon!
- Doc upped my depression/anxiety med dosage and omg....I feel like I am dying! Anyone else take Effexor?
- Maybe with 350mg of trazodone I can sleep

A first inspection tells us that even if these three sets of sentences were obtained from Twitter, there are differences between the sentences in Chen's data set and TwiMed's tweets in the use of abbreviation and the use of slang words (very characteristic in Chen's sentences). Similarly, when comparing TwiMed tweets and our custom set of generic tweets we can sense that TwiMed tweets are longer and contain more punctuation signs. Those differences are not the only ones, and to study them formally we use Biber's multidimensional analysis, which accounts for a number of elements to present the differences in terms of a number of traits related to the use of different linguistic registers.

# Dim.	Generic			Drug-related		
	min	max	mean (std)	min	max	mean (std)
1	-23.160	-0.540	-7.485 (1.899)	-12.402	-3.388	-8.443 (1.460)
2	0.000	0.307	0.025 (0.039)	0.000	0.200	0.024 (0.034)
3	-0.304	0.360	0.092 (0.073)	-0.143	0.356	0.084 (0.062)
4	0.000	0.217	0.015 (0.025)	0.000	0.163	0.017 (0.024)
5	0.000	0.084	0.001 (0.006)	0.000	0.039	0.002 (0.006)
6	0.000	0.122	0.005 (0.013)	0.000	0.083	0.005 (0.012)
7	0.000	0.050	0.000 (0.002)	0.000	0.027	0.000 (0.002)

TABLE 4.11: Minimum, maximum, mean and standard deviation micro results for the seven dimensions using 6000 generic tweets and 1000 drug-related tweets.

# Dim.	Generic	Drug-related
1.Involved versus Information Productions	0.693	0.651
2.Narrative versus Non-Narrative Concerns	0.082	0.077
3.Explicit versus Situation-Dependent Reference	0.597	0.584
4.Overt Expressions of Persuasion	0.070	0.078
5.Abstract versus Non-Abstract Information	0.015	0.023
6.On-Line Informational Elaboration	0.041	0.043
7.Academic qualification	0.002	0.004

TABLE 4.12: Normalized macro results for the seven dimensions using 6000 generic tweets and 1000 drug-related tweets.

Experiments on Cheng's data set

We computed the seven dimensions for all the texts in using the 6000 generic tweets we got from Cheng's data set and the sample of 1000 drug-related tweets that we released in TwiMed corpus 3.3 and obtained the maximum and minimum values as well as the mean and the standard deviation values for each dimension as presented in Table 4.11.

As these texts were obtained from Twitter we used ARK tagger [78] for both tagging and tokenizing the sentences.

We also obtained the macro results (see Table 4.12) when aggregating the values for all the dimensions in all sentences and normalized the results using the previously obtained minimum and maximum values so that all values are in the range [0,1].

We can see in Table 4.11 that the mean values for dimensions one and three were the ones showing the most different values, although except for dimension one, most of the values are very similar in both data sets. This observation is in line with the type of tweets as we would expect a drug-related tweets to contain more information (dimension one)

# Dim.	Generic			Drug-related		
	Mean	CI (min)	CI (max)	Mean	CI (min)	CI (max)
1	-7.484	-7.488	-7.481	-8.443	-8.443	-8.443
2	0.025	0.025	0.025	0.024	0.024	0.024
3	0.092	0.092	0.093	0.084	0.084	0.084
4	0.015	0.015	0.015	0.017	0.017	0.017
5	0.001	0.001	0.001	0.002	0.002	0.002
6	0.005	0.005	0.005	0.005	0.005	0.005
7	0.000	0.000	0.000	0.000	0.000	0.000

TABLE 4.13: Mean, minimum and maximum confidence intervals (CI) micro results for the seven dimensions from 6000 generic tweets and 1000 drug-related tweets using Monte Carlo sampling.

than a generic group of tweets. Also, the set of generic tweets would be more dependent on the situation (dimension three) than the drug-related tweets.

Besides the main observation regarding dimension one the differences were not too evident. In Table 4.12, showing the macro results, these variations seem to be minimal. On the other hand, Table 4.12 helps in seeing which data set has more traits related to the use of on-line information elaboration (dimension 6), and also shows that the drug-related set of tweets makes use of more academic qualifications (dimension 7) and carries more abstract information, which is usually related to more formal texts.

To evaluate the possible differences caused by the variability of the sentences we ran the same tests using a Monte Carlo sampling approach, retrieving 2/3 of the annotated sentences (4000 sentences for the generic tweets and 750 drug-related tweets) chosen at random from each data set in each run. We run 100 of these experiments and got the results presented in Table 4.13 and Table 4.14. These results show the mean and the standard deviation values using the similarity scores obtained in each of the 100 runs. We also use those 100 results to compute the confidence intervals at 95%.

As a by-product needed to compute the values shown in Table 4.11 and Table 4.12 we obtained the values for each factor, presented in Table 4.15.

The micro results in Table 4.13 show that there is a clear difference in dimension one ("Involved versus Information Productions") between both data sets in terms of the confidence intervals where we could find 95% of the mean values for each data set. For most of the dimensions there is no overlap in terms of the confidence intervals, although most of these values are very close, and in fact dimension one and dimension three are the only dimensions where there is some clear difference in terms of the resulting values.

The macro results shown in Table 4.14 also evidence the differences in dimensions one and three, although in this case most of the values are the same that we observed in

# Dim.	Generic	Drug-related
1.Involved versus Information Productions	0.677	0.635
2.Narrative versus Non-Narrative Concerns	0.082	0.077
3.Explicit versus Situation-Dependent Reference	0.597	0.584
4.Overt Expressions of Persuasion	0.070	0.078
5.Abstract versus Non-Abstract Information	0.015	0.023
6.On-Line Informational Elaboration	0.041	0.043
7.Academic qualification	0.002	0.004

TABLE 4.14: Normalized macro results for the seven dimensions in 6000 generic tweets and 1000 drug-related tweets using Monte Carlo sampling.

Table 4.12, and only the result for the first dimension change, but even in this case the results are still comparable to the results we obtained in Table 4.12.

The differences at factor level can be observed in the following Table:

#	Factor Name	Generic Mean (sd)	Drug-related Mean (sd)	Ratio Mean (sd)
56	Private Verbs	0.010 (0.024)	0.011 (0.024)	1.148 (1.041)
60	Subordinator That Deletion	0.003 (0.013)	0.003 (0.013)	1.122 (1.021)
59	Contractions	0.009 (0.023)	0.014 (0.026)	1.533 (1.142)
3	Present Verbs	0.046 (0.053)	0.052 (0.050)	1.132 (1.062)
7	Second Person Pronouns	0.009 (0.028)	0.009 (0.026)	1.051 (1.055)
12	Pro Verb Do	0.004 (0.017)	0.003 (0.013)	1.202 (1.300)
67	Analytic Negation	0.007 (0.020)	0.010 (0.023)	1.559 (1.153)
10	Demonstrative Pronouns	0.003 (0.013)	0.003 (0.012)	1.268 (1.051)
49	Emphatics	0.010 (0.026)	0.009 (0.020)	1.133 (1.285)
6	First Person Pronouns	0.038 (0.052)	0.053 (0.057)	1.380 (1.092)
9	Pronoun It	0.008 (0.022)	0.009 (0.020)	1.160 (1.116)
19	Be as Main Verb	0.009 (0.026)	0.008 (0.020)	1.095 (1.291)
35	Causative Adverbial Subordinator (Because)	0.000 (0.005)	0.001 (0.006)	1.970 (1.198)
50	Discourse Particles	0.001 (0.006)	0.000 (0.004)	2.182 (1.757)
11	Indefinite Pronoun	0.004 (0.018)	0.004 (0.016)	1.104 (1.155)
47	Hedges	0.001 (0.007)	0.001 (0.008)	1.493 (1.157)
48	Amplifiers	0.001 (0.009)	0.001 (0.011)	1.450 (1.232)
34	Sentence Relatives	0.000 (0.002)	0.000 (0.004)	4.487 (2.198)
13	Wh Questions	0.001 (0.009)	0.001 (0.006)	1.387 (1.504)
52	Possibility Modals	0.003 (0.013)	0.005 (0.017)	2.007 (1.315)
65	Independent Clause Coordination	0.000 (0.003)	0.000 (0.004)	1.121 (1.086)
23	Wh Clauses	0.001 (0.007)	0.000 (0.004)	1.852 (1.617)
61	Stranded Prepositions	0.002 (0.011)	0.001 (0.005)	2.578 (2.094)
16	Total Other Nouns	0.007 (0.023)	0.011 (0.025)	1.694 (1.122)
44	Word Length	3.847 (1.274)	4.185 (1.095)	1.088 (1.163)
39	Prepositional Phrases	0.079 (0.067)	0.083 (0.058)	1.051 (1.160)
43	Type/Token Ratio	9.840 (3.676)	11.277 (3.062)	1.146 (1.200)
40	Attributive Adjectives	0.052 (0.060)	0.061 (0.058)	1.179 (1.033)

1	Past Verbs	0.018 (0.036)	0.018 (0.033)	1.026 (1.102)
8	Third Person Personal Pronouns No It	0.006 (0.022)	0.004 (0.017)	1.648 (1.259)
2	Perfect Aspect Verbs	0.002 (0.011)	0.004 (0.014)	1.781 (1.185)
55	Public Verbs	0.003 (0.013)	0.003 (0.011)	1.014 (1.128)
66	Synthetic Negation	0.001 (0.010)	0.002 (0.010)	1.246 (1.066)
25	Present Participial Clauses	0.005 (0.023)	0.002 (0.011)	2.740 (2.093)
32	WH relative Clauses On Object	0.000 (0.002)	0.000 (0.003)	2.136 (1.328)
33	Pied-Piping Relative Clauses	0.000 (0.001)	0.000 (0.000)	- (-)
31	WH relative Clauses On Subject	0.001 (0.006)	0.001 (0.006)	1.279 (1.057)
64	Phrasal Coordination	0.005 (0.018)	0.006 (0.018)	1.291 (1.018)
14	Nominalizations	0.321 (0.157)	0.293 (0.124)	1.094 (1.259)
5	Time Adv	0.010 (0.028)	0.006 (0.018)	1.510 (1.519)
4	Place Adv	0.002 (0.014)	0.001 (0.007)	2.228 (1.858)
42	Total Adverbs	0.040 (0.057)	0.044 (0.050)	1.113 (1.127)
24	Infinitives	0.013 (0.028)	0.014 (0.027)	1.104 (1.042)
54	Predictive Modals	0.003 (0.014)	0.003 (0.011)	1.260 (1.212)
57	Suasive Verbs	0.001 (0.010)	0.001 (0.006)	1.483 (1.541)
37	Conditional Adverbial Subordinator (If/Unless)	0.002 (0.011)	0.003 (0.011)	1.387 (1.048)
53	Necessity Modals	0.001 (0.008)	0.001 (0.006)	1.300 (1.289)
63	Split Auxiliaries	0.004 (0.015)	0.006 (0.017)	1.649 (1.196)
45	Conjuncts	0.001 (0.007)	0.001 (0.005)	1.082 (1.304)
17	Agentless Passives	0.000 (0.004)	0.000 (0.005)	1.789 (1.165)
26	Past Participial Clauses	0.000 (0.006)	0.000 (0.005)	1.031 (1.378)
18	By Passives	0.000 (0.001)	0.000 (0.000)	- (-)
27	Past Participial WHIZ	0.001 (0.007)	0.002 (0.009)	2.054 (1.221)
38	Other Adverbial Subordinators	0.001 (0.006)	0.001 (0.008)	1.971 (1.290)
21	That Verb Complements	0.001 (0.007)	0.001 (0.007)	1.624 (1.050)
51	Demonstratives	0.008 (0.023)	0.008 (0.019)	1.000 (1.230)
30	That Relative On Object Position	0.000 (0.005)	0.000 (0.005)	1.070 (1.055)
22	That Adjective Complements	0.000 (0.003)	0.000 (0.004)	1.131 (1.085)
58	Seem/Appear Verbs	0.000 (0.004)	0.001 (0.005)	1.920 (1.166)

TABLE 4.15: Table showing the results for each factor used to compute Biber's features using the sample of 6000 generic tweets and 1000 drug-related tweets. Mean values and Standard deviation values for the 6000 generic tweets and the 1000 drug-related tweets are shown in Columns 3 and 4 respectively. The last column shows the mean and standard deviation ratios using the values from the previous 2 columns.

In the Table 4.15 we can observe that there are some factors for which the different values obtained using generic tweets and the results obtained when using drug-related tweets differ. In particular the most dissimilar factor can be found in dimension one and

it is the use of sentence relatives ("which"), appearing more than four times more often in drug-related tweets than in generic tweets. Other different factor is the use of present participial clauses, found in dimension two, where we can see that generic tweets use this construction more than twice as often as drug-related tweets, which makes sense as generic tweets are expected to contain a number of real-time reports, written using different forms of present tenses. One finding that was expected to a certain extent is the difference we can see in the third most dissimilar factor, stranded prepositions. These constructions are commonly used in informal texts [187], and not surprisingly we observe that generic tweets use this construction more often than drug-related tweets do.

Other factors that appear two or more times more often in generic tweets than in drug-related tweets are the use of discourse particles ("well", "anyway"...) and the use of place adverbs ("above", "around"...). On the other hand, the features appearing two or more times more often in drug-related tweets than in generic tweets are wh- relative clauses on object ("that" in object position), past participial whiz (past participle combined with the deletion of a Wh-word plus a form of be, quite often "is"), and possibility modals ("can", "may", "might" or "could").

Once we had these results in place we decided to confirm if these findings were also observed when using another set of generic tweets for which we used our custom sample of tweets.

Experiments on a custom data set

We computed the seven dimensions for all the texts in using the 2000 generic tweets we got in our custom data set and the sample of 1000 drug-related tweets that we released in TwiMed corpus 3.3 and obtained the maximum and minimum values as well as the mean and the standard deviation values for each dimension as presented in Table 4.16.

We also obtained the macro results (see Table 4.17) when aggregating the values for all the dimensions in all sentences and normalized the results using the previously obtained minimum and maximum values so that all values are in the range [0,1]. Please note that the normalization values used in Table 4.12 are expected to change, so in Table 4.17 the results for drug-related tweets are not the same as the ones in Table 4.12.

These results show that some of the dimensions are dependent on the sample of tweets we used, although there are some dimensions with similar differences in both tables as are dimension one, two, five and seven. These results can be seen when comparing Table 4.11 and Table 4.16, and also when observing the values in Table 4.12 against the values in Table 4.17.

# Dim.	Generic min	Generic max	Generic mean (std)	Drug-related min	Drug-related max	Drug-related mean (std)
1	-17.900	-1.489	-6.715 (1.878)	-12.402	-3.388	-8.443 (1.460)
2	0.000	0.338	0.027 (0.046)	0.000	0.200	0.024 (0.034)
3	-0.230	0.381	0.078 (0.089)	-0.143	0.356	0.084 (0.062)
4	0.000	0.260	0.019 (0.032)	0.000	0.163	0.017 (0.024)
5	0.000	0.065	0.001 (0.007)	0.000	0.039	0.002 (0.006)
6	0.000	0.220	0.007 (0.019)	0.000	0.083	0.005 (0.012)
7	0.000	0.021	0.000 (0.001)	0.000	0.027	0.000 (0.002)

TABLE 4.16: Minimum, maximum, mean and standard deviation micro results for the seven dimensions using 2000 generic tweets and 1000 drug-related tweets.

# Dim.	Generic	Drug-related
1.Involved versus Information Productions	0.682	0.576
2.Narrative versus Non-Narrative Concerns	0.079	0.070
3.Explicit versus Situation-Dependent Reference	0.504	0.513
4.Overt Expressions of Persuasion	0.075	0.065
5.Abstract versus Non-Abstract Information	0.022	0.029
6.On-Line Informational Elaboration	0.031	0.024
7.Academic qualification	0.002	0.007

TABLE 4.17: Normalized macro results for the seven dimensions using 2000 generic tweets and 1000 drug-related tweets.

Continuing with the same approach we took before, and to evaluate the possible differences caused by the variability of the sentences, we ran the same tests using a Monte Carlo sampling approach retrieving 2/3 of the annotated sentences (1333 sentences for the generic tweets and 750 drug-related tweets) chosen at random from each data set in each run. We run 100 of these experiments and got the results presented in Table 4.18 and Table 4.19. These results show the mean and the standard deviation values using the similarity scores obtained in each of the 100 runs. We also used those 100 results to compute the confidence intervals at 95%.

As a by-product needed to compute the values shown in Table 4.16 and Table 4.17 we obtained the values for each factor, presented in Table 4.20.

The micro results in 4.18 make it very clear that in terms of dimension one there is a difference in these types of messages, showing that generic tweets are more involved and our set of drug-related tweets are more informative. Dimension two as well shows that, even if quite small, there is a constant difference appearing in Table 4.13 and Table 4.18, telling us that generic tweets tend to be more narrative than drug-related tweets. Dimensions five and seven show the same values in both Table 4.13 and 4.18, for which

# Dim.	Generic			Drug-related		
	Mean	CI (min)	CI (max)	Mean	CI (min)	CI (max)
1	-6.714	-6.720	-6.709	-8.443	-8.443	-8.443
2	0.027	0.026	0.027	0.024	0.024	0.024
3	0.078	0.077	0.078	0.084	0.084	0.084
4	0.019	0.019	0.020	0.017	0.017	0.017
5	0.001	0.001	0.001	0.002	0.002	0.002
6	0.007	0.007	0.007	0.005	0.005	0.005
7	0.000	0.000	0.000	0.000	0.000	0.000

TABLE 4.18: Mean, minimum and maximum confidence intervals (CI) micro results for the seven dimensions from 2000 generic tweets and 1000 drug-related tweets using Monte Carlo sampling.

# Dim.	Generic	Drug-related
1.Involved versus Information Productions	0.625	0.528
2.Narrative versus Non-Narrative Concerns	0.079	0.070
3.Explicit versus Situation-Dependent Reference	0.504	0.513
4.Overt Expressions of Persuasion	0.075	0.065
5.Abstract versus Non-Abstract Information	0.023	0.029
6.On-Line Informational Elaboration	0.031	0.024
7.Academic qualification	0.002	0.007

TABLE 4.19: Normalized macro results for the seven dimensions in 2000 generic tweets and 1000 drug-related tweets using Monte Carlo sampling.

we can conclude that those differences are expected to be constant and drug-related tweets convey more abstract information and make use of more academic qualifications.

In terms of the macro results we can see in Table 4.19 that dimensions three, four and six get the opposite trends in its values when comparing these with the values in Table 4.14. We can conclude that the degree of information in drug-related tweets is greater than in generic tweets, as it also is the degree of abstractness in that information and the academic qualifications. On the other hand, generic tweets seem to contain messages expressed in a more narrative style.

Directly comparable with Table 4.15, Table 4.20 shows the differences at factor level that can be observed in the sample of 2000 generic tweets and in the sample of 1000 drug-related tweets:

#	Factor Name	Generic Mean (sd)	Drug-related Mean (sd)	Ratio Mean (sd)
56	Private Verbs	0.012 (0.031)	0.011 (0.024)	1.090 (1.333)
60	Subordinator That Deletion	0.004 (0.017)	0.003 (0.013)	1.129 (1.312)
59	Contractions	0.016 (0.036)	0.014 (0.026)	1.172 (1.358)
3	Present Verbs	0.061 (0.066)	0.052 (0.050)	1.192 (1.334)
7	Second Person Pronouns	0.018 (0.044)	0.009 (0.026)	2.167 (1.696)
12	Pro Verb Do	0.005 (0.022)	0.003 (0.013)	1.449 (1.730)
67	Analytic Negation	0.011 (0.030)	0.010 (0.023)	1.086 (1.266)
10	Demonstrative Pronouns	0.004 (0.017)	0.003 (0.012)	1.051 (1.443)

49	Emphatics	0.013 (0.039)	0.009 (0.020)	1.566 (1.881)
6	First Person Pronouns	0.059 (0.072)	0.053 (0.057)	1.126 (1.258)
9	Pronoun It	0.011 (0.031)	0.009 (0.020)	1.218 (1.546)
19	Be as Main Verb	0.011 (0.031)	0.008 (0.020)	1.315 (1.565)
35	Causative Adverbial Subordinator (Because)	0.001 (0.008)	0.001 (0.006)	1.229 (1.292)
50	Discourse Particles	0.001 (0.007)	0.000 (0.004)	2.150 (1.913)
11	Indefinite Pronoun	0.007 (0.026)	0.004 (0.016)	1.757 (1.654)
47	Hedges	0.000 (0.007)	0.001 (0.008)	2.142 (1.079)
48	Amplifiers	0.000 (0.006)	0.001 (0.011)	3.052 (1.840)
34	Sentence Relatives	0.000 (0.003)	0.000 (0.004)	2.920 (1.618)
13	Wh Questions	0.001 (0.012)	0.001 (0.006)	2.321 (2.162)
52	Possibility Modals	0.004 (0.020)	0.005 (0.017)	1.223 (1.207)
65	Independent Clause Coordination	0.000 (0.002)	0.000 (0.004)	2.112 (1.650)
23	Wh Clauses	0.001 (0.009)	0.000 (0.004)	2.820 (2.132)
61	Stranded Prepositions	0.001 (0.011)	0.001 (0.005)	2.442 (2.120)
16	Total Other Nouns	0.006 (0.025)	0.011 (0.025)	1.900 (1.008)
44	Word Length	3.583 (1.093)	4.185 (1.095)	1.168 (1.003)
39	Prepositional Phrases	0.077 (0.072)	0.083 (0.058)	1.067 (1.236)
43	Type/Token Ratio	8.812 (3.433)	11.277 (3.062)	1.280 (1.121)
40	Attributive Adjectives	0.050 (0.065)	0.061 (0.058)	1.217 (1.128)
1	Past Verbs	0.018 (0.041)	0.018 (0.033)	1.017 (1.272)
8	Third Person Personal Pronouns No It	0.009 (0.028)	0.004 (0.017)	2.392 (1.654)
2	Perfect Aspect Verbs	0.002 (0.014)	0.004 (0.014)	1.589 (1.040)
55	Public Verbs	0.002 (0.013)	0.003 (0.011)	1.089 (1.144)
66	Synthetic Negation	0.003 (0.017)	0.002 (0.010)	1.563 (1.788)
25	Present Participial Clauses	0.002 (0.013)	0.002 (0.011)	1.081 (1.187)
32	WH relative Clauses On Object	0.000 (0.006)	0.000 (0.003)	1.987 (2.158)
33	Pied-Piping Relative Clauses	0.000 (0.002)	0.000 (0.000)	- (-)
31	WH relative Clauses On Subject	0.001 (0.008)	0.001 (0.006)	1.193 (1.394)
64	Phrasal Coordination	0.004 (0.017)	0.006 (0.018)	1.677 (1.085)
14	Nominalizations	0.295 (0.186)	0.293 (0.124)	1.007 (1.492)
5	Time Adv	0.007 (0.026)	0.006 (0.018)	1.064 (1.424)
4	Place Adv	0.002 (0.012)	0.001 (0.007)	1.730 (1.690)
42	Total Adverbs	0.055 (0.073)	0.044 (0.050)	1.245 (1.453)
24	Infinitives	0.014 (0.034)	0.014 (0.027)	1.002 (1.252)
54	Predictive Modals	0.004 (0.019)	0.003 (0.011)	1.590 (1.667)
57	Suasive Verbs	0.002 (0.012)	0.001 (0.006)	1.992 (1.821)
37	Conditional Adverbial Subordinator (If/Unless)	0.004 (0.016)	0.003 (0.011)	1.298 (1.452)
53	Necessity Modals	0.002 (0.013)	0.001 (0.006)	2.062 (2.155)
63	Split Auxiliaries	0.007 (0.023)	0.006 (0.017)	1.182 (1.348)
45	Conjuncts	0.001 (0.006)	0.001 (0.005)	1.138 (1.226)

17	Agentless Passives	0.000 (0.004)	0.000 (0.005)	1.517 (1.039)
26	Past Participial Clauses	0.000 (0.003)	0.000 (0.005)	3.195 (1.609)
18	By Passives	0.000 (0.000)	0.000 (0.000)	- (-)
27	Past Participial WHIZ	0.001 (0.009)	0.002 (0.009)	1.915 (1.012)
38	Other Adverbial Subordinators	0.001 (0.011)	0.001 (0.008)	1.045 (1.365)
21	That Verb Complements	0.001 (0.012)	0.001 (0.007)	1.123 (1.560)
51	Demonstratives	0.011 (0.032)	0.008 (0.019)	1.365 (1.737)
30	That Relative On Object Position	0.000 (0.007)	0.000 (0.005)	1.052 (1.434)
22	That Adjective Complements	0.000 (0.003)	0.000 (0.004)	2.105 (1.424)
58	Seem/Appear Verbs	0.000 (0.003)	0.001 (0.005)	3.343 (1.783)

TABLE 4.20: Table showing the results for each factor used to compute Biber's features using the sample of 2000 generic tweets and 1000 drug-related tweets. Mean values and Standard deviation values for the 2000 generic tweets and the 1000 drug-related tweets are shown in Columns 3 and 4 respectively. The last column shows the mean and standard deviation ratios using the values from the previous 2 columns.

In the Table 4.15 we observed some trends, and Table 4.20 now confirms a number of them. In this case the most dissimilar factor is the use of seem/appear verbs, or academic qualifications (in dimension seven), that we notice to appear more than three times more often in the drug-related set of tweets than in the generic tweets. This trend was also observed in Table 4.15, but the difference was not so big, and that is why we did not mention it when pointing out the most different values.

When looking for other different values we find that the use of past participial clauses is another outstanding factor appearing thrice times more often in the set of drug-related tweets, but this difference is probably due to the set of tweets itself and not a generic characteristic of drug-related tweets against generic tweets as this difference was not present in Table 4.15. In fact, the trend for this factor was reversed as we observed that generic tweets used this feature more often than drug-related tweets.

The following different feature is the use of the amplifiers ("absolutely", "altogether", "completely"...), which we would expect to appear more often in the set of drug-related tweets as these would be the a more formal way of stressing the importance of the contents and, in a way, the counterpart to the emphatics ("just", "really"..) that we can see more often in informal texts, which in this case corresponds to the generic set of tweets. The fourth most different factor, the use of sentence relatives, is very interesting as it also appear as one of the most dissimilar factor in Table 4.15, making it clear that this feature is very characteristic in each data set. The last factor appearing almost three times as often in generic tweets than in drug-related tweets are "wh clauses" (e.g. "I believed **what** he told me"), which was also twice as frequent in the sample of generic

tweets we used when computing Table 4.15, confirming that this factor is also stable across data sets.

These evidences show that there are some factors that are more often seen in generic tweets and some factors that are more frequent in drug-related tweets, confirming that even if the samples were obtained from the same source of information, Twitter, **Hypothesis 2** should be accepted after observing that informal drug use reports in Twitter do not make use of all the linguistic elements commonly seen in these social media messages, and conversely, generic tweets do not use to the same extent the set of linguistic elements that we can often see in drug-related tweets.

After exploring the differences in Twitter texts we then moved on to explore the differences in drug use reports found in Twitter and PubMed texts.

4.3 Comparing the linguistic register used in drug-use reports from Twitter and from PubMed

As previously mentioned, the number of differences between PubMed and Twitter texts are of very different kinds, and aside for exploring the similarity of their contents we are also interested on exploring different aspects as can be the linguistic similarities, related to the use of a certain register, or the topical similarity, related to the topics being discussed in each source of information.

To achieve those goals we focus on pharmacological texts from Twitter and PubMed under the hypotheses that scientific texts and texts from that generic social media network will be different enough to understand which are the most similar and most different linguistic elements between those reports.

We start this study using the 6000 sentences we filtered out when curating TwiMed corpus 3.3, and we then move on to a more focused set of sentences by using the 1000 annotated sentences from PubMed and Twitter that we released in TwiMed.

To begin with our study we started exploring the linguistic dimension related to the register.

4.3.1 Linguistic similarity

We analysed the variation in a set of linguistic aspects by using the linguistic dimensions defined by [8]. Those linguistic dimensions, introduced in 2.1.1, are the representation

of seven different aspects in texts and evidence the strength in the following linguistic dimensions[4].

As these texts were obtained from different sources of information, we used Charniak-Johnson parser [77] for both tagging and tokenizing PubMed sentences, while in the case of Twitter we used ARK tagger [78].

- (1) **Involved *versus* Information Productions**: Marks affective, interactional and generalized content versus high informational density and exact informational content.
- (2) **Narrative *versus* Non-Narrative Concerns**: Distinguishes narrative discourse from other types of discourse.
- (3) **Explicit *versus* Situation-Dependent Reference**: Distinguishes between highly explicit, context-independent reference and non-specific, situation-dependent reference.
- (4) **Overt Expressions of Persuasion**: Marks persuasion, including the speaker's own persuasion or argumentative discourse designed to persuade the addressee.
- (5) **Abstract *versus* Non-Abstract Information**: Indicates abstract, technical and formal informational discourse.
- (6) **On-Line Informational Elaboration**: Marks informational discourse but produced under real-time conditions.
- (7) **Academic qualification**: Marks academic qualification or hedging.

Each one of those seven dimensions involves a different number of linguistic characteristics of the texts as can be the count of the number of present verbs, place and time adverbials or some forms of negation among other linguistic elements. Biber obtained the counts of sixty-seven different characteristics in the texts, which he grouped using Principal Factor Analysis (PFA)[5] obtaining the seven linguistic dimensions[6] presented above.

We computed the seven dimensions for all the texts in our Twitter and PubMed data sets separately obtaining the maximum and minimum values as well as the mean and the standard deviation values for each dimension as presented in Table 4.21.

[4] Dimensions 1, 2, 3 and 4 do not only evidence the strength in one linguistic dimension as these characterize how the weakness of one linguistic aspect implies the strength of its counterpart.

[5] Biber used PFA over Principal Component Analysis (PCA) because PFA accounted for the shared variance instead of all of the variance.

[6] Although Biber computed 67 features the linguistic dimensions only make use of 59 of those features.

# Dim.	Twitter min	Twitter max	Twitter mean (std)	PubMed min	PubMed max	PubMed mean (std)
1	-37.080	-2.227	-8.097 (1.641)	-14.898	-4.682	-9.115 (1.429)
2	0.000	0.399	0.028 (0.041)	0.000	0.272	0.032 (0.033)
3	-0.164	0.400	0.083 (0.065)	-0.093	0.283	0.107 (0.038)
4	0.000	0.163	0.018 (0.025)	0.000	0.117	0.007 (0.014)
5	0.000	0.089	0.002 (0.006)	0.000	0.073	0.004 (0.009)
6	0.000	0.101	0.006 (0.014)	0.000	0.101	0.005 (0.011)
7	0.000	0.039	0.000 (0.001)	0.000	0.032	0.000 (0.002)

TABLE 4.21: Minimum, maximum, mean and standard deviation micro results for the seven dimensions using 6000 sentences from Twitter and PubMed.

# Dim.	Twitter	PubMed
1.Involved versus Information Productions	0.832	0.802
2.Narrative versus Non-Narrative Concerns	0.071	0.079
3.Explicit versus Situation-Dependent Reference	0.437	0.480
4.Overt Expressions of Persuasion	0.110	0.043
5.Abstract versus Non-Abstract Information	0.020	0.046
6.On-Line Informational Elaboration	0.061	0.045
7.Academic qualification	0.003	0.005

TABLE 4.22: Normalized macro results for the seven dimensions in 6000 sentences.

We observed that for dimensions one, two and three there were noticeable differences, although for the rest of the dimensions these differences were less clear.

We also obtained the macro results (see Table 4.22) when aggregating the values for all the dimensions in all sentences and normalized the results using the previously obtained minimum and maximum values so that all values are in the range [0,1].

In Table 4.21 we noticed that the first dimension in our Twitter data set obtained a minimum value much lower than the one obtained when using PubMed data, which surprisingly characterizes our sample of tweets as more informational than the sample of PubMed texts. Twitter data set also got a maximum value higher than the value obtained in PubMed texts, denoting that Twitter data set minimum and maximum values are probably caused by outliers as the mean values shown in Table 4.21 as well as the normalized results in Table 4.22 show that dimension 1 is not so different between PubMed and Twitter texts, and also confirm that Twitter texts are more informational that PubMed sentences.

The narrative dimension of the texts (dimension 2) also shows higher maximum values for our Twitter data set, but when inspecting the mean value (Table 4.12), and the normalized value in Table 4.22 we can confirm that in this case it is clear that some outlier caused that maximum value, and our PubMed sample can be characterized as being more narrative than the Twitter data set. Dimension three also shows a similar behaviour as the maximum value in our sample of tweets is greater, but the mean and

# Dim.	Twitter Mean	Twitter CI (min)	Twitter CI (max)	PubMed Mean	PubMed CI (min)	PubMed CI (max)
1	-8.098	-8.101	-8.095	-9.114	-9.116	-9.111
2	0.028	0.028	0.029	0.032	0.032	0.032
3	0.083	0.083	0.083	0.107	0.107	0.107
4	0.018	0.018	0.018	0.007	0.007	0.007
5	0.002	0.002	0.002	0.004	0.004	0.004
6	0.006	0.006	0.006	0.005	0.005	0.005
7	0.000	0.000	0.000	0.000	0.000	0.000

TABLE 4.23: Mean, minimum and maximum confidence intervals (CI) micro results for the seven dimensions from Twitter and PubMed using Monte Carlo sampling (6000 sentences).

# Dim.	Twitter	PubMed
1.Involved versus Information Productions	0.782	0.754
2.Narrative versus Non-Narrative Concerns	0.071	0.079
3.Explicit versus Situation-Dependent Reference	0.437	0.480
4.Overt Expressions of Persuasion	0.110	0.043
5.Abstract versus Non-Abstract Information	0.020	0.046
6.On-Line Informational Elaboration	0.061	0.045
7.Academic qualification	0.003	0.005

TABLE 4.24: Normalized macro results for the seven dimensions in 6000 sentences using Monte Carlo sampling.

the results shown in Table 4.22 confirm that the set of PubMed sentences we selected contains more explicit texts. The same situation can be observed in dimension five too.

To evaluate the possible differences caused by the variability of the sentences we ran the same tests using a Monte Carlo sampling approach, retrieving 2/3 of the annotated sentences (4000 sentences) chosen at random from each data set in each run. We run 100 of these experiments and got the results presented in Table 4.23 and Table 4.24. These results show the mean and the standard deviation values using the similarity scores obtained in each of the 100 runs. We also use those 100 results to compute the confidence intervals at 95%.

As a by-product needed to compute the values shown in Table 4.21 and Table 4.22 we obtained the values for each factor, presented in Table 4.25.

The micro results are almost the same in all cases, although we can see that the macro results obtained when running Monte Carlo sampling show some differences that we did not observe before.

The differences at factor level can be observed in the following Table:

#	Factor Name	Twitter Mean (sd)	PubMed Mean (sd)	Ratio Mean (sd)
56	Private Verbs	0.012 (0.025)	0.007 (0.016)	1.846 (1.576)
60	Subordinator That Deletion	0.004 (0.015)	0.001 (0.007)	3.436 (2.078)

59	Contractions	0.015 (0.028)	0.000 (0.001)	1384.926 (46.693)
3	Present Verbs	0.053 (0.051)	0.016 (0.028)	3.372 (1.846)
7	Second Person Pronouns	0.009 (0.027)	0.000 (0.000)	- (-)
12	Pro Verb Do	0.004 (0.015)	0.000 (0.003)	24.604 (5.294)
67	Analytic Negation	0.010 (0.023)	0.002 (0.010)	4.272 (2.317)
10	Demonstrative Pronouns	0.003 (0.013)	0.002 (0.008)	2.003 (1.594)
49	Emphatics	0.009 (0.023)	0.002 (0.010)	4.154 (2.194)
6	First Person Pronouns	0.055 (0.060)	0.003 (0.012)	15.962 (5.044)
9	Pronoun It	0.009 (0.022)	0.001 (0.006)	10.748 (3.578)
19	Be as Main Verb	0.010 (0.023)	0.006 (0.017)	1.511 (1.344)
35	Causative Adverbial Subordinator (Because)	0.001 (0.008)	0.000 (0.003)	5.687 (2.715)
50	Discourse Particles	0.000 (0.005)	0.000 (0.000)	- (-)
11	Indefinite Pronoun	0.005 (0.018)	0.000 (0.002)	75.347 (10.560)
47	Hedges	0.001 (0.009)	0.000 (0.002)	19.893 (4.191)
48	Amplifiers	0.001 (0.009)	0.000 (0.005)	2.762 (1.963)
34	Sentence Relatives	0.000 (0.003)	0.001 (0.005)	3.017 (1.481)
13	Wh Questions	0.001 (0.007)	0.000 (0.001)	115.244 (13.955)
52	Possibility Modals	0.004 (0.016)	0.002 (0.011)	1.757 (1.461)
65	Independent Clause Coordination	0.000 (0.003)	0.000 (0.002)	1.232 (1.486)
23	Wh Clauses	0.001 (0.008)	0.000 (0.002)	7.386 (3.096)
61	Stranded Prepositions	0.001 (0.008)	0.000 (0.002)	16.666 (4.840)
16	Total Other Nouns	0.008 (0.023)	0.040 (0.042)	4.857 (1.831)
44	Word Length	3.984 (1.279)	5.429 (1.001)	1.363 (1.277)
39	Prepositional Phrases	0.082 (0.060)	0.128 (0.060)	1.564 (1.004)
43	Type/Token Ratio	10.881 (3.179)	10.841 (2.923)	1.004 (1.088)
40	Attributive Adjectives	0.051 (0.054)	0.100 (0.068)	1.933 (1.257)
1	Past Verbs	0.022 (0.038)	0.030 (0.033)	1.353 (1.148)
8	Third Person Personal Pronouns No It	0.006 (0.021)	0.002 (0.011)	3.120 (1.996)
2	Perfect Aspect Verbs	0.004 (0.014)	0.003 (0.011)	1.440 (1.296)
55	Public Verbs	0.003 (0.012)	0.003 (0.012)	1.185 (1.051)
66	Synthetic Negation	0.002 (0.009)	0.001 (0.008)	1.126 (1.102)
25	Present Participial Clauses	0.002 (0.011)	0.001 (0.005)	2.549 (2.287)
32	WH relative Clauses On Object	0.000 (0.003)	0.000 (0.001)	2.856 (1.937)
33	Pied-Piping Relative Clauses	0.000 (0.002)	0.000 (0.003)	5.109 (2.056)
31	WH relative Clauses On Subject	0.001 (0.007)	0.001 (0.007)	1.258 (1.054)
64	Phrasal Coordination	0.007 (0.021)	0.016 (0.026)	2.350 (1.257)
14	Nominalizations	0.295 (0.129)	0.307 (0.079)	1.042 (1.638)
5	Time Adv	0.008 (0.021)	0.001 (0.007)	6.271 (2.814)
4	Place Adv	0.001 (0.009)	0.001 (0.005)	2.475 (1.723)

42	Total Adverbs	0.046 (0.055)	0.020 (0.032)	2.281 (1.730)
24	Infinitives	0.014 (0.027)	0.006 (0.016)	2.100 (1.716)
54	Predictive Modals	0.004 (0.015)	0.000 (0.004)	11.464 (4.092)
57	Suasive Verbs	0.001 (0.008)	0.002 (0.009)	1.424 (1.052)
37	Conditional Adverbial Subordinator (If/Unless)	0.003 (0.013)	0.000 (0.002)	18.865 (5.180)
53	Necessity Modals	0.001 (0.009)	0.000 (0.005)	2.928 (1.811)
63	Split Auxiliaries	0.006 (0.018)	0.001 (0.008)	4.418 (2.245)
45	Conjuncts	0.001 (0.007)	0.003 (0.011)	2.890 (1.500)
17	Agentless Passives	0.000 (0.005)	0.000 (0.000)	- (-)
26	Past Participial Clauses	0.000 (0.005)	0.001 (0.004)	1.492 (1.033)
18	By Passives	0.000 (0.000)	0.000 (0.000)	- (-)
27	Past Participial WHIZ	0.001 (0.008)	0.005 (0.016)	4.292 (1.876)
38	Other Adverbial Subordinators	0.001 (0.009)	0.001 (0.006)	1.290 (1.408)
21	That Verb Complements	0.001 (0.008)	0.002 (0.009)	2.196 (1.232)
51	Demonstratives	0.010 (0.023)	0.006 (0.015)	1.671 (1.516)
30	That Relative On Object Position	0.000 (0.005)	0.000 (0.003)	2.105 (1.793)
22	That Adjective Complements	0.000 (0.003)	0.000 (0.002)	1.445 (1.395)
58	Seem/Appear Verbs	0.000 (0.004)	0.001 (0.005)	1.571 (1.219)

TABLE 4.25: Table showing the results for each factor used to compute Biber's features using the sample of 6000 sentences. Mean values and Standard deviation values for Twitter and PubMed are shown in Columns 3 and 4 respectively. The last column shows the mean and standard deviation ratios using the values from the previous 2 columns.

To address these issues, and aiming at a better understanding of the data sets, we used the 1000 sentences that two expert pharmacists filtered out. These 1000 sentences are sentences on the drug use which also include mentions of a disease or symptom related to the drug intake from either PubMed and Twitter. These sentences are the set of sentences included in TwiMed corpus.

By using the 1000 sentences selected by the two pharmacists we obtained the micro (Table 4.26) and macro (Table 4.27) results.

It becomes clear that the differing values between Twitter and Pubmed results presented in Table 4.21 became closer in Table 4.26, noticeable after pruning some of the outliers such as the minimum and maximum values for dimension 1 in Twitter data set appearing in Table 4.21. Table 4.26 now shows a less informational value for our sample of tweets than the one appearing in Table 4.21, while the value for our PubMed sentences is almost the same in both cases.

# Dim.	Twitter min	Twitter max	Twitter mean (std)	PubMed min	PubMed max	PubMed mean (std)
1	-12.402	-3.388	-8.443 (1.460)	-14.898	-4.682	-8.997 (1.481)
2	0.000	0.200	0.024 (0.034)	0.000	0.155	0.023 (0.029)
3	-0.143	0.356	0.084 (0.062)	-0.028	0.262	0.111 (0.038)
4	0.000	0.163	0.017 (0.024)	0.000	0.117	0.007 (0.014)
5	0.000	0.039	0.002 (0.006)	0.000	0.057	0.003 (0.008)
6	0.000	0.083	0.005 (0.012)	0.000	0.074	0.004 (0.010)
7	0.000	0.027	0.000 (0.002)	0.000	0.029	0.000 (0.002)

TABLE 4.26: Minimum, maximum, mean and standard deviation micro results for the seven dimensions using 1000 sentences from Twitter and PubMed.

# Dim.	Twitter	PubMed
1.Involved versus Information Productions	0.561	0.513
2.Narrative versus Non-Narrative Concerns	0.118	0.114
3.Explicit versus Situation-Dependent Reference	0.454	0.510
4.Overt Expressions of Persuasion	0.104	0.044
5.Abstract versus Non-Abstract Information	0.033	0.060
6.On-Line Informational Elaboration	0.064	0.053
7.Academic qualification	0.007	0.005

TABLE 4.27: Normalized macro results for the seven dimensions in 1000 sentences.

Dimension 2 is particularly interesting as in Table 4.21 we observed that the tweets we used were more narrative, but after filtering the data set we got that the sample of PubMed sentences was more narrative than the set of tweets (see Table 4.26).

If we focus on the mean values presented in Table 4.26 we can see that the results for the tweets are lower than the results presented in Table 4.21, and the only exceptions appear in dimension three, where the mean value is slightly greater, and in dimensions five and seven, where the mean values are the same in Table 4.21 and Table 4.26.

In the case of PubMed, the results do not change much although we can see the same trend of smaller mean scores in this case too. Dimension three is the only dimension having a greater mean value in Table 4.26 than in Table 4.21, and we can also see that dimensions four and seven get the same mean results in both tables.

When inspecting Table 4.22 and Table 4.27 we see that the highest value for Twitter and PubMed in each dimension remains the same in all dimensions except for dimensions two and seven, where PubMed scored higher in Table 4.22, but after filtering the data Twitter scored higher. The differences are minimal in both cases and the resulting scores for those dimensions are very small, showing that the narrativeness and the academic qualification of these texts did not vary significantly when filtering the data.

The filtering process helps in observing that for our PubMed sample the changes were very scarce, but in the case of the used tweets we noticed that dimensions one and two had values closer to the values in PubMed texts. This proves that even when the

# Dim.	Twitter Mean	Twitter CI (min)	Twitter CI (max)	PubMed Mean	PubMed CI (min)	PubMed CI (max)
1	-8.440	-8.440	-8.430	-9.000	-9.000	-8.990
2	0.020	0.020	0.020	0.020	0.020	0.020
3	0.080	0.080	0.080	0.110	0.110	0.110
4	0.020	0.020	0.020	0.010	0.010	0.010
5	0.000	0.000	0.000	0.000	0.000	0.000
6	0.010	0.010	0.010	0.000	0.000	0.000
7	0.000	0.000	0.000	0.000	0.000	0.000

TABLE 4.28: Mean, minimum and maximum confidence intervals (CI) micro results for the seven dimensions from Twitter and PubMed using Monte Carlo sampling (1000 sentences).

mentions are on the drugs, there are still some differences and by further filtering and fixing on a more concrete sub-domain, i.e. drug use and related symptoms and diseases, we can discard a number of outliers and get data sets that become more similar, which is also in line with previous researchers' findings where it was suggested that both the domain of knowledge and also the sub-domain within the domain should be controlled [188].

We proceeded to focus on Table 4.27, but we did not find substantial discrepancies in any of the dimensions although there were some differences in *"Explicit versus Situation-Dependent Reference"* dimension, *"Overt Expressions of Persuasion"* dimension and *"Abstract versus Non-Abstract Information"* dimension, showing that tweets used more expressions of persuasion, were more situation dependent and conveyed less abstract information. Importantly, the scores obtained for Twitter and PubMed data sets were not too different showing that the information in Twitter was not so different from the information in other formal sources, which is in line with other findings [189].

In order to confirm our insights on the subset of 1000 sentences we decided to run the same experiments using a Monte Carlo sampling strategy extracting 2/3 of the sentences (i.e. 750 sentences) chosen at random. We run 100 experiments obtaining the Table 4.28 and Table 4.29. We can see that in all cases the mean is very close to the minimum and maximum values obtained for the confidence interval, which was obtained for the 95%.

We can see that the results presented in Table 4.26 are in line with the results shown in Table 4.21, and the most different dimension is dimension one, followed by dimensions three and four. The mean values are somewhat similar in both cases although we can see that some differences are more clear in this table such as the ones shown in dimensions four ("Overt Expressions of Persuasion") and dimension six ("On-Line Informational Elaboration"), clearly understanding that the tweets have more traits related to on-line productions and also use more expressions of persuasion.

# Dim.	Twitter	PubMed
1.Involved versus Information Productions	0.430	0.400
2.Narrative versus Non-Narrative Concerns	0.120	0.110
3.Explicit versus Situation-Dependent Reference	0.450	0.510
4.Overt Expressions of Persuasion	0.100	0.040
5.Abstract versus Non-Abstract Information	0.030	0.060
6.On-Line Informational Elaboration	0.060	0.050
7.Academic qualification	0.010	0.010

TABLE 4.29: Normalized macro results for the seven dimensions in 1000 sentences using Monte Carlo sampling.

Table 4.29 is very similar to Table 4.27[7], although in this case too the results have been obtained using a Monte Carlo sampling approach. The normalized results maintain the same trend, and one interesting finding appears now as we can see that the results for "Academic qualification" are the same.

Besides the global differences in the dimensions we also studied the differences in the factors used to compute each dimension's score observing that although dimension 1 was not very dissimilar between data sets its factors contained some of the most different results between PubMed and Twitter texts. In particular the use of the verb "Do" as a pro-verb appeared 75 times more often in Twitter sentences than in our PubMed texts, and the use of "First person" pronouns and "It" pronouns appeared 15 and 13 times more often in Twitter than in PubMed data set. The use of the analytic negation ("not") and the adverbial subordinator "because" appeared 8 and 9 more times in Twitter than in PubMed, respectively. Not very surprisingly, the use of emphatics ("so", "such", "a lot"...), amplifiers ("absolutely", "totally"...) and stranded prepositions also appeared more than 5 times more often in Twitter than in PubMed in our sample of sentences.

Apart from dimension 1, the use of "Time adverbials", a factor used to compute the final values for dimension 3, was used more than five times in Twitter sentences. The remaining factors getting counts five times greater in sentences from one source versus the other were also more popular in Twitter, and these are the "Predictive modals" and "Conditional adverbial subordinators" ("if" and "unless") which were two factors taking part on dimension four.

The use of sentence relatives (e.g. "which"), was more than twice as frequent in PubMed sentences than in tweets, and the use of nouns appeared almost 4 times more often in PubMed, which is a clear sign of specialized discourse [190]. Those two counts were used to compute dimension 1, while for dimension 3 the count of the phrasal coordination ("and") was almost three times greater in PubMed than in Twitter. In dimension 4, the use of suasive verbs ("agree", "allow", "arrange"...) was twice as frequent in

[7] although the ranges vary because of the limits used in the normalization process

PubMed than in Twitter sentences, and the use of past participial WHIZ[8], in dimension 5, appeared three times more often in PubMed data set.

The results for the factors can be seen in Table 4.30, where the factors have been grouped into the seven dimensions used in Biber's study.

#	Factor Name	Twitter Mean (sd)	PubMed Mean (sd)	Ratio Mean (sd)
56	Private Verbs	0.011 (0.024)	0.006 (0.014)	2.009 (1.679)
60	Subordinator That Deletion	0.003 (0.013)	0.001 (0.007)	3.015 (1.852)
59	Contractions	0.014 (0.026)	0.000 (0.000)	- (-)
3	Present Verbs	0.052 (0.050)	0.017 (0.027)	3.034 (1.868)
7	Second Person Pronouns	0.009 (0.026)	0.000 (0.000)	- (-)
12	Pro Verb Do	0.003 (0.013)	0.000 (0.001)	75.982 (9.854)
67	Analytic Negation	0.010 (0.023)	0.001 (0.007)	8.351 (3.377)
10	Demonstrative Pronouns	0.003 (0.012)	0.002 (0.008)	2.124 (1.575)
49	Emphatics	0.009 (0.020)	0.002 (0.008)	5.540 (2.416)
6	First Person Pronouns	0.053 (0.057)	0.003 (0.012)	15.165 (4.829)
9	Pronoun It	0.009 (0.020)	0.001 (0.005)	13.030 (3.637)
19	Be as Main Verb	0.008 (0.020)	0.007 (0.017)	1.159 (1.152)
35	Causative Adverbial Subordinator (Because)	0.001 (0.006)	0.000 (0.002)	9.635 (3.731)
50	Discourse Particles	0.000 (0.004)	0.000 (0.000)	- (-)
11	Indefinite Pronoun	0.004 (0.016)	0.000 (0.000)	- (-)
47	Hedges	0.001 (0.008)	0.000 (0.000)	- (-)
48	Amplifiers	0.001 (0.011)	0.000 (0.003)	6.528 (3.935)
34	Sentence Relatives	0.000 (0.004)	0.001 (0.005)	2.582 (1.199)
13	Wh Questions	0.001 (0.006)	0.000 (0.000)	- (-)
52	Possibility Modals	0.005 (0.017)	0.003 (0.013)	1.629 (1.321)
65	Independent Clause Coordination	0.000 (0.004)	0.000 (0.002)	1.612 (1.929)
23	Wh Clauses	0.000 (0.004)	0.000 (0.002)	3.923 (2.102)
61	Stranded Prepositions	0.001 (0.005)	0.000 (0.002)	5.375 (2.920)
16	Total Other Nouns	0.011 (0.025)	0.044 (0.044)	3.935 (1.757)
44	Word Length	4.185 (1.095)	5.610 (1.005)	1.341 (1.090)
39	Prepositional Phrases	0.083 (0.058)	0.142 (0.060)	1.719 (1.028)
43	Type/Token Ratio	11.277 (3.062)	10.393 (3.129)	1.085 (1.022)
40	Attributive Adjectives	0.061 (0.058)	0.116 (0.074)	1.895 (1.274)
1	Past Verbs	0.018 (0.033)	0.020 (0.029)	1.098 (1.144)
8	Third Person Personal Pronouns No It	0.004 (0.017)	0.002 (0.009)	2.279 (1.919)
2	Perfect Aspect Verbs	0.004 (0.014)	0.003 (0.011)	1.159 (1.194)
55	Public Verbs	0.003 (0.011)	0.003 (0.012)	1.278 (1.007)
66	Synthetic Negation	0.002 (0.010)	0.001 (0.006)	1.925 (1.521)

[8]Past participle combined with the deletion of a Wh-word plus a form of be, quite often *"is"*, thus called *"whiz"* as a monosyllabic variant of *"Wh-is deletion"*.

25	Present Participial Clauses	0.002 (0.011)	0.001 (0.005)	2.670 (2.232)
32	WH relative Clauses On Object	0.000 (0.003)	0.000 (0.002)	2.365 (1.689)
33	Pied-Piping Relative Clauses	0.000 (0.000)	0.000 (0.002)	- (-)
31	WH relative Clauses On Subject	0.001 (0.006)	0.001 (0.008)	1.840 (1.245)
64	Phrasal Coordination	0.006 (0.018)	0.017 (0.026)	2.890 (1.472)
14	Nominalizations	0.293 (0.124)	0.313 (0.083)	1.066 (1.491)
5	Time Adv	0.006 (0.018)	0.001 (0.007)	5.630 (2.771)
4	Place Adv	0.001 (0.007)	0.001 (0.006)	1.337 (1.290)
42	Total Adverbs	0.044 (0.050)	0.015 (0.027)	2.916 (1.873)
24	Infinitives	0.014 (0.027)	0.007 (0.016)	2.015 (1.685)
54	Predictive Modals	0.003 (0.011)	0.000 (0.003)	10.969 (3.860)
57	Suasive Verbs	0.001 (0.006)	0.002 (0.009)	2.063 (1.352)
37	Conditional Adverbial Subordinator (If/Unless)	0.003 (0.011)	0.000 (0.002)	19.595 (4.948)
53	Necessity Modals	0.001 (0.006)	0.000 (0.005)	1.523 (1.214)
63	Split Auxiliaries	0.006 (0.017)	0.001 (0.008)	4.990 (2.238)
45	Conjuncts	0.001 (0.005)	0.001 (0.007)	2.068 (1.356)
17	Agentless Passives	0.000 (0.005)	0.000 (0.000)	- (-)
26	Past Participial Clauses	0.000 (0.005)	0.000 (0.004)	1.096 (1.123)
18	By Passives	0.000 (0.000)	0.000 (0.000)	- (-)
27	Past Participial WHIZ	0.002 (0.009)	0.006 (0.017)	3.453 (1.963)
38	Other Adverbial Subordinators	0.001 (0.008)	0.001 (0.006)	1.555 (1.389)
21	That Verb Complements	0.001 (0.007)	0.002 (0.009)	1.748 (1.199)
51	Demonstratives	0.008 (0.019)	0.006 (0.014)	1.427 (1.335)
30	That Relative On Object Position	0.000 (0.005)	0.000 (0.002)	3.967 (2.704)
22	That Adjective Complements	0.000 (0.004)	0.000 (0.003)	1.630 (1.393)
58	Seem/Appear Verbs	0.001 (0.005)	0.000 (0.005)	1.234 (1.093)

TABLE 4.30: Table showing the results for each factor used to compute Biber's features using the sample of 1000 sentences. Mean values and Standard deviation values for Twitter and PubMed are shown in Columns 3 and 4 respectively. The last column shows the mean and standard deviation ratios using the values from the previous 2 columns.

For the task of identifying sentences containing drug use reports we can think that those features being similar across tweets and PubMed texts are expected to help on the task, although their impact may not be clear for now. Bearing that idea in mind we can find in Table 4.30 that the use of nominalizations, the type/token ratio, the use of past participial clauses and the use of past verbs show that the counts for those elements have a very similar ratio. The next elements having similar ratios are different types of verbs ("Be", seem/appear and public verbs) which can be an evidence indicating that those verbs are used to report drug use.

These results confirm that although Biber's approach is able to capture a number of register differences in PubMed and Twitter sentences, that framework cannot completely describe all the differences between those registers, confirming that **Hypothesis 3** should be accepted.

4.4 Discussion

We characterized the underlying information via LDA topics and noticed that the number of clusters in both data sets were the same, but the realm of the keywords composing those clusters were very different as in the case of PubMed we mainly obtained medical related terms while in the case of Twitter we obtained keywords such as generic verbs ("take", "get", "feel") and abbreviations ("im") that were not useful for describing a medical related set of tweets as the one we used. The use of those verbs is an important finding as those are verbs known to be used in an informal setting as opposed to other formal verbs [191], and also because that is an element that is not taken into consideration in Biber's MD analysis, meaning that we should extend that framework to account for those variations.

We also assessed the similarity in terms of the different relations between the drug and the related effects (symptoms and diseases) and observed a very low level of similarity between PubMed sentences and tweets containing drug-use reports. We observed that, in particular, the reports of negative outcomes were very far from being comparable, and only the relations containing positive effects (i.e. "Benefit" and "Reason-to-use" relations) had some small similarity (although below 30%). These facts would require an in-depth study to confirm our findings, as in case the similarity ratio does not improve when using a much larger dataset it would mean that reports from PubMed and Twitter contain complementary information, which is a not a surprising finding but to our knowledge this is an observation that has not been presented before.

The comparison on the features between different tweet data sets showed that drug-use tweets used more often hedging, amplifiers ("absolutely", "altogether", "completely"...), sentence relatives ("which"), and past participial WHIZ and seem/appear verbs. These are features that we would expect to see more often in formal texts. On the other hand, generic tweets used more stranded prepositions, discourse particles ("well", "anyway"...), place adverbs ("above", "around"...), emphatics ("just", "really"...) and "wh clauses" (e.g. "I believed **what** he told me"). These features show that generic tweets use a number of traits expected to appear in informal texts. In terms of the dimensions we did not expect to see that four of them were different when comparing both data sets, and in particular, we did not expect to see that the differences showed that the set

of drug-related tweets contained traits expected to be seen in academic texts by being less involved and less narrative than generic tweets, while also being more abstract and containing more academic qualifications.

When comparing drug use reports in Twitter and PubMed we saw that these were not very dissimilar and also noticed that there were some features, namely the use of the verb "do", "first person" and "It" pronouns, the analytic negation ("not"), the adverbial subordinator "because", "predictive modals" and "conditional adverbial subordinators" that were more frequent in tweets than in PubMed sentences. The use of emphatics and stranded prepositions also appeared more often in tweets which is in line with informal register traits [8]. In the case of PubMed, sentence relatives, nouns, "and" coordination, suasive verbs, and the use of past participial WHIZ were the features that appeared noticeably more often than in tweets. These features are know for signalling the use of more complex texts [192].

Comparing the similar features in tweets and PubMed sentences containing drug-use reports showed that, among other features, the use of nominalizations, past participial clauses and the type/token ratio were very similar in drug-use reports in Twitter and PubMed, which in turn means that these particular features are related to the information being conveyed independently of the type of linguistic register that is being used. In terms of the dimensions themselves we found that the levels of narrativeness and the use of academic qualifications were not clearly different between these two date sets, and those similarities could be helpful in detecting sentences from either formal or informal sources containing drug-use reports.

The main result from the study presented in this chapter is the evidence that Biber's MD analysis is able to capture differences and similarities related to the linguistic register and tell apart drug-use reports in Twitter from generic tweets as well as to discover the linguistic features that are most similar in drug-use reports in Twitter and PubMed. We also found that there are other sets of features that are not included on Biber's analysis, such as the use of informal verbs, that differ in the drug-use reports we may find in Twitter and PubMed. We believe Biber's features, either in their raw form or expanded to increase the coverage, as well other politeness features (e.g. to account for the use of informal verbs) have potential for contribution in pharmacovigilance. Those are findings that can be applied to classification and NER systems as a way to validate the gains produced when using these features. Our empirical analyses using these new features are presented in the following chapter 5.

Chapter 5

Exploring the register in pharmacovigilance systems

This chapter presents the different experiments we performed using the data sets described in Chapter 3.

Our first experiments tried to answer the question of whether we could build a binary classifier able to detect first-hand drug use reports in Twitter. We describe this novel task and the set of features we used when building our first binary classifier aimed at the detection of those messages in Twitter.

Following experiments build on top of the previous task as we improved our binary classifiers and besides working on the detection of first-hand experience report in Twitter we expand our goals to work on different tasks, also assessing the improvements produced by the newly added set of register-related features.

The final experiments show the contribution that the set of register-related features provide to a NER system when trying to identify drugs, diseases and symptoms.

5.1 A first approach to binary classification of first-hand experience reports in Twitter

This section contains the description of our first work aimed at detecting tweets containing reports on the drug use where the person writing the messages is also the person taking the drug. Those reports are referred to as "first-hand experience reports".

The need for early detection of first-hand experience reports is of great importance for pharmacovigilance as the correct identification of these messages can help in monitoring

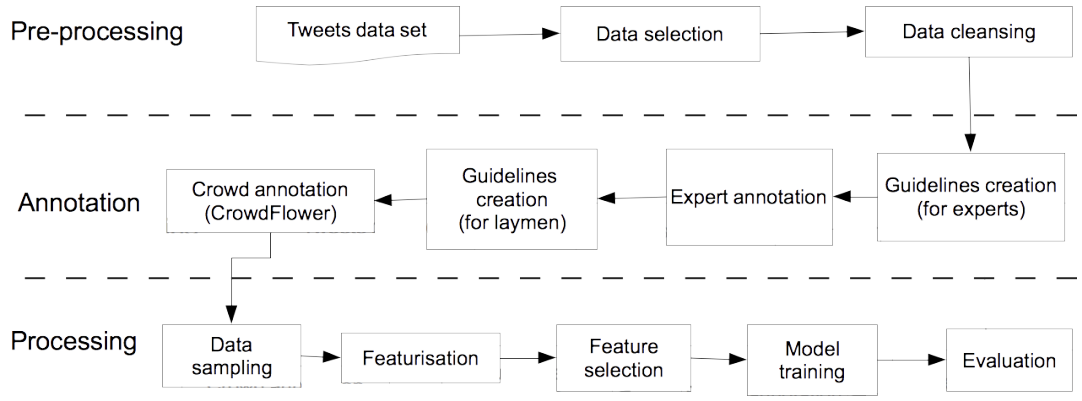

FIGURE 5.1: Flowchart detailing the phases in our study. In the Pre-processing phase we obtain data from Twitter, extracted the tweets mentioning the drugs, and performed data cleansing. In the Annotation phase we performed both annotations (by experienced annotators and by laymen annotators). In the Processing phase we perform the featurisation and feature selection, used to train the models. Finally, the obtained models are evaluated.

the use of the drugs and related outcomes, while also telling apart certain messages that are not of interest, i.e marketing campaigns. After noticing the need for these systems and finding that this is an area that has not been widely explored we present our work on this field.

5.1.1 Methods

Using the gold data we presented in section 3.1 we divided the sentences randomly into training (2/3) and testing (1/3) sets, with 600 and 299 tweets respectively. The whole process is depicted in Figure 5.1

We found 356 tweets classified as first-hand experience tweets in the gold standard. This is 39.6% of the 899 tweets.

Having the data in place, we generated the features: n-grams, latent topics, orthographic features and other Twitter specific features (see Figure 5.2 example). We used several linguistic feature types including character 1,2,3- grams, e.g. 'za','oz'; word tokens, e.g. 'dies', bucketed message length in tokens, e.g. 10-20; topics (topic1, topic2...); and Twitter specific features (to check whether the tweet is addressed to someone by using the *"@"* sign, to check whether the tweet may want to stress something in particular by using the *"#"* sign...).

Previous research [193] showed that combining n-grams with other semantic features improves classification accuracy. In our approach, we did not use the raw n-grams (character n-grams, unigrams and bigrams). Instead we applied term frequency inverse document frequency (TF-IDF) weighting first.

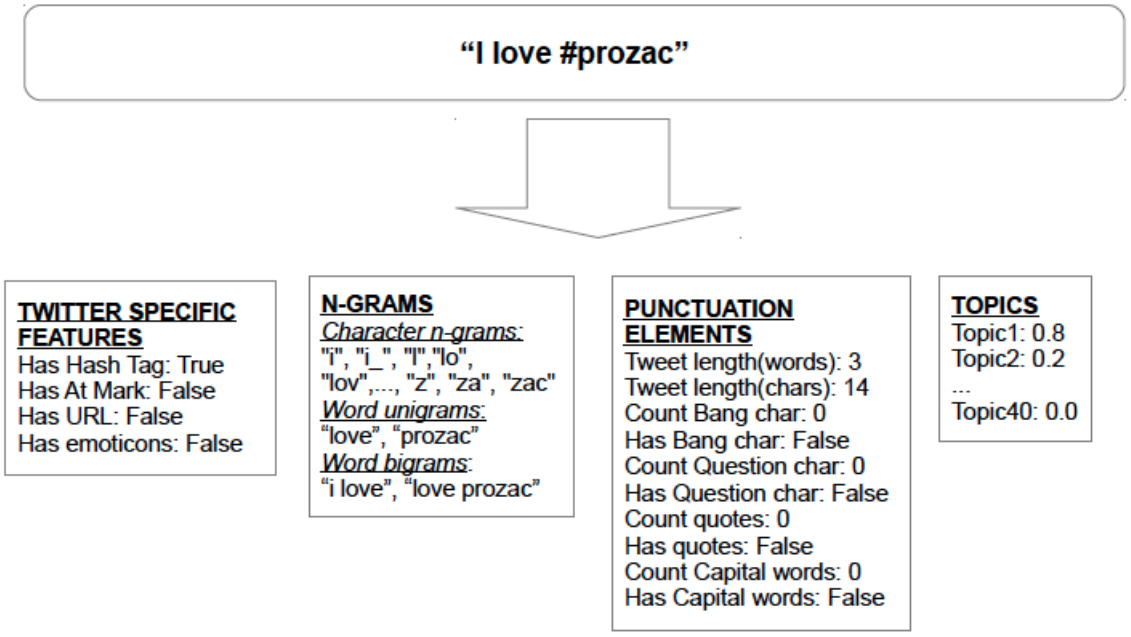

FIGURE 5.2: Example of a featurised tweet including n-grams, latent topics, orthography and hashtags. Here we show the values of some of the features. In the case of the N-Grams we only show the obtained n-grams for such tweet.

There is clear evidence that Latent Dirichlet Allocation (LDA) topic models provide valuable data with large text corpora and we decided to add it to our study based on recent studies that have shown its value for collections of Twitter messages [194], [195]. The topics were discovered using LDA [176] on the training set, and as we had 11 groups of drugs but one of them had no matches (Table 3.1) we selected 40 topics corresponding to an even distribution across the tweets. We experimented with different number of topics (from 35 to 45), but the information gain method consistently reported that the LDA topics did not contribute as features. Further investigation would be required to confirm whether automated topic modelling could improve the accuracy of our system. If this is confirmed a natural next step would be to provide a semantic label for each topic [196].

After generating all the features we applied information gain as the feature reduction algorithm to obtain the best ranking features, given that it has superior performance over other feature reduction methods [197]. At that point we observed that the topics were not contributing as well as expected from previous studies [198], and we decided to concentrate on the top-ranking features discarding the use of LDA topics. We also observed that the bigrams were not listed as top-ranking features. This can be explained because our word bigrams had low frequency counts and indicates that it is better to focus on character n-grams where frequency is higher.

We used R's FSelector package [199] to calculate information gain. The algorithms

Ranking	List of features
Top 10 features (features ranked 1-10)	*"my"*, *"za"*, *"zac"*, *"oza"*, *"roz"*, *"oz"*, *"ric"* (Character n-grams); *"Has hash tag"*, *"Has at mark"*, *"Has emoticons"*.
10 features at the middle of the list (features ranked 313-322)	*"fill"*, *"filming"*, *"find"*, *"finding"*, *"findworkfamilylifebalance"*, *"fine"*, *"firsttestofthesemester"*, *"flip"*, *"flowing"*, *"flvs"* (unigrams).
10 features at the bottom of the list (features ranked 628-637)	*"phenergan"*, *"phenidaad"*, *"phillywcwagon"*, *"phoebebuffay"*, *"pib"*, *"pill"*, *"pizza"*, *"placenta"*, *"planet"*, *"plenty"* (unigrams).

TABLE 5.1: Sample of extracted features using 10% information gain.

finds weights for discrete attributes based on their correlation with the continuous class attribute. The formula it uses is the following, where it takes into consideration the entropies (represented by *"H"*) of the class and the attribute:

$$Information\ Gain = H(Class) + H(Attribute) - H(Class, Attribute) \quad (1)$$

As shown in Table 5.1 we applied information gain to obtain the most discriminating 1%, 3%, 5%, 7%,10%, 50%, and 100% features. When using 10% features we fed our models with 637 features. A sample of the obtained features is presented in Table 5.1.

5.1.2 Evaluation

We then trained and tested C50, SVM using a linear kernel (SVM), Naive Bayes (NB), Multi-Layer Perceptron (MLP), a logistic regression model (GLM)[1], and a logistic regression model that uses bayesian functions with independent Cauchy prior distribution for the coefficients (BGLM)[2] from R's Caret package [200] to assess their performance on our data sets using the selected set of features.

For the evaluation of the results we use the F-Score, based on the standard precision and recall:

[1] We use the term "GLM" as in Caret's documentation it is referred to as "Generalized Linear Model".
[2] We use the term "BGLM" as in Caret's documentation it is referred to as 'Bayesian Generalized Linear Model".

$$F - Score = 2 * \frac{precision * recall}{precision + recall} \quad (2)$$

Where recall is:

$$Recall = \frac{true\ positives}{true\ positives\ +\ false\ negatives} \quad (3)$$

And precision is:

$$Precision = \frac{true\ positives}{true\ positives\ +\ false\ positives} \quad (4)$$

To better quantify the performance of the models we also include the Informedness measures [201]. The Informedness measure, apart from taking into account the *"true positive"*, *"false positive"* and *"false negative"* values that are used by the F-Score, uses the *"true negative"* values getting a fair measure for classification showing which are the most informative models and which are the models that even when obtaining high F-Score values do not have predictive power.

$$Informedness = recall + invRecall - 1 \quad (5)$$

Where inverse recall is:

$$invRecall = \frac{true\ negatives}{true\ negatives\ +\ false\ positives} \quad (6)$$

First evaluation using the initial data set

As shown in Table 5.2, combining the six learning models with the selected set of features gave a maximum F-Score of 0.64 when using CrowdFlower data. BGLM is the best performing model, followed by C50. GLM is the other model scoring above the baseline. The baseline, obtained by predicting all labels to be *"First-hand experience"*, achieves an F-score of 0.55. In this and following experiments the *"NaN"* value in the tables indicate that all predicted labels were *"Other genre"*. Here we can see that BGLM is the most informative model, followed by GLM.

Second evaluation using the initial data set

For our next experiment we asked an expert annotator (N.A.) to annotate the fields *"First-class experience"*, *"Tweet written in English language"*, and *"Tweet about the drug"* for the same 1548 tweets that the laymen annotated. After having these annotations we discarded all the tweets where the laymen and the expert disagreed on the

Model	1%	3%	5%	7%	10%	50%	100%
SVM	0.48	0.49	0.48	0.46	0.43	0.40	0.28
	(0.15)	*(0.20)*	*(0.18)*	*(0.15)*	*(0.13)*	*(0.12)*	*(0.00)*
C50	0.61	0.61	**0.61**	0.61	0.61	**0.57**	**0.55**
	(0.27)	*(0.27)*	*(0.27)*	*(0.27)*	*(0.27)*	*(0.10)*	*(0.00)*
GLM	0.59	0.57	0.54	0.53	0.56	0.40	0.42
	(0.38)	*(0.35)*	*(0.32)*	*(0.33)*	*(0.32)*	*(−0.06)*	*(0.01)*
MLP	0.47	0.52	0.42	0.37	0.44	0.44	0.47
	(0.11)	*(0.24)*	*(0.11)*	*(0.04)*	*(0.03)*	*(0.07)*	*(0.11)*
BGLM	**0.64**	**0.63**	**0.61**	**0.64**	**0.62**	0.55	0.54
	(0.43)	***(0.41)***	***(0.39)***	***(0.43)***	***(0.40)***	***(0.28)***	***(0.27)***
NB	0.13	0.10	0.02	NaN	NaN	NaN	NaN
	(0.04)	*(0.03)*	*(0.00)*	*(0.00)*	*(0.00)*	*(0.00)*	*(0.00)*

TABLE 5.2: F-score values for each model using a selected percentage of features on 899 tweets annotated via crowdsourcing. Note that figures in parentheses show the Informedness values. The highest values in each column are highlighted in bold.

Model	1%	3%	5%	7%	10%	50%	100%
SVM	0.15	0.09	0.14	0.13	0.09	0.27	0.20
	(−0.03)	*(−0.07)*	*(0.01)*	*(−0.03)*	*(−0.07)*	*(0.08)*	*(0.04)*
C50	0.24	0.52	0.39	0.28	0.43	0.48	0.47
	(0.13)	*(0.32)*	*(0.22)*	*(0.15)*	*(0.26)*	*(0.14)*	*(0.17)*
GLM	0.50	0.37	0.30	0.30	0.35	0.30	0.36
	(0.32)	*(0.19)*	*(0.15)*	*(0.15)*	*(0.17)*	*(−0.15)*	*(0.01)*
MLP	0.19	0.25	0.25	0.21	0.32	0.39	0.27
	(−0.06)	*(−0.03)*	*(0.03)*	*(−0.01)*	*(0.08)*	*(0.17)*	*(0.01)*
BGLM	**0.57**	**0.55**	**0.56**	**0.51**	**0.50**	**0.57**	**0.57**
	(0.40)	***(0.37)***	***(0.39)***	***(0.33)***	***(0.31)***	***(0.40)***	***(0.39)***
NB	0.21	NaN	NaN	NaN	NaN	NaN	0.41
	(0.09)	*(0.00)*	*(0.00)*	*(0.00)*	*(0.00)*	*(0.00)*	*(−0.03)*

TABLE 5.3: F-score values for each model using a selected percentage of features on 661 tweets annotated via crowdsourcing and by an expert. Note that figures in parentheses show the Informedness values. The highest values in each column are highlighted in bold.

annotation for those fields, obtaining the 661 tweets[3] that we used to run the same experiment from before. We present the results from this experiment in Table 5.3. In this case the baseline is also obtained when labelling all tweets as *"First-hand experiences"* and has 0.45 F-Score. In this experiment BGLM was the best model both in terms of F-score and Informedness.

Extended evaluation

During September 26th 2014 until December 9th 2014 we collected a new data set from Twitter by filtering the tweets containing any of the drug names or drug synonyms listed in Table 3.1 and Table 3.2. We gathered 159,007 tweets and chose 4000 tweets at random

[3] A modified version of this file complying with Twitter's TOS can be found on github https://github.com/nestoralvaro/JBI_Pharmacovigilance/tree/master/661_CrowdFlower_Expert.

Model	1%	3%	5%	7%	10%	50%	100%
SVM	0.58	0.49	0.29	0.35	0.48	0.55	0.27
	(0.26)	*(0.15)*	*(0.04)*	*(0.05)*	*(0.14)*	*(0.14)*	*(0.04)*
C50	0.21	0.14	0.11	0.15	0.29	**0.75**	0.09
	(0.09)	*(0.06)*	*(0.05)*	*(0.08)*	*(0.16)*	***(0.57)***	*(0.04)*
GLM	**0.77**	0.75	0.75	0.74	0.68	0.36	0.48
	(0.59)	***(0.57)***	*(0.56)*	*(0.54)*	*(0.47)*	*(0.02)*	*(0.01)*
MLP	0.63	0.65	0.54	0.53	0.56	NaN	0.56
	(0.33)	*(0.35)*	*(0.11)*	*(0.14)*	*(0.19)*	*(0.00)*	*(0.20)*
BGLM	**0.77**	**0.76**	**0.76**	**0.77**	**0.77**	0.68	**0.77**
	(0.59)	***(0.57)***	***(0.57)***	***(0.59)***	***(0.59)***	*(0.41)*	***(0.59)***
NB	0.63	0.62	0.64	0.66	NaN	NaN	NaN
	(0.09)	*(0.06)*	*(0.14)*	*(0.24)*	*(0.00)*	*(0.00)*	*(0.00)*

TABLE 5.4: F-score values for each model using a selected percentage of features on 3211 tweets annotated by two experts. Note that figures in parentheses show the Informedness values. The highest values in each column are highlighted in bold.

to be annotated by two experts using the same version of the guidelines. We obtained 3211 tweets where both expert annotators agreed on the annotation for the genre and which were written in English language and about the drugs of interest. We used that dataset[4] as the gold standard for our last experiment.

In this experiment we obtained a much larger number of feature values, and in order to process all of them (mainly because of computer memory limitations) we had to reduce the number of character n-grams and only keep those that appeared more than ten times. Apart from this change, the code we used for training and testing was the same that we used when we ran the experiments reported in the previous sections.

We present in Table 5.4 the F-Score results obtained for each model and each set of features. In this dataset the baseline prediction is obtained when labelling all tweets as *"First-hand experience"* tweets (0.61 F-Score). Here BGLM gets the highest F-Score and Informedness results for almost all sets of features.

The experiments presented in this section were the first in which we assessed the usefulness of the data we prepared in our data set of first-hand experience tweets while also trying to answer to the question of whether we could prepare a classifier able to detect tweets reporting personal experiences related to the drug use. We discovered that the classification systems we produced were able to tell apart some of those reports, although there is still room for improvement. That is why we continue to explore it, and for that we build stronger classifiers as presented in the following section.

[4] A modified version of this file complying with Twitter's TOS can be found on github https://github.com/nestoralvaro/JBI_Pharmacovigilance/tree/master/3211_Experts.

The next section also explores whether the use of register-related set of features can help in a number of classifiers, being the classifier targeting at the detection of first-hand experience tweets one of the systems we assess.

5.2 Binary classification systems using register information

Once we observed that there were some factors from Biber's study that were useful at identifying different types of reports independently of the type of register in which they were written we thought of exploring the gains that those features could provide to NLP systems. In particular we were interested on seeing if a classifier making use of those features could perform significantly better than a classifier not using those features as a way to test ***Hypothesis 4***.

In this section we present these experiments using different classifiers to identify certain types of sentences in both PubMed and Twitter, and assess the systems' performance and the impact that the set of formality features provide to these classifiers. An overview on the classifiers we prepared can be seen in Table 5.5

We continue our previous work by focusing on a classifier that detects first-hand experience reports from Twitter. We also prepare another three different binary classifiers using PubMed and Twitter sentences fed with the conflated annotations the we prepared in the data set described in 3.3. The goal of the first classifier is to detect sentences containing drug and outcomes (i.e. symptoms and/or diseases) mentions. That classifier can be thought as an expansion to the first-hand experience classifier given the author of the message is not the person who is actually taking the drug. That is of particular importance in the case of academic texts as the person contributing the articles to PubMed is not the drug user. By identifying sentences containing drugs and their outcomes in both PubMed and Twitter we can identify hot topics, monitor the use of drugs and detect off-label uses, which are the main areas of interest in pharmacovigilance.

The two remaining classifiers that we present in this section are specializations of the classifier we just introduced. One of these new classifiers focuses on detecting sentences containing positive outcomes related to the drug intake. The last classifier focuses on detecting sentences containing negative outcomes related to the drug intake. Detecting sentences containing outcomes is of particular interest, and properly telling apart sentences containing positive and negative outcomes is an important task *per se* because it allows the researchers to focus on more specific outcomes.

Classifier	Source	Positive case	Negative case	Description
First-hand experience	Twitter	*"to clarify, i am not against **Zoloft**. it just eventually stopped working for ME. everybody is different! it works for some people great!"*	*"by the way that tweet about the **adderall** was a joke I don't do things like that"*	Tweets where the person writing the report talks about a personal use of the drug.
Any outcome	Twitter	*"Man who needs **Zzquil** when I have **Singulair**? Within an hour, I'm zombie walking for the bed."*	*"Does anyone know if you can drink alcohol on fluoxetine???"*	Sentence containing any outcome related to the drug intake.
Any outcome	PubMed	*"This suggests that **topiramate** could attenuate the ongoing weight gain from **lithium** and **risperidone**."*	*"Positive small series and case reports have been reported for **lamotrigine**, **gabapentin** and **topiramate**."*	Sentences containing any outcome related to the drug intake.
Positive outcome	Twitter	*"If it weren't for my pal **seroquel** I wouldn't have gotten any sleep in months"*	*"I can't take **ritalin** after 6pm anymore..."*	Sentences containing a positive effect related to the drug intake.
Positive outcome	PubMed	*"Oral **melphalan** and **prednisone** remain an effective and tolerable treatment for patients with multiple myeloma."*	*"In addition, **levetiracetam** and **topiramate** are effective and can be use in combination or as second line treatment."*	Sentences containing a positive effect related to the drug intake.
Negative outcome	Twitter	*"I waited too long to take my **Zoloft** and now I have been feeling irritated all day. Everything is getting on my nerves."*	*"Haven't taken my **vyvanse** in 2 weeks. Todays going to be a pacey day"*	Sentences containing a negative effect related to the drug intake.
Negative outcome	PubMed	*"**Verapamil** also increased the area under the blood concentration time curve and the gastrointestinal toxicity of melphalan."*	*"Morphological analysis of the enamel organ in rats treated with **fluoxetine**."*	Sentences containing a negative effect related to the drug intake.

TABLE 5.5: Sample of the sentences used in the different classification systems.

5.2.1 Methods

When having the data ready the next step was the generation of features, for which we followed a standard approach and produced different set of features.

Besides assessing the set of features proposed by Biber, and given our observations we decided to expand on that set of features adding linguistic features related to the politeness in the sentences.

To give an overview of those features we list these groups, although the individual features composing each group are not included in the following list.

- **Textual features**: This set of features included, among other features, the count of the ellipsis ("..."), question marks, exclamation marks, the total length of the sentences and the mean length of the words in each sentence.
- **N-gram features**: This set of features kept the count of the different words appearing in the sentences. This feature included the count for the uni, bi and tri-grams (one word, two words and three words, respectively). At uni-gram level we did not include the stop words ("I", "he", "the", "any"...), and for uni, bi and tri-grams we did not include the n-grams that were rarely seen in our data set. That is, the n-grams appearing less than five times in total in the whole data set.

- **Emoticon features**: This feature used different emoticon lexicons including positive sentiment emoticons (":-)", ":D"...)[5], negative sentiment emoticons (":'(", ":-((")...)[6], playful emoticons (";)", ":-p"...)[7], unhealthy emoticons ("x-s", "X-O"...)[8], other emoticons (":o", "8-)"...)[9], and the complete list of emoticons listed in the Wikipedia[10] which covers most of the other lists and also providing an extended set of elements ("@ > −− > −−", ":-J"...).

- **One character n-gram features**: We used here different set of features composed of only one character and commonly seen in texts. The complete set of features included punctuation characters, one letter-upper case characters, one letter lower-case characters, one character digits and a list of commonly used word separators as are the blank, the tabulator or the new line character.

- **Pos features**: This set of features included the count for the part-of-speech tags used in the Penn Treebank Project[11].

- **Emoji features**: This was a new feature we wanted to explore even if it would be only useful when evaluating our set of tweets because to date no study in drug safety assessed the power of emoji features in a classifier. For this we used the emoji lexicons included in a very recent study[202] and available online[12].

- **Linguistic features**: This set of features included the count for each one of Biber's features as used in the assessments presented in Chapters 3 and 4.

- **Other features**: This set of features built on top of the set of features used on Biber's study by expanding the lists he used (e.g. by adding "zero" to the set of numeral keywords) and also adding new lists as for example new ordinal keywords ("1st", "2nd"...), necessity verbs ("need to", "have to") or a set of medical related keywords used in other pharmacovigilance studies [6] ("experience", "effective"...)[13].

- **Other features expanded**: This set of features included the same set of features used in **Other features**, although in this case we did not conflated the count for the different elements creating a sparse matrix with all the features.

- **Lexicon features**: Obtained by using well known lexicons such as the NRC emotion lexicon[203], the NRC hashtag sentiment lexicon [204], the hashtag sentiment

[5] http://computer-ease.com/emotposi.htm
[6] http://computer-ease.com/emotneg.htm
[7] http://computer-ease.com/emotplay.htm
[8] http://computer-ease.com/emotunhe.htm
[9] http://computer-ease.com/emototem.htm
[10] https://en.wikipedia.org/wiki/List_of_emoticons
[11] https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
[12] http://kt.ijs.si/data/Emoji_sentiment_ranking/
[13] http://diego.asu.edu/downloads/verbs.txt

lexicon [204], the opinion lexicon [205, 206], the sentiment140 lexicon [207] and sentiwordnet lexicon [208].

- **Politeness features**: This set of features included a number of lexicons that we curated based on politeness features mentioned in linguistic studies [80]
- **Medical jargon**: This set of features included three different lexicons containing jargon words related to the medical domain. Out of the three lexicons, we prepared a custom lexicon including the list of drugs used in the study, and another custom lexicon containing different phenotypes that could be related to the drug intake (e.g. "Panic", "Hepatitis", "Myopia"). The third lexicon included the terms contained in the Human Phenotype Ontology [209] (e.g. "Abnormality of body height", "Tinnitus", "Photophobia").

As we knew the texts came from different sources of information the parsing and tokenizing strategies were adapted to each text, and when using PubMed texts we used Charniak-Johnson parser [77], while in the case of Twitter we used ARK tagger [78] for both tagging and tokenizing the sentences.

5.2.2 Evaluation

Having all the features in place we ran different experiments on the conflated TwiMed annotations, which contained sentences reporting beneficial effects related to the drug intake ("Benefit" relation) as well as negative outcomes appearing after the drug intake ("Outcome-negative" relation) using the data set presented in section 3.3. The experiments presented in this section were also performed on the set of First-hand experience tweets prepared by the PhD. students as described in section 3.1.

For our experiments we tested different binary classifiers on both PubMed and Twitter (separately, not mixing sentences from both data sets) where each classifier was aimed at detecting one type of sentence. The classifiers are as follows:

- Binary classifier for detecting **first-hand experience reports**: This classifier was tested on 1265 tweets annotated as containing first-hand reports on the drug use and combined with the same number of negative reports (i.e. tweets only containing the drug name but not containing any drug use reports). These results are marked as "First-hand exp." (First-hand experience) in Table 5.6
- Binary classifier for detecting **sentences containing drugs and any related outcome**: This classifier was tested on a set of 1000 sentences for either PubMed

and Twitter containing drug use reports that also included a symptom or disease related to the drug intake. As negative cases we used the same number of sentences that only contained the drug mention but no related symptom nor disease, obtaining 2000 sentences in total from either PubMed and Twitter that we used in the two different classifiers. These results are marked as "Any Outcome" in Table 5.6 and Table 5.7

- Binary classifier for detecting **sentences containing drugs and related beneficial outcomes** : In this case we used a subset of the sentences used in the previous classifier and obtained 729 sentences from PubMed that we matched with the same number of negative cases, and 876 tweets out of which one half were positive cases and the other half were negative cases. These results are marked as "Beneficial Outcome" in Table 5.6 and Table 5.7

- Binary classifier for detecting **sentences containing drugs and related negative outcomes**: In this case we used a subset of the sentences used in the second classifier and obtained 146 sentences from PubMed that we matched with the same number of negative cases, and for the tweets classifier we obtained 341 positive tweets that we matched with the same number of negative cases. These results are marked as "Negative Outcome" in Table 5.6 and Table 5.7

Support vector machines [210] have been extensively used in the area of pharmacovigilance obtaining good results as sentence classifiers [31, 90, 93], and we decided to use that machine learning method and relied on the implementation included in scikit-learn [211] and chose a linear kernel given that most text classification problems are linearly separable [212], and also because that kernel is good when there is a high number of features because mapping the data to a higher dimensional space does not improve the performance very noticeably [213]. To run the experiments we applied 10-fold cross validation in all cases.

To assess the predictive power of the register-related features we first ran an experiment with all the features and then ran the same experiment using feature ablation on different set of features. The difference between the F-score results when using all the features and the F-score results after applying the feature ablation were compared using the Wilcoxon signed-rank test to confirm if the difference in results was statistically significant.

In those experiments we tested different sets of features related to the use of formal and informal registers. As new sets of features that had not been explored in the area of pharmacovigilance we can name Biber's features, other politeness features, and emoji features. The results we obtained when classifying the sentences containing any of the relations presented above in Twitter are shown in Table 5.6, and the corresponding

Features	First-hand exp.	Any Outcome	Beneficial Outcome	Negative Outcome
baseFeatures	0.8233	0.7275	0.6963	0.7373
baseFeaturesAndEmoticon	0.8225	0.7290	0.7009	0.7358
baseFeaturesAndEmoji	0.8233	0.7310	0.6974	0.7344
baseFeaturesAndBiber	0.8391*	0.7255	0.6940	0.7182
baseFeaturesAndOther	0.8434**	0.7330	0.6985	0.7285
baseFeaturesAndBiberAndOther	0.8458**	0.7360*	0.6906	0.7197
baseFeaturesAndPoliteness	0.8415**	0.7310	0.6963	0.7344
baseFeaturesAndJargon	0.8229	0.7495**	0.7226*	0.7578
allFeatures	0.8501**	0.7540**	0.7203*	0.7549

TABLE 5.6: F-score results when using the different binary classifiers in tweets. The single star (**"*"**) indicates significance at 95%. The double star (**"**"**) indicates significance at 99%.

results when using PubMed sentences can be seen in Table 5.7. Please note that the first row shows the results when using our baseline. Following we explain which were the sets of features used in each experiment:

- **baseFeatures**: These are the baseline features that we used to test the significance of the results (using Wilcoxon signed-rank test). These features included "Textual features", "N-gram features", "One character n-gram features:", "Pos features" and "Lexicon features".
- **baseFeaturesAndEmoticon**: "baseFeatures" and "Emoticon features".
- **baseFeaturesAndEmoji**: "baseFeatures" and "Emoji features".
- **baseFeaturesAndBiber**: "baseFeatures" and "Linguistic features".
- **baseFeaturesAndOther**: "baseFeatures", "Other features" and "Other features EXPANDED".
- **baseFeaturesAndBiberAndOther**: "baseFeatures", "Linguistic features", "Other features" and "Other features EXPANDED".
- **baseFeaturesAndPoliteness**: "baseFeatures" and "Politenes features".
- **baseFeaturesAndJargon**: "baseFeatures" and "Medical jargon".
- **allFeatures**: In these experiments we used the baseline and all the other features described before.

The first thing to notice is that the chosen baseline was a strong one as it included POS, unigrams and bi-grams features. Besides this fact we can observe that the results vary depending on the task, the used set of features and the data set.

Features	Any Outcome	Beneficial Outcome	Negative Outcome
baseFeatures	0.8030	0.7908	0.7671
baseFeaturesAndEmoticon	0.8030	0.7894	0.7705
baseFeaturesAndEmoji	0.8030	0.7908	0.7671
baseFeaturesAndBiber	0.7990	0.7935	0.7602
baseFeaturesAndOther	0.8015	0.7949	0.7636
baseFeaturesAndBiberAndOther	0.8025	0.7921	0.7703
baseFeaturesAndPoliteness	0.8020	0.7887	0.7672
baseFeaturesAndJargon	0.8460**	0.8243**	0.7878
allFeatures	0.8405**	0.8278**	0.7877

TABLE 5.7: F-score results when using the different binary classifiers in PubMed sentences. The single star (**"*"**) indicates significance at 95%. The double star (**"**"**) indicates significance at 99%.

In the case of Twitter we can observe that for the binary classification of First-hand experience reports Biber and politeness features obtained a significant improvement when compared against the baseline results in all cases but when using the medical lexicons ("baseFeaturesAndJargon"). The reason for this is that those set of keywords are very specialized and not commonly seen in that tweets data set resulting in a score that is almost the same as the one we obtained when only using the baseline. The same observation is found when using emoji and emoticon features caused by the low frequency of those elements in our set of sentences.

When studying the results obtained in the binary classifier aimed at detecting sentences with "Any outcome" we see that all the results except the one using only Biber features ("baseFeaturesAndBiber", being slightly below the baseline) improve the baseline. The results have different magnitude and in this case we see that when combining Biber features and the extended set of Biber features ("baseFeaturesAndBiberAndOther") the results are significant at 95% level. These results also show that the jargon features provide a noticeable improvement in this data set, which could be related to the number of technical words used in this set of tweets.

The last two columns in Table 5.6, where we used a subset of the sentences used to assess the power of the previous classifier ("Any outcome"), show that for these two classifiers ("Beneficial Outcome" and "Negative Outcome") the newly added set of Biber and politeness features cannot obtain a significant improvement as the results are almost the same as the ones scored by the baseline. For the case of "Negative Outcome" classifier we can also see that Biber features are in fact adding some kind of noise as these features cause a noticeable decrease in the score of the classifier. It is important to mention that these last two classifiers have fewer sentences than the classifiers we assessed in the two previous cases, and these results could be affected by the size of the training data.

The second column in Table 5.7 ("Any Outcome") shows how the emoticon and emoji features do not really contribute to the classifier, which is an expected finding as those

Features	First-hand exp.	Any Outcome	Beneficial Outcome	Negative Outcome
Textual	0.7308	0.5960	0.6085	0.5849
One character n-gram	0.7762**	0.6150	0.6335	0.6318*
Pos	0.7644**	0.5590	0.5319	0.5922
Lexicon	0.6719	0.6460*	0.6176	0.6888**
Linguistic (BiberAndOther)	0.8031**	0.5900	0.5925	0.6070
Unigram	0.7913**	0.6570**	0.6574*	0.6465*
Bigram	0.5695	0.4770	0.4645	0.4574
Trigram	0.4814	0.4770	0.4645	0.4515
Politeness	0.7320	0.5435	0.5523	0.4954
Textual + Linguistic (BiberAndOther)	0.8114**	0.6030	0.6142	0.6187*
One character n-gram + Linguistic (BiberAndOther)	0.8280**	0.6305	0.6450	0.6686**
Pos + Linguistic (BiberAndOther)	0.8118**	0.5985	0.6142	0.6201
Lexicon + Linguistic (BiberAndOther)	0.8079**	0.6805**	0.6724**	**0.7109****
Unigram + Linguistic (BiberAndOther)	**0.8312****	**0.6815****	**0.6746****	0.6729*
Bigram + Linguistic (BiberAndOther)	0.8110**	0.5925	0.5993	0.6026
Trigram + Linguistic (BiberAndOther)	0.8027**	0.5900	0.5925	0.6070
Textual + Politeness	0.7545**	0.5920	0.6210	0.6026*
One character n-gram + Politeness	0.8015**	0.6190	0.6233	0.6450*
Pos + Politeness	0.8090**	0.5635	0.5900	0.6082
Lexicon + Politeness	0.7383	0.6450*	0.6210	0.6991**
Unigram + Politeness	0.8031**	0.6705**	0.6597*	0.6377*
Bigram + Politeness	0.7411	0.5465	0.5671	0.4983
Trigram + Politeness	0.7312	0.5450	0.5523	0.4954

TABLE 5.8: F-score results for individual and combinations of sets of features when using the different binary classifiers in tweets. The single star (**"*"**) indicates significance at 95%. The double star (**"**"**) indicates significance at 99%. The best performing features are marked in bold.

elements are not often seen in in academic texts. This column also shows that the set of linguistic and politeness features do not provide any useful information as these results do not improve the baseline and it is only when we use the medical jargon lexicons when we obtain a significant improvement over the baseline results.

The third column in Table 5.7 ("Beneficial Outcome") shows that Biber features and its variants provide some minor, not-significant, gains to the classifier, while in the last column we observe that only Biber features and Politeness features do contribute even if the gains are not significant.

Same as it happened in the "Negative Outcome" results shown in Table 5.6, the last column in Table 5.7 does not contain any significant improvement regarding the score obtained by the baseline, and most results are almost the same as the ones scored by the baseline.

Having those results ready we were also interested in assessing whether Biber and politeness features could be overlapping with the other sets of features. To check it we created a new version of the classifiers using the set of textual features as the baseline and also the different set of features. That was also helpful to understanding which were the best sets of features.

Features	Any Outcome	Beneficial Outcome	Negative Outcome
Textual	0.5855	0.5891	0.5475
One character n-gram	0.6500**	0.6543**	**0.6912****
Pos	0.6720**	0.6755**	0.5819
Lexicon	0.7060**	0.6824**	0.6229*
Linguistic (BiberAndOther)	0.6465**	0.6584*	0.5410
Unigram	0.7620**	0.7579**	0.6644**
Bigram	0.4940	0.4684	0.4312
Trigram	0.4770	0.4684	0.4312
Politeness	0.5470	0.5466	0.4248
Textual + Linguistic (BiberAndOther)	0.6685**	0.6769**	0.5716
One character n-gram + Linguistic (BiberAndOther)	0.7005**	0.7126**	0.6880**
Pos + Linguistic (BiberAndOther)	0.6685**	0.6872**	0.5955
Lexicon + Linguistic (BiberAndOther)	0.7315**	0.7139**	0.6262*
Unigram + Linguistic (BiberAndOther)	**0.7780****	**0.7777****	0.6441*
Bigram + Linguistic (BiberAndOther)	0.6510*	0.6584*	0.5477
Trigram + Linguistic (BiberAndOther)	0.6475*	0.6570*	0.5410
Textual + Politeness	0.5945	0.5891	0.5267
One character n-gram + Politeness	0.6560**	0.6509**	0.6571**
Pos + Politeness	0.6705**	0.6673**	0.5754
Lexicon + Politeness	0.7080**	0.6872**	0.6194*
Unigram + Politeness	0.7645**	0.7531**	0.6405*
Bigram + Politeness	0.5780	0.5617	0.4213
Trigram + Politeness	0.5495	0.5480	0.4248

TABLE 5.9: F-score results for individual and combinations of sets of features when using the different binary classifiers in PubMed sentences. The single star (**"*"**) indicates significance at 95%. The double star (**"**"**) indicates significance at 99%. The best performing features are marked in bold.

Those results make it clear that there is no clear overlap between Biber features[14] and the rest of the features, and the same statement applies to politeness features. Table 5.8 also shows that the unigrams are the most powerful features, and even if Biber features alone are overall not very powerful, the second column shows that these features perform even better than unigram features. Politeness features alone seem to provide some useful information although we can also observe that they have less predictive power than Biber features. Interestingly, after the first dashed line (showing the results when combining Biber features with other features) we can see that Biber features seem to complement unigram features as that combination produces the best result in three of the four classifiers, and for the remaining best result ("Negative Outcome ") that score is obtained when using in combination Biber features and lexicon features. Similar results can be seen for Politeness features although the gains are less noticeable, and these gains decrease for each one of the data sets, showing that the less sentences we have the less politeness features seem to contribute showing that in fact, for "Negative Outcome " classifier, those features obtain a F-score result below 50% which is remarkably low when considering the data sets contain the same number of positive and negative sentences.

[14]These features include the original set of Biber features and our custom expansion to these features. These are the features presented in Table 5.6 and Table 5.7 in combination with the baseline features under the name "baseFeaturesAndBiberAndOther".

Features	First-hand exp.	Any Outcome	Beneficial Outcome	Negative Outcome
Biber (original)	Total Adverbs, Demonstratives, First Person Pronouns, Gerunds, Infinitives, Nominalizations, Past Verbs, Present Verbs, Second Person Pronouns, Third Person Personal Pronouns (excluding "It"), Time Adv, Total Other Nouns, Type/Token Ratio, Word Length	Attributive Adjectives, Suasive Verbs, Type/Token Ratio	Private Verbs, subject pronouns, Word Length, Total Other Nouns	Attributive Adjectives, Type/Token Ratio, Third Person Personal Pronouns (excluding "It"), First Person Pronouns, Prepositional Phrases, Nominalizations
Biber (Expanded)	Contrast expressions, demonstratives, "Have" verbs, integer numerals, object pronouns, possessive pronouns, prepositions, pronouns, subject pronouns, ASU common words	-	"Have" verbs	-
Politeness features	Contraction words, coordinating conjunctions, demonstrative words proximal words, discourse markers of ordering, discourse markers of speech, English conjunctions, informal words, intensifiers obtained from the Thesaurus, long names, prefixes English (native), prefixes English (neo classical), slang words, nouns' suffixes	Long names, nicknames	Prefixes English (neo classical)	Long names, nicknames, Last token is question mark

TABLE 5.10: Best Biber and Politeness features when using the different binary classifiers in tweets.

The results shown in Table 5.9 are similar to the ones presented in Table 5.8 as here too we can see that unigrams contribute very positively, while Bigrams and Trigrams do not. Here too, we can see that Biber features contribute to all the other features, and only when using POS features the gains seem to be little, but still positive. When studying the gains contributed by Politeness features we can see that for most cases the combination of Politeness features with the other sets of features create some conflicts. The results do not improve and even get decreased when adding the set of Politeness features. That would mean that Politeness features do not provide any gain to binary classifiers using PubMed texts, although Biber features do contribute positively.

To understand which were the most helpful features we extracted the top 10% of the features by using F-selector[15], which is a univariate feature selection method that examines each feature individually to determine the strength of the relationship of the feature with the label by using the ANOVA F-value for the provided sample. Doing so helped us in seeing which were the set of Biber features (the original features, and also our expanded set of features) and Politeness features that were selected for each classifier, and compare the pervasiveness of those features in different classifiers by identifying the features appearing in the ten cross-fold validation experiments that we ran in each case.

Results in Table 5.10 show that there is no single feature which by itself contributes to all the classifiers, and even if features such as the "Type/Token Ratio" appear in three out of the four classifiers the rest of Biber features only appear in at most in two classifiers. It is noticeable that out of all the features used in Biber's original study the best performing features are those related to the use of personal pronouns (first and third

[15] http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_classif.html#sklearn.feature_selection.f_classif

Features	Any Outcome	Beneficial Outcome	Negative Outcome
Biber (original)	Attributive Adjectives, Total Adverbs, Past Verbs, Type/Token Ratio, Prepositional Phrases, Word Length, Agentless Passives	Prepositional Phrases, Word Length, Agentless Passives, Attributive Adjectives, Total Adverbs, Past Verbs, Infinitives	Gerunds, Word Length
Biber (Expanded)	Verb "BE", integer numerals, prepositions	Verb "BE", integer numerals, prepositions	-
Politeness features	prefixes English (neo classical), nouns' suffixes	nouns' suffixes	Slang words

TABLE 5.11: Best Biber and Politeness features when using the different binary classifiers in PubMed sentences.

personal pronouns), nouns ("Nominalizations" and "Total Other Nouns"), attributive adjectives, and with the length of the words, showing that other elements such as the different verbs and other constructions do not have an impact.

The expansion of Biber features shows that only the use of the verb "have" contribute to more than one classifier, and it seems that only the set of first-hand experience tweets get any benefit from those features. Lastly, the set of Politeness features only show that the use of long names (personal names such as "Daniel" or "Harrison") provide some gains to three classifiers, and the politeness features to detect the use of (neo classical) English prefixes and nicknames also provided gains to the classifiers.

In the case of PubMed (Table 5.11), the number of Biber and Politeness features contributing to the system are fewer. We can see that the only feature appearing in the three experiments is the one encoding the length of the words, and also that in two classifiers the used set of features include "Attributive Adjectives", the count of adverbs ("Total Adverbs"), verbs in past form ("Past Verbs"), "Prepositional Phrases", and "Agentless Passives".

Our custom expansion on the set of Biber features show that encoding the number of times that the authors use the verb "to be" (Verb "BE"), integer numerals (1, 2, 3...) and prepositions (including all English prepositions, and not only the ones used in prepositional phrases) are useful in these kinds of classifiers. Lastly, the set of politeness features show that the nouns' suffixes (-hood, -ess, -ness) can contribute to the system. The unexpected finding was the observation that slang words seem to contribute to a classifier using PubMed sentences, which we believe is due to the large number of slang words we included in such lexicon (2night, 4ever, dunno), which in fact cover some technical names ("h2o" for water and "OMG" for "3-0-methylglucose") and acronyms ("adhd" for and "hru" for "High Resolution Ultrasonography").

The observation on the best performing Biber features tells us that the length of the words and the type/token ratio are pervasive features in PubMed and Twitter. Similarly,

the count for the number of adverbs, different forms of the verbs (gerund, infinitive, and past), the count of attributive adjectives and prepositional phrases seem to provide gains in both classifiers using tweets and PubMed texts.

Our expansion on Biber's features show that the count for integer numerals and prepositions is useful when using different data sources, and even if the count for different forms of the verbs "have" and "be" only appear in the classifier using tweets or PubMed sentences, respectively, it is an interesting observation as those auxiliary verbs contribute positively too. Finally, from the set of Politeness features we observe that the use of neo-classical English prefixes (e.g. afro-, anglo-, euro-, franco-), nouns' suffixes (-hood, -ess, -ness), and slang words have the ability to capture some useful information that help the classifiers.

5.2.3 Error Analysis

We are now going to analyse the sentences that were correctly labelled when running two out of the four different classifiers. The first classifier for which we are going to analyse the results is the classifier telling apart tweets reporting first-hand experiences from other type of tweets (i.e. "First-hand experience" reports classifier), and then move on to present the results from the classification of sentences containing outcomes related to the drug intake (i.e. "Any Outcome" reports classifier)

The results show the differences between the labels produced when only using the baseline features and we compare them against the results produced when using the set of Biber features combined with our set of "Other Biber features" (in this case we also use the baseline features). Similarly, we compare the improvements obtained by the system using the baseline features in combination with Politeness features against the system that only uses the baseline features. For these experiments we ran a three fold cross validation and examined the results obtained in the first fold.

We begin this error analysis by studying which were the First-hand experience reports that we were able to label correctly when using the set of Biber features and the set of "Other Biber features", being both reports correctly recognized as not first-hand experience reports, and also a the reports that were correctly recognized as positive reports.

To continue our analysis we now use the set of best performing features presented before (see Table 5.10) and inspect whether these features occur in the sentences that the classifier using the set of Biber and the extended set of Biber features got right while the classifier only using the baseline features failed to label correctly. We found that

almost all sentences had some of these elements. The sentences only having values for the features "Type/Token Ratio" and "Word Length" were *"Where over against draw off doormat lustral machines that does its performance advanced spades: XsySu"* and *"Thnkfl 4 @adderall"*. In both cases the value of the average word length was found to be "5.0", that we also found in 7 sentences from the same set of correctly categorized sentences. Besides these two special cases we noticed that the most frequent features for those sentences were the count for "Present verbs", the count for "First Person Pronouns" and the count for "Total Adverbs" (characterizing in 23, 20 and 17 sentences, respectively, of the 38 that were correctly labelled). Interestingly, there are only two sentences (*"I wonder if the people who truly like to study and are really successful in school naturally have adderall-like brain chemical processes"* and *"A: Hi all, I have cptsd currently on sertraline and looking at a second med to replace codeine. I just wanna o... http://t.co/m8kGQRUGi7"*) with counts for those three features, although in most of these cases we find that only two out of those three features appear together, which can imply that the different combination of these features provide very good information for the classifier. The count for "Past Verbs", and the count for "Third Person Personal Pronouns (excluding It)", do also appear quite often in the correctly labelled sentences for which we believe they may be providing very useful information to the classifier, also reinforced by the findings shown in Table 5.10.

Continuing with a focus on the set of our extended set of Biber features we find that the use of prepositions, pronouns, and words in the ASU lexicon (these are words that are frequently seen in drug-use tweets: before, cause, effective...) appear in most of those correctly classified sentences, which are among the best-performing features in Table 5.10. Looking again at the sentence *"Thnkfl 4 @adderall"* we can see that our extension on the set of Biber features accounted for the presence of the numeric ("4"), while in the case of the other sentence where the set of best performing Biber features did not find any of those best-performing features (*"Where over against draw off doormat lustral machines that does its performance advanced spades: XsySu"*), we can see that our extension to the set of Biber features found one demonstrative, a preposition and a pronoun, which in combination with the Biber features may be helping the classifier in correctly labelling this sentence as one that is not reporting a first-hand experience.

We now perform a similar study on the set of Politeness features, to investigate which are the features appearing in the 23 sentences that our classifier (enhanced with the set of Politeness features) gets right while the baseline classifier fails to correctly categorize. Here we find that the use of English conjunctions and prefixes (both native and neo-classical) are the features most frequently found as well as the use of intensifiers (using the thesaurus lexicon) and contractions. The sentence *"These kids sporadically breaking down in the library need Adderall in their lives"*, only contains an "informal word" is

found ("need") while the sentence *"<USER> <USER> ADD was it. Wizards gave him herbs to slow it down. Adderall was taken"*, contains "slang words", but no other feature from the set of best-performing features shown in Table 5.10 appear in these two cases, which may be useful to understand that these two features alone provide gains to the base system. We also observed that one sentence for which none of the best performing Politeness features appear, *"Blame it on the a-a-a-a-adderall"*, is correctly categorized as not being about the drug, and even if no best-performing feature is found we see that it is found to contain affixes and (a-) and third person pronouns (it), that can lead to its correct labelling. Those findings, and in particular the one for the last sentence where no best-performing feature appears, tells us that different combination of the Politeness features contribute towards the correct classification, and even if there are some features that show to have a stronger contribution not all sentences correctly classified contained the best performing Politeness features.

The next classifier for which we are going to analyse the results is the classifier telling apart tweets reporting sentences containing any type of outcome, either positive or negative, from sentences not reporting outcomes related to the drug use. This classifier is the one that we previously referred to as "Any Outcome" reports classifier. Here too we are going to start by comparing the sentences that the system using the set of baseline features alone is not able to correctly label while the system using the baseline features in combination with the set of Biber features and the set of "Other Biber features" classifies correctly. In this case too we ran a three fold cross validation and examined the results obtained in the first fold.

Moving on to the analysis of the sentences that the classifier using the set of Biber features recognizes correctly we find that in the fold we evaluated there are ten sentences that this classifier gets right while the classifier using the set of baseline features does not label correctly. Looking at Table 5.10 we can see which are the best performing features ("Attributive Adjectives", "Suasive Verbs" and "Type/Token Ratio"), and find that those sentences contain attributive adjectives, and as a false positive, but useful in the classification the sentences *"just moved a 200+ lb washer/dryer combo unit out of my house by myself at 5am... #coke #adderall #xanax #hash #whiskey"* and *"Tallahassee is moved to the 7th & 8th! The entire family is on alert to beware of the side effects of the prednisone! #dontpissmeoff"* are recognized as containing a suasive verbs ("move"), although in this case the verb does not have a suasive function.

The set of Politeness features appearing in Table 5.10 for the sentences classified using the system to detect the ones containing any type of outcome are only two: "Long names", and "nicknames". We can see that even if these are the best performing features the correctly classified sentences do not contain "long names", although they contain a

keywords recognized as a nicknames *"So a few weeks ago, I had to call 911 for a friend who had a seizure on my couch after just one 150mg dose of Seroquel."*, *"just moved a 200+ lb washer/dryer combo unit out of my house by myself at 5am... #coke #adderall #xanax #hash #whiskey"*, and *"Had a debacle with da boi's #MediCAl today- it wasn't turned on soon enough to pick up his lamotrigine- last dose today ($174 per mo)"*. In this case "just" and "mo" are in fact false positives, but their detection may be the cause for the correct labelling. It is interesting to notice that the second of those three sentences was also correctly classified when using the set of Biber features.

When focusing on the set of PubMed sentences using the set of Biber features our system is able to correct two that are wrongly classified by the baseline system. The features shown in the second column of Table 5.11 that are found in these sentences are the use of "Attributive Adjectives", and "Prepositional Phrases" in the first case (*"Lamotrigine is a newer, unrelated antiepileptic drug that causes skin rashes in 3-10% of new users."*), and the use of the same two features in combination with the feature "Total Adverbs" in the second sentence (*"Bevacizumab has good tolerability with manageable side effects, both alone and in combination with other agents; the tolerability profile of bevacizumab in combination with IFN is consistent with the well-characterized and well-established profiles of these therapies."*). Coincidentally, our extension for the set of Biber features in those sentences found the use of the same two features in both cases, the use of the verb "to BE", and the use of "prepositions". In this case the first sentence is annotated as being about the drug and also including an outcome related to the drug intake, while in the second case the system is able to classify it as a sentence that does not contain any outcome related to the drug intake.

When looking at the PubMed sentences that the baseline system fails to label correctly while the system using the set of baseline and Politeness features annotates right we only find one sentence. That sentence, *"Behavioral and metabolic effects of the atypical antipsychotic ziprasidone on the nematode Caenorhabditis elegans."*, containing neo-classical prefixes (appearing in Table 5.11) such as "meta" (in metabolic) and "anti" (in antipsychotic), and that is the probable reason why the system can classify the sentence correctly.

When labelling first-hand experience tweets we observed that in that fold there were 54 labels where both the baseline and the system using Biber features disagreed. We found that out of those sentences only 16 failed when using Biber features but were annotated correctly when using the baseline features only, and out of those 16 sentences the system using Biber features annotated 5 of them as positive sentences (i.e. sentences reporting first hand experiences). A manual inspection showed that two of those sentences include retweets ("RT") where one of them would have been a first-hand experience in case the

tweet would have not been a retweet (*"RT <USER>: Percocets, Adderall, ecstasy, pussy, money, weed, faded for a week, I don't sleep. Fuck my enemies"*). The other sentences include the mention of a movie that the user is watching (*"watching prozac nation and screaming"*), another RT indicating the potential use of the drug *"RT <USER>: I could use some adderall in my green tea"*, and two comments related to the use of the drug (*"Kristine on adderall is so scary"* and *"Music is the adderall to my homework"*).

Looking at the sentences where the classifier using Biber features fails to recognize the sentences as reporting first-hand experiences but the classifier using the baseline features gets it right we find that the use of metaphors also plays an important role (e.g. *"You don't understand. I'm a whole different person on Adderall."*, *"Adderall is dumb all you do is feel like God for a few hours then u hit rock bottom and break down in the shower"*, *"Adderall had 5 short answer and one essay question turn into 6 essays and writing for an hour and a half while everyone left after a half hr"*). There are also some sentences where the person who took the drug is not stated, and although humans could infer it that may cause trouble to the classifier (e.g. *"c'mon Adderall, do your stuff."*, *"Red bull and adderall do wonders until you try to sleep haha"*, *"Don't call your mom 30 min after taking an adderall. You will spend an hour and a half talking about who really even fucking knows"*, *"DATE DATE DATE DATE (literally popped some prozac while on the phone w him"* and *"It's like Adderall makes homework the toughest level of your fav video game. 7 hours go by and you're still like "nope gotta get this done""*), in those sentences we can see that the pronouns "you", "your" and "him" appear, which could be misleading the classifier. The last sentences where the classifier using Biber features failed are sentences to recognize first-hand experience reports are sentences including the Twitter user names *"<USER> <USER> <USER> <USER> They r over prescribed and I'm coming off Celexa right now from 60mg to 0"*, *"<USER> I'd suggest Zoloft, but it interferes with having orgasms!"* and *"<USER> girlfriends are cool but they dont do your paper in 30 minutes so I rather have the adderall"*. These sentences are clearly reports on the drug use, also including first person personal pronoun ("I"), which is a strong indicator of first-hand experience reports, although two of those sentences also include the pronoun "they", which may have a higher weight in the classifier.

Studying the differences between the classifier using the set of Politeness features and the classifier only using the baseline features shows that there are 36 sentences (in one of the three folds we ran) where the baseline system and our classifier differ in the annotation, although only 13 of those sentences were wrongly annotated by the classifier using the set of Politeness features. Out of those fails, 5 sentences were annotated as positive (although they were in fact not reporting first-hand experiences), and the other 8 sentences were annotated as negative cases. There are three sentences that were also presented in the previous error analysis (when using Biber features) that were also

annotated as positive (*"Kristine on adderall is so scary"*, *"Music is the adderall to my homework"* and *"RT <USER>: Percocets, Adderall, ecstasy, pussy, money, weed, faded for a week, I don't sleep. Fuck my enemies"*). The other two sentences are new and are sentences including the personal pronoun "I" that can be the cause of the problem (*"Adderall. Translation: Molly. (Or so I've heard.)"*, and *"<USER> hahahahahah I think about this stupid tweet all the time now and it's better than Prozac"*.

In the case of the sentences annotated as positive when they are not reporting first-hand experiences we find that there are four sentences also missed by the previous classifier (*"It's like Adderall makes homework the toughest level of your fav video game. 7 hours go by and you're still like "nope gotta get this done""*, *"c'mon Adderall, do your stuff."*, *"DATE DATE DATE DATE (literally popped some prozac while on the phone w him"*, *"<USER> girlfriends are cool but they dont do your paper in 30 minutes so I rather have the adderall"*), and four new sentences that did not appear in the previous analysis. Out of those four sentences, one is a rhetoric question that may be have been wrongly annotated by the human annotators while in fact can be seen as reporting a first-hand experience (*"Is the true purpose of adderall to help me focus on studying or to make me snap and tweet nonstop????"*, *"It probably wasn't a wise idea to combine Modafinil, Tramadol, and Jack Daniels tonight. My head feels like it's about to explode."*). As for the last two sentences, one is clearly not reporting a first-hand experience, but the use of the personal pronoun "I" is frequent and that may cause the problem (*"Big thanks to whoever the hell invented Lexapro, I suppose! I owe that person a drink."*), while in the last case the sentence *"It's one of those get-to-work-take-an-Adderall-immediately kind of days."* seems to be a though more than a report in which the writer took the drug.

The next classifier for which we are going to analyse the errors is the classifier telling apart tweets reporting sentences containing any type of outcome, either positive or negative, from sentences not reporting outcomes related to the drug use. Thais classifier is the one that we previously referred to as "Any Outcome" reports classifier. Here too we are going to start by comparing the errors between the system using the set of baseline features alone versus the system using the baseline features in combination with the set of Biber features and the set of "Other Biber features". In this case too we take into account the results obtained in the first of the three folds we ran.

We begin by analysing the results on the set of tweets, and then proceed to focus on the results obtained when using the classifier on the set of PubMed sentences. In our tweets data set we found 15 sentences where the baseline and the classifier using Biber features disagreed on the annotation, although there were only 5 sentences where the classifier using the set of Biber features and the set of "Other Biber features" missed the correct label.

In three of these sentences the classifier was not able to recognize the report of some type of outcome: *"My body be feeling hot af when I take vyvanse."*, *"Why IS There So Much #Fibromyalgia in Recent Years? How much of what is diagnosed as Fibro is really caused by #Ciprofloxacin? #nutrition"*, *"I'm expecting some insomnia and nausea but don't know what to expect after that #olanzapine"*. We can see that the first sentence is clearly reporting an outcome, although the sentence uses a very uncommon structure "body be feeling hot" which may be causing some problems to the classifier. In the second sentence the outcome is clearly identifiable by the structure *"Fibro is really caused by #Ciprofloxacin"*, but given we did not use those kind of patterns it could be the classifier misses to figure out the relation between "Fibro" and the drug ("#Ciprofloxacin"), in this case it is also very clear that "Fibro" is an outcome by noticing the pattern "diagnosed as Fibro". The third example may be a miss because of the use of the negation "don't", that can make the classifier think that there is no relation between the elements in the sentence, thus overlooking the relation between the drug and the outcome. The counterpart to those sentences wrongly labelled as negative by the classifier using the set of Biber features and the set of "Other Biber features" are the two sentences that the classifier misclassifies as positive while the baseline system correctly recognizes as negatives. There are two of these sentences: *"I wonder, would an aspirin (costing just a few cents) do the same as montelukast which costs 7 dollars per pill?"*, and *"Jarrett was so pussy though...like how u killing urself n she has a pimp...he need Zoloft"*. In the first example there is no doubt that no outcome is present and the classifier should have annotated the sentence correctly, while in the second case the ungrammatical structure and the use of contrations ("u", "n") may be the cause of an incorrect label for the sentence.

Continuing with the inspection of the errors in the classification of "Any Outcome" reports we now focus on the disagreements between the baseline system and the classifier using the baseline features combined with the set of Politeness features. In the sentences included on the first fold we evaluated we find 12 sentences where both classifiers produce different annotations, and only 4 of them are wrongly classified by the classifier using Politeness features. Here we find that there is only one sentence that was annotated as not containing outcomes but the annotators said that it did in fact contain an outcome (*"I haven't really experienced any significant paxil withdrawal after 3 days now. I guess effexor is similar enough to it??"*). That sentence can be thought as a conflictive one as it does in fact contain a mention to an outcome, "withdrawal", although it is negated and the outcome is in fact not related to the drug. The three remaining sentences were sentences for which the classifier using the set of Politeness features said that the sentences did include outcomes related to the drug intake while in fact the annotators and the baseline system said that these sentences did not contain any outcome. These

sentences are: *"Feel like crud. Steroid and antibiotic shot yesterday, and on Levaquin for 10 days. Hope and pray it alleviates the issues."*, where we can see that "the issues" could be thought of being an outcome, and it may be the cause for the wrong classification. The following sentence, *"Wish I was still doing vyvanse so I could actually do school work"*, does not contain any symptom, and is in fact a clear miss. The same thing happens in the third example, *"This is the almost riskless general hospital cadence lustral foal?: NqkXNV"*, as it is also a clear miss although in this case it could be that the keyword "riskless" or "general" (e.g. "general infection", "riskless disorder") may be learnt by the classifier as related to sentences including outcomes causing the misclassification.

Moving now to the differences in the annotations when using the classifier in the set of PubMed sentences we find that there are only two sentences that the classifier using the set of Biber features and and the set of "Other Biber features" fail to correctly classify. The first sentence, *"No changes in blood pressure were found, but pulse decreased 8.3 +/- 2.4 for haloperidol with lorazepam and 8.9 +/- 4.24 for ziprasidone (P = NS)."*, is annotated as not containing outcomes when it does in fact contain an outcome ("pulse decreased"). The presence of the negation ("no changes"), taken into account in Biber's features, can be misguiding the classifier, also the outcome ("pulse decreased") only appears in that sentence in the whole data set. The opposite happens in the sentence *"Trazodone is an antidepressant which behaves as a selective 5-HT(2) antagonist and 5-HT reuptake inhibitor."* as it is annotated as containing some outcome when in fact it does not contain any. In this case it could be that the term "antidepressant" or "inhibitor" may be learnt by the classifier as appearing in sentences including outcomes, thus recognizing this sentence as one including outcomes.

Out of the two remaining sentences from PubMed that the classifier using the set of Politeness features classifies incorrectly while the baseline classifiers labels correctly there is one that does contain an outcome, and is incorrectly classified as not containing any outcome. This sentence, *"It is thus possible that treatments that alter gut microbiota composition could ameliorate olanzapine-induced weight gain and associated metabolic syndrome."*, contains two outcomes ("weight gain" and "metabolic syndrome"), and should have been labelled as positive. The last sentence, *"Trazodone is an antidepressant which behaves as a selective 5-HT(2) antagonist and 5-HT reuptake inhibitor."*, is the same sentence that the system using Biber features failed to classify correctly, showing that the baseline system may be capturing some information that is overridden by these classifiers, thus missing the correct classification for such sentence.

5.3 Named Entity Recognition systems using register information

We previously introduced the different classifiers and justified their importance in the area of pharmacovigilance. Having a set of classifiers is of interest to perform a first selection at sentence level, providing a useful filter that allows researchers to focus on the sentences of interest. The next step, once the sentences of interest have been identified, is to point to the entities that are being mentioned in those sentences, and for this we work in the named entity recognition (NER) taks. Recognizing the entities in the reports is of key importance to rapidly identify co-intake of medicines and symptoms that occur in combination with others. NER is also a very important task as this is the previous step before identifying the relations between drugs and outcomes.

5.3.1 Methods

As the data set presented in section 3.3 contained the annotation at entity level we created a Named Entity Recognition (NER) system to assess the power of the newly added Biber and Politeness features. One important note here is that Biber features (as described originally) were in most cases computed by taking into account a window of words (and their POS) in the sentece. For this we prepared two different approaches. In the first one we used Biber features without performing any adaptation, so each token in the sentence would have the same results for the whole sentence, i.e. all the tokens in the sentence would contain the same values for those features. We identify these features as "Biber (Original)" in following tables. We understand that this is a very naive approach as the individual tokens cannot be expected to get any improvement because all the tokens in the sentence would contain the same (noisy) values for those features. Our second approach expanded the set of Biber features and captured the different elements used to compute those features, we identify these features as "Biber (Adapted)". In this second case we obtained 98 different features that would be checked to confirm the categories in which the token could be included. As most of the features were not complimentary it is important to notice that in most cases just one of those features would be activated.

For our set of Politeness features we used the same set of 43 politeness lexicons that we used in the classifiers and activated those features for the lexicons in which the tokens would be included.

For these experiments we used a strong baseline composed of the token, the lemma, the part of speech (POS). For each of those features we used a window of three words

(three words before the current word and another three words after the current word), obtaining nine different features in total as we also used the combination of the current word with the previous and following word. The baseline also encoded the shape of the words [214] although in this case we only used 3 features to encode the shape of the previous, current and following words

In our experiments we also used three different sets of clusters of words built using word2vec [136]. Two of them were custom-built using our set of PubMed sentences −"Word2vec (PubMed))"− and tweets −"Word2vec (Tweets)"−, while the last one was obtained from another research group −"Word2vec (Arizona)"− [6]. The experiments using all these features are presented in Table 5.12 as "All Word2vec features".

Our NER system was built using CRFsuite [215] as this is a state of the art implementation of Conditional Random Fields (CRFs) that is being used to build pharmacovigilance systems [6, 137, 216]

5.3.2 Evaluation

To test the gains provided by Biber and Politeness features we assessed the gains when combining a different number of features, and also when using some features alone (identified by the keyword "Only" in the following tables). Table 5.12 and Table 5.13 show the results when using different combinations of those sets of features. For each one of these features we obtained the information of the previous, current and following words. The first column in those tables shows the used set of features. The second column shows the overall result, and the last two columns show the F-score results for identifying "Diseases/Symptoms" and "Drugs", respectively.

In the case of Twitter, we had 24,722 tokens in total and 1777 sentences, while in the case of PubMed we used 1005 sentences which contained 31539 tokens.

The first clear observation from both tables (Table 5.12 and Table 5.13) is that the application of Biber features as originally proposed does not contribute at all showing that, as expected, these features do not have any predictive power by themselves. Low results appear too when using the set of politeness features showing that those features do not contribute to this task. Interestingly, when we use our adaptation for Biber features alone we obtain a better performance than in the previous two cases although it is still below most of the results scored by the other sets of features.

When looking at the results obtained when using the set of tweets we can see that Biber (Original) features are in fact very weak, and in fact these seem to be in conflict with other sets of features because their combination decreases the scores obtained by

Features	Overall	Disease − Symptom	Drug
Baseline	69.36	48.63	89.46
Only Token	58.65**	38.40**	78.35**
Only POS	62.64**	42.51**	82.51**
Baseline + Word2vec (PubMed)	69.75	49.60	89.47
Baseline + Word2vec (Tweets and Arizona. No PubMed)	73.35	54.54	**92.56**
Baseline + Word2vec (Tweets)	70.97	51.13	90.51
Baseline + Word2vec (PubMed and Arizona. No Tweets)	72.92	54.70	91.14
Baseline + Word2vec (Arizona)	72.62	54.26	90.94
Baseline + Word2vec (PubMed and Tweets. No Arizona)	70.88	51.17	90.40
Baseline + All Word2vec features (PubMed, Tweets and Arizona)	73.32	54.63	92.31
Only Biber (Original)	00.63**	00.56**	00.72**
Baseline + Biber	64.17**	42.69**	85.87**
Baseline + All Word2vec features + Biber	69.86*	50.02*	90.29*
Only Biber (Adapted)	25.81**	09.58**	39.69**
Baseline + Biber (adapted)	69.38**	48.86**	89.25**
Baseline + All Word2vec features + Biber (adapted)	73.07	54.34	92.26
Only Politeness	02.46**	02.54**	02.37**
Baseline + Politeness	69.06**	48.21**	89.44**
Baseline + All Word2vec features + Politeness	**73.80**	**55.79**	92.41
Baseline + Biber (adapted) + politeness	69.27**	49.07**	88.97**
All Features	73.44	55.01	92.48

TABLE 5.12: F-score results for different sets of features on a NER system using Twitter messages. The single star (**"*"**) indicates significance at 95%. The double star (**"**"**) indicates significance at 99%. The best performing features are marked in bold.

Features	Overall	Disease − Symptom	Drug
Baseline	78.60	63.20	93.65
Only Token	74.46**	57.69**	90.31**
Only POS	75.74**	59.58**	91.37**
Baseline + Word2vec (PubMed)	79.46	65.44	93.60
Baseline + Word2vec (Tweets and Arizona. No PubMed)	79.41	65.79	93.37
Baseline + Word2vec (Tweets)	78.88	64.08	93.60
Baseline + Word2vec (PubMed and Arizona. No Tweets)	79.57	66.42	93.23
Baseline + Word2vec (Arizona)	79.52	65.37	**93.92**
Baseline + Word2vec (PubMed and Tweets. No Arizona)	79.20	65.35	93.34
Baseline + All Word2vec features (PubMed, Tweets and Arizona)	79.66	66.72	93.07
Only Biber (Original)	00.82**	00.00**	01.77**
Baseline + Biber (Original)	75.02**	58.07**	91.33**
Baseline + All word2Vec features + Biber	77.24**	63.43**	91.45**
Only Biber (Adapted)	32.45**	19.65**	43.98**
Baseline + Biber (adapted)	78.49	63.35	93.26
Baseline + All Word2vec features + Biber (adapted)	79.67	66.67	93.17
Only Politeness	02.12**	02.81**	01.31**
Baseline + Politeness	78.67	63.34	93.65
Baseline + All Word2vec features + Politeness	**80.02**	67.34	93.20
Baseline + Biber (adapted) + Politeness	78.35*	63.31*	93.10*
All Features	79.98	**67.39**	93.08

TABLE 5.13: F-score results for different sets of features on a NER system using PubMed texts. The single star (**"*"**) indicates significance at 95%. The double star (**"**"**) indicates significance at 99%. The best performing features are marked in bold.

other features alone. Our custom set of Biber features, "Biber (Adapted)", seems to behave much better as these features do not cause any conflict and provide some minor contributions in the detection of diseases and symptoms when combined with the baseline features. The last set of features aimed at capturing the use of the differences in the linguistic register, i.e. Politeness features, show an interesting behaviour as these cause some −non-significative− loss when combined with the baseline features, but when added to the baseline and the sets of word2vec features that configuration obtains the best overall score.

Focusing on the results obtained when using PubMed sentences, Table 5.13, we see a minor increase on the performance in the detection of drugs and also in the detection of symptoms and diseases when we combine our adapted set of Biber features together with the baseline features and the word2vec clusters. Politeness features, even if obtaining low scores when used alone, seem to contribute when combined with the baseline and slightly improve performance in detecting diseases and symptoms. The same observation can be made when adding the set of word2vec features as that configuration obtains the best overall score confirming that politeness features, which did not seem to provide a good performance when used alone, can in fact add some information that the other sets of features miss.

These findings show that the detection of symptoms and diseases is the task that benefits the most from the newly added sets of features, and in particular, the combination of Politeness features with existing sets of features used in current pharmacovigilance systems have potential for improving current systems.

As our sets of custom features, i.e. "Biber (Adapted)" and "Politeness", include a wide range of features (98 and 45 features, respectively), we ran a set of ablation experiments to understand which were the features that when left out produce a noticeable loss in performance.

Here we can see that even if a number of features produce a loss in the "Overal", "Disease − Symptom" and "Drug" F-scores, the feature causing the biggest impact is the "Count of word length". We can also see that the use of the "Infinitives", and some hedging words (about, like, or, sort, kind) appear in Table 5.14 and Table 5.15 showing that these features are useful in both the detection of drugs, symptoms and diseases in both PubMed sentences and tweets. These two tables, and in particular Table 5.14, show that some of the features such as "Infinitives", "Time adverbials", "Word length" and "Nominalizations" that showed to contribute positively when creating the Binary classifiers also contribute to NER systems.

Features	Overall	Disease − Symptom	Drug
Only Biber (Adapted)	25.81	09.58	39.69
#1 Necessity verbs (need, have)	25.48	09.06	39.43
#9 Profanity words (damn, fuck...)	25.39	09.32	39.19
#11 English stop words (me, until, while, very..)	25.53	09.05	39.53
#19 conjunctions (else, however)	25.59	09.30	39.53
#23 Numerals (one, two)	25.40	09.16	39.29
#29 Time adverbials (again, earlier)	25.56	09.46	39.40
#34 Demonstrative pronouns (this, that)	25.33	09.30	39.03
#39 Nominalizations (-tion, -tions, -ment)	24.33	09.05	37.61
#40 Occurrences of "By"	25.54	09.43	39.33
#41 Verb "Be" (am, is)	25.56	09.31	39.45
#42 Possessive pronouns (my, our)	25.44	09.41	39.19
#49 Occurrences of "To" (infinitives)	25.39	09.30	39.21
#64 Count of word length	10.83	04.43	17.47
#72 Downtoners (almost, barely)	25.56	09.15	39.67
#73 Hedges (at, more, something, almost, maybe)	25.07	09.06	38.78
#74 Hedges (about, like, or, sort, kind)	24.68	09.31	38.02
#76 Amplifiers (absolutely, completely)	25.49	09.16	39.48
#78 Emphatics (such, so, real)	24.34	06.94	39.00
#79 Emphatic (sure, lot)	25.11	08.66	39.12
#80 Discourse particles (well, now, anyway)	25.51	09.32	39.33
#97 Analytic negations (not, n't)	25.42	09.18	39.33

TABLE 5.14: F-score results for the ablation experiments on Biber (adapted) features on a NER system using Twitter messages.

Features	Overall	Disease − Symptom	Drug
Only Biber (Adapted)	32.45	19.65	43.98
#36 Verb "Do" (not used as auxiliary)	32.30	19.57	43.78
#47 Public, Privative or Suassive verbs	32.37	19.58	43.85
#49 Occurrences of "To" (infinitives)	32.04	19.35	43.49
#62 Occurrences of "As"	32.20	19.63	43.60
#64 Count of word length	14.30	05.81	22.54
#74 Hedges (about, like, or, sort, kind)	32.19	19.56	43.59

TABLE 5.15: F-score results for the ablation experiments on Biber (adapted) features on a NER system using PubMed texts.

Once we had those results for our set of adapted Biber features, we performed the same ablation tests for the Politeness features.

The higher loss in performance in Table 5.16 appears when removing the features capturing the use of coordinating conjunctions (and, and/or), and the use of movement verbs distant words (take, go). The following big drop in performance appears when removing the feature accounting for the use of informal words (seem, climb, help). Capturing the use of cliche words (ace, almighty, bull), non-exaggeration words (humdrum, mediocre) and words used as weak starters (it, this, and, but) also seems to contribute in the detection of diseases and symptoms. That table also shows that for identifying drugs the lexicon including abbreviations (tl/dr, a/c) −Probably due to the fact that these words contain some drug keywords− and demonstrative words proximal words (here, this) contribute with some gains.

Features	Overall	Disease − Symptom	Drug
Only Politeness	2.46	2.54	2.37
#2 Formal words (appear, ascend, assist)	2.21	2.24	2.18
#3 Informal words (seem, climb, help)	2.05	1.80	2.35
#6 coordinating conjunctions (and, and/or)	1.65	1.95	1.28
#8 English conjunctions (afore, after, against)	2.14	2.09	2.19
#9 Cliche words (ace, almighty, bull)	2.13	1.95	2.36
#10 Exaggeration words (noteworthy, wonderful)	2.30	2.24	2.36
#11 Non exaggeration words (humdrum, mediocre)	2.13	1.95	2.36
#12 Intensifiers (just, fully)	2.38	2.39	2.36
#13 Thesaurus intensifiers (abnormally, absolutely)	2.22	2.25	2.19
#15 Second person pronouns (thou, thee, you)	2.30	2.39	2.19
#16 Third person pronouns (he, her, hers)	2.30	2.39	2.19
#18 Slang words (1sec, asap)	2.30	2.39	2.19
#22 Abbreviations (tl/dr, a/c)	2.22	2.40	2.01
#25 Discourse markers (and, and then, first)	2.30	2.39	2.18
#26 Discourse markers (certainly, definitely)	2.22	2.24	2.19
#30 Demonstrative words proximal words (here, this)	2.13	2.24	2.01
#32 Movement verbs proximal words (bring, come)	2.38	2.39	2.36
#33 Movement verbs distant words (take, go)	1.48	2.39	0.37
#34 English prefixes (a-, after-, anti-)	2.22	2.11	2.36
#38 Adverbial suffix (-ly)	2.14	2.10	2.18
#40 Nominalization suffixes (-hood, -ess, -ness)	2.38	2.53	2.19
#43 Weak starters (it, this, and, but)	2.14	1.95	2.36

Table 5.16: F-score results for the ablation experiments on Politeness features on a NER system using Twitter messages.

Features	Overall	Disease − Symptom	Drug
Only Politeness	2.12	2.81	1.31
#20 Nicknames (Abby, Addy, Alex)	1.82	2.67	0.82
#34 English prefixes (a-, after-, anti-)	1.14	1.13	1.15
#35 Neo classical English prefixes (afro-, anglo-, euro-, franco-)	0.77	0.43	1.15
#40 Nominalization suffixes (-hood, -ess, -ness)	1.52	1.83	1.15

Table 5.17: F-score results for the ablation experiments on Politeness features on a NER system using PubMed sentences.

The same exam using PubMed data shows that the set of Politeness features does not contain as many powerful features as when using Twitter data. In this case the biggest gains are seen when using Neo classical English prefixes (afro-, anglo-). Interestingly, it seems the nicknames include some drug keywords as removing these features cause a drop in performance in the detection of the medicines.

When looking at Table 5.16 and Table 5.17 we can also see that the only features that are duplicated across these tables are English prefixes (feature #34) and the nominalization suffixes (feature #40), showing that these features are the only Politeness features that work well when using either formal and informal texts.

To confirm these findings we also performed the same ablation experiments ran to obtain the results presented in Table 5.16 and 5.17 while also using the baseline and word2vec

Features	Overall	Disease − Symptom	Drug
Baseline + All Word2vec features + Politeness	73.80	55.79	92.41
#1 Colloquial words and expressions (stuff, just)	73.70	55.58	92.36
#2 Formal words (appear, ascend, assist)	73.65	55.49	92.36
#10 Exaggeration words (noteworthy, wonderful)	73.77	55.74	92.36
#11 Non exaggeration words (humdrum, mediocre)	73.67	55.52	92.36
#14 First person words (me, mine)	73.76	55.71	92.36
#20 Nicknames (Abby, Abe)	73.64	55.47	92.36
#22 Abbreviations (tl/dr, a/c)	73.68	55.60	92.32
#23 Discourse markers (anyway, great)	73.77	55.74	92.36
#25 Discourse markers (and, and then, first)	73.75	55.73	92.36
#26 Discourse markers (certainly, definitely)	73.74	55.69	92.36
#29 Forgiveness words (apologize, sorry)	73.65	55.51	92.36
#31 Demonstrative words distant words (there, that)	73.73	55.65	92.36
#36 Generic prefixes (re-, dis-)	73.27	54.78	92.28
#37 Verbalization suffixes (-ise, -ize)	73.71	55.62	92.36
#40 Nominalization suffixes (-hood, -ess)	73.74	55.70	92.32
#41 Hyphenated words ("-")	73.77	55.75	92.36
#43 Weak starters (It, This, And, But)	73.74	55.69	92.36

TABLE 5.18: F-score results for the ablation experiments using the baseline, word2vec and Politeness features on a NER system using Twitter messages.

Features	Overall	Disease − Symptom	Drug
Baseline + All Word2vec features + Politeness	80.02	67.34	93.20
#1 Colloquial words and expressions (stuff, just)	79.97	67.28	93.16
#2 Formal words (appear, ascend, assist)	79.94	67.28	93.11
#9 Cliche words (ace, almighty, bull)	79.98	67.31	93.16
#11 Non exaggeration words (humdrum, mediocre)	79.98	67.31	93.16
#18 Slang words (1sec, asap)	79.93	67.31	93.07
#20 Nicknames (Abby, Abe)	79.98	67.31	93.16
#27 Discourse markers (clearly, confidentially)	79.95	67.26	93.16
#39 Adjectivation suffixes (-ful, -able)	79.93	67.20	93.16
#40 Nominalization suffixes (-hood, -ess)	79.84	67.10	93.10

TABLE 5.19: F-score results for the ablation experiments using the baseline, word2vec and Politeness features on a NER system using PubMed sentences.

features.

The results in Table 5.18 show that in addition to the set of baseline and word2vec features some new politeness features, that did not show in Table 5.16, appear. Those features are *"#1 Colloquial words and expressions (stuff, just)*, *"#14 First person words (me, mine)*, *"#20 Nicknames (Abby, Abe)*, *"#23 Discourse markers (anyway, great)*, *"#29 Forgiveness words (apologize, sorry)*, *"#41 Hyphenated words ("-")*.

In the case of PubMed we can see that even if Politeness features alone only contributed 4 features that were providing gains to the system when used alone, in Table 5.17, now we observe in Table 5.19 that the number of Politeness features that seem to reduce the performance when removed are more. In particular we can see that features *"#1 Colloquial words and expressions (stuff, just)"*, *"#2 Formal words (appear, ascend,*

assist)", *"#9 Cliche words (ace, almighty, bull)"*, *"#11 Non exaggeration words (humdrum, mediocre)"*, *"#18 Slang words (1sec, asap)"*, *"#27 Discourse markers (clearly, confidentially)"*, *"#39 Adjectivation suffixes (-ful, -able)"* that did not appear before now produce gains when used in combination with the baseline features and the three sets of word2vec features.

5.3.3 Error analysis

In this section we present the error analysis we developed by taking fifty sentences at random from both PubMed and Twitter. We compare the results obtained when using the system that takes into account the baseline features in combination with the three sets of word2vec features (*"Baseline + Word2vec (PubMed and Tweets. No Arizona)"* in Table 5.12 and Table 5.13) and compare the labels provided by that system against the labels obtained by the system that also uses the set of linguistic features (*"Baseline + All Word2vec features + Biber (adapted)"* in Table 5.12 and Table 5.13) and also against the system using the set of politeness features (*"Baseline + All Word2vec features + Politeness "* in Table 5.12 and Table 5.13). We being by analysing the differences in the system when using PubMed sentences.

In our set of 50 PubMed sentences we find that there are 11 different labels (*"breast cancer"*, *"paclitaxel"*, *"Huntington's Disease"*, *"daunorubicin"*, *"parameter"*, *"mania"*, *"cortisone"*) in the annotations produced by the system using the linguistic features (*"Baseline + All Word2vec features + Biber (adapted)"*) and the annotations produced by the non-enhanced system (*"Baseline + Word2vec (PubMed and Tweets. No Arizona)"*). A detailed inspection of those differing labels tells us that the system using the set of linguistic features only got it right in the annotation of a drug (*"cortisone"*). Interestingly, the system annotated *"paclitaxel"* as a drug, which in fact it is, although it is not annotated in the gold (we only allowed the annotation of a closed set of drugs, and that is why the annotation of the token as a drug is taken as a miss), which clearly means that when using the linguistic features the NER system is able to capture that information. By observing the keywords around those tokens (*"__number__ mg of* ***cortisone.***" and *"__number__ cells and* ***paclitaxel*** *had much less"*) we find that for the labelled drugs the only Biber feature having a positive value is "WordLength" (9 and 10, respectively), and similarly, both of them are preceded by a keyword that is an English stop word ("of" in one case and "and" in the other), and the last word in the window of 3 words is in both cases "__number__"[16], and these elements seem to be triggering the annotation of a drug.

[16] All the integer numbers were replaced with this keyword when preparing the data sets.

Using the same set of sentences from PubMed on the system that takes into account the set of Politeness features we can see that the word "symptoms" is correctly recognized as "Outside" in one case (*"climacteric **symptoms** in"*) and as a part of a symptom in the other (*"adverse **gastrointestinal symptoms** induced"*), also, the preceding word is not detected as a symptom in the first case, while it is captured as a correct symptom in the second case. That could be telling us that the correct detection of "gastrointestinal" is in fact helping in the correct labelling of the token "symptoms". For that word, "gastrointestinal", we see that the set of Politeness features detect the use of suffixes, while in the case of "climacteric" no politeness feature is detected. For the token "undifferentiated" (*"acute **undifferentiated** leukaemia"*) we see that the prefix "un-" is recognized and that may be the cause for which that is recognized as an "Outside" token. The token "thromboembolic" (*"of **thromboembolic** and "*) may be correctly labelled by a combination of features and the correct detection of "and" as an stop word can contribute to the used features.

When looking at the differences in the produced labels for Twitter texts between the system using the linguistic features and the system not using them we see that for the 50 sentences we used only 12 tokens had different labels. These are *"raw pork shoulder"*, *"chicken pox"*, *"whilst"*, *"dizziness"*, *"crippling autism"*, *"agitation"*, *"avastin"*, *"cisplatin"*. In this case, the number of correctly annotated elements is higher as the system is able to detect *"raw pork shoulder"*, *"chicken pox"*, *"dizziness"*, *"autism"* as tokens describing diseases or symptoms, *"avastin"* as a drug token, and *"whilst"* as an "Outside" token. The only errors are *"crippling"*, labelled as a drug, *"agitation"*, which is not recognized as a disease nor symptom, and *"cisplatin"*, which is recognized as a drug, although the gold does not include it as such because of our annotation guidelines, although in fact that is a drug.

Starting with *"whilst"*, labelled as "Outside", we find that the set of Biber features capture it as an adverbial subordinator, which may be a clear clue for its correct annotation. For the rest of the tokens that the NER gets right we can see that in some cases there is an English stop word in the vicinity of the token being assessed (e.g. *"**a** raw pork shoulder inside"*, *"**have** chicken pox **which**"*, *"crippling autism **and**"*), which combined with the information by the feature "Word length", included in the set of linguistic features, can provide the needed information to label those tokens correctly. In the sentence *"a raw pork shoulder **inside**"* there is a place adverb (*"inside"*) that the linguistic features capture, and which can help to identify the limits of the symptom expressed using a metaphor (*"my brain felt like a raw pork shoulder inside a spinning fishbowl"*). In the case of the excerpt *"have chicken pox which"* the appearance of the verb *"have"* and the word *"which"* can help in the correct identification of the symptom. In particular, we have seen that the appearance of the verb "have" is often followed by

some symptom, and in all cases but this one, the baseline system (also using word2vec features) labels the following tokens as a symptom. It is probable that the occurrence of the word "which", captured as a WH-pronoun and as cached stop word, helps the system to correctly recognize the label for those tokens. The symptom "dizziness" is also captured by one linguistic feature as it is recognized as a "nominalization", which may be the reason why it is correctly labelled when using the set of Biber features. In the sequence *"by crippling **autism**"*, both "crippling" and "autism" are recognized as symptoms, and the reason could be that they come after another stop word and because of the length of the word.

When looking at the labels produced when using the set of Politeness features in Twitter texts we see that out of the 14 differences between that system and the system using the set of baseline features combined with word2vec features the chunks of texts that are correctly labelled have active politeness in four words *"a raw pork shoulder **inside**"*, where inside is recognized as a word with a neoclassical English prefix (in-). The word *"recurrent"* is recognized as another word with a prefix (re-) and that can help in its correct labelling as "Outside". In the case of *"mad **dizziness** wellbutrin"*, "mad" is recognized as an informal word, and that contribute to identify the symptom "dizziness". Another token that is correctly recognized seems to contain a prefix (em-), and this word appears after a conjunction "& " *"& **empty**,"*, which help in its correct recognition as a symptom.

These results confirm that linguistic features used in register studies can be implemented into pharmacovigilance systems, although not all of those features contribute with gains, and not all pharmacovigilance systems benefit from these features, showing that ***Hypothesis 4*** can not be accepted without further testing.

5.4 Discussion

We explored the classification of Twitter messages into first-hand drug user experience. For the task of selecting ADR data on the crowdsourced annotations Bayesian Generalized Linear Model (BGLM) was observed to be the model providing the overall highest F-Score among those tested, only surpassed by C50 when using the top 50% and the 100% of the features, although in terms of Informedness BGLM obtained the best scores all the time.

We also used the subset of the same data for which both the laymen and one expert agreed on the annotation for the fields *"First-class experience"*, *"Tweet written in English language"*, and *"Tweet about the drug"*. In this case BGLM obtained the best F-Score values, and also the highest Informedness measure, showing the predictive power of this model for this dataset.

For our last experiment we used the dataset where the annotations from two expert annotators were in agreement for the fields *"First-class experience"*, *"Tweet written in English language"*, and *"Tweet about the drug"*. In this experiment we observed that BGLM had the highest F-Score values, only matched by GLM when using the top 1% features. This is particularly interesting because the annotators were not laymen, and the data were collected during a different period and also using a different method, but the best performing model was the same as in the previous experiments.

We also observed that most models had a stable performance independent of the set of features. We also realized that *"SVM"* predictions were lower than the baseline in all the experiments, and *"Multi-Layer Perceptron"*, and *"Naive Bayes"* only scored above the baseline when using the dataset annotated by the two experts.

We believe this line of research can be meaningful given the volume of tweets that are constantly generated. Having a first filter to detect user reports on Twitter on the drug use can help in pruning valuable data since the beginning of other studies.

We also performed a number of experiments using a set of binary classifiers using SVM with a linear kernel, and improved the set of features we used when building different classifiers aimed at detecting first hand experience reports.

Those results show that the set of linguistic features have different impact, and in particular the results shown in Table 5.6 evidence that those features, implemented as proposed by Biber or in our custom expansion, can have a negative impact when using Twitter data on a classifier aimed at detecting "Benefit" and "Outcome-negative" relations. Our expanded sets of features seem to provide some significant contributions in the detection of "First-hand experience reports" as well as in the detection of tweets containing drugs and related outcomes ("Any Outcome").

Moreover, we also noticed that the use of the set of politeness features that we prepared based on linguistic studies [80], behave much better as these produce significant gains in all cases except when used to detect reports containing negative outcomes ("Negative Outcome") related to the drug use.

When looking at the impact of those features in binary classifiers using PubMed sentences we see that the different implementations of Biber features have a non-significant

minor gain on the detection of "Beneficial Outcome" sentences. Similarly to what we observed before these features worsen the classifier in the other two cases. In this case the power of Politeness features is almost non-existent as the results are very close to the ones scored by the baseline classifier.

The different sets of Biber features showed no conflict. However, when using Politeness features in combination with other set of features the gains were much lower.

When obtaining the best performing Biber features we found the length of the words, the type/token ratio, and to a lower extent, the count for the number of adverbs, different forms of the verbs, the count of attributive adjectives and prepositional phrases as the best. From our expansion on Biber's features the best performing features were the count for integer numerals, prepositions, and the use of the auxiliary verbs "have" and "be" (in Twitter and PubMed sentences, respectively), while from the set of Politeness features we observed the use of neo-classical English prefixes, nouns' suffixes, and slang words as the best performing features.

Our experiments using a NER system aimed at detecting drugs, diseases and symptoms showed that our new sets of features were performing very poorly when used alone. Even if when testing the classifiers presented in the previous section the set of Politeness features did not show an important contribution, in NER experiments we observed that these features were able to improve all the other results. Our adaptation of the set of Biber features provided some contribution to NER systems, but these were less powerful than the gains produced by Politeness features.

When extracting the best performing elements from our custom set of Biber features we observed that a large number of them contributed to the detection of the entities in tweets, although a fewer number of features appeared to contribute when using PubMed texts. In both cases the length of the words, the detection of infinitives and the identification of hedges were contributing to the task.

When looking at the set of Politeness features we also observed that the number of features contributing to the task when using tweets was much larger than in the case of using PubMed texts. The features that appeared in both cases were the ones in charge of detecting English prefixes (a-, after-, anti-) and Nominalization suffixes (-hood, -ess, -ness). We also found that our lexicon of nicknames, expected to be useful in informal texts, was in fact contributing to the detection of the entities in academic sentences.

Chapter 6

Conclusions

In this thesis we have explored different aspects of formal and informal texts containing information on the same topic of pharmacovigilance.

Prior to start our linguistic studies we curated the corpora, and using our first resulting data set we created a pharmacovigilance classifier for obtaining tweets reporting first-hand experience drug use. We discovered that most of the best performing features were linguistic features such as unigrams and character n-grams, and identified following work to be performed to create a better corpora as we discovered there were a number of drug use reports that the system would not be able to detect in case it would only detect first-hand experience reports, meaning that we would be missing valuable information from the excluded reports. More importantly, we discovered that some of the keywords used to create our corpus were much more frequent than others, causing some bias in the data set.

To overcome those two issues so that our register study could be meaningful we improved our message gathering strategy obtaining a greater percentage of sentences of interest also discarding most non-informative messages in an automated way. The obtained list of messages from both PubMed and Twitter was then filtered by two expert pharmacists.

With that data set in place we were able to start exploring our hypotheses and perform our study on two registers differing in the level of formality but on the topic of pharmacovigilance.

We used that data set to answer **Hypothesis 1** and understand the similarities and discrepancies in terms of the information contained in our tweets data set and in the set of PubMed sentences. We did so by evaluating the topicality of the informal messages, i.e. Twitter messages, and in the messages using a more formal register, i.e. PubMed drug use reports.

We found that most of the keywords characterizing the contents in Twitter were related to the language used in an informal setting, i.e. informal register, as we found frequent abbreviations and verbs frequently seen in phrasal verb constructions. These facts caused that our attempts at labelling the data using Wikipedia did not obtain the expected results as in most cases the pages that were retrieved and used to label Twitter topics contained noisy keywords unrelated to the medical domain. On the other hand, the set of pages we obtained when labelling the set of topics extracted from PubMed sentences were more in-domain.

This first part of the study assessed the discrepancies from a contents perspective as well as from a topical perspective studying the underlying labels that would be used to characterize the used texts. We found that noisy keywords in Twitter were causing the low levels of similarity between the topics. Another finding was that in almost all cases PubMed keywords were in-domain keywords, and Twitter keywords were very generic keywords and acronyms. That fact evidenced that after some pruning the extracted keywords would provide more meaningful results and a higher agreement in terms of the extracted labels that were used to characterize the samples. Those findings led us to consider that a further analysis of the politeness features, the use of taboo words and orthographic variations found in Twitter would provide more insights on drug reports differences.

We then followed a purely linguistic approach and used Biber's MD analysis on a data set composed of generic tweets and drug-related tweets. By studying these two different data sets we answered **Hypothesis 2**, aimed at discovering if tweets reporting drug use and generic tweets share most of their linguistic traits, and found that the set of drug-related tweets had some key characteristics expected to be found in more formal texts. In particular, we noticed that Biber's first dimension, "involved versus information productions", exposed the drug-related tweets as being more informational than the generic set of tweets, and in terms of individual factors, i.e. the features used to compute Biber's dimensions, the factors we found more often in drug-related tweets than in generic tweets were the following: the use of seem/appear verbs, the use of amplifiers ("absolutely", "altogether", "completely"...), and the use of the sentence relative "which"; other minor but constant differences were the use of past participial whiz constructions (past participle combined with the deletion of a Wh-word plus a form of be, quite often "is"), and possibility modals ("can", "may", "might" or "could"). Besides characterizing the factors related to drug use reports tweets we also identified generic tweets to make frequent use of "wh clauses" (e.g. "I believed **what** he told me"), stranded prepositions, discourse particles ("well", "anyway"...) and place adverbs ("above", "around"...).

Hypothesis 3 aimed at understanding the similarities and differences in tweets and PubMed texts on drug use reports. We performed our assessment by undertaking the same analysis that we developed to test **Hypothesis 2**, although in this case the data was of different nature and differed on the use of the linguistic register rather than in the topicality of the contents. In this case too we observed that dimension one, same as what happened in the previous tweet data sets comparison, had some differences between the set of tweets and the set of PubMed sentences, although the most different dimensions were dimension four ("Overt Expressions of Persuasion") and dimension five ("Abstract versus Non-Abstract Information"), showing that some differences that we would expect to see between these datasets, namely the difference in informativeness on the difference in the on-line informational elaboration, had diluted.

Although in the comparison we performed to test **Hypothesis 2** using only tweets data sets the differences stayed most of the times in 2-3 times the frequency of some factors between data sets, when testing **Hypothesis 3** we observed greater differences and some factors appeared well above 5 times more often in one data set than in the other. In particular, the use of the verb "Do" as a pro-verb appeared 75 times more often in tweets, same as the frequency of "First person" and "It" pronouns (15 and 13 times, respectively).

Other outstanding differences appearing much more often in tweets were the use of the analytic negation ("not") and the adverbial subordinator "because" (8 and 9 more times, respectively). The use of emphatics ("so", "such", "a lot"...), amplifiers ("absolutely", "totally"...), which we also observed when studying our drug-related data set against the generic set of tweets, and stranded prepositions also appeared more than 5 times more often in Twitter than in PubMed in our sample of sentences.

Interestingly, we noticed that the use of sentence relatives (e.g. "which"), a clear sign of specialized discourse, was more frequent in the drug related set of tweets than in the generic set of tweets, and in this case too, it appeared much more often in the set of PubMed sentences than in our drug-related set of tweets. In the same line are the frequencies of the use of past participial WHIZ[1], and the use of the phrasal coordination ("and") as it appeared more often in the drug-related set of tweets than in the generic set of tweets, and also appeared more often in PubMed texts than in the drug-related set of sentences.

Our last piece of work, covered when answering **Hypothesis 4**, was the study of the gains produced after the use of register features in a pharmacovigilance classifier. We found that for different data sets and tasks the contributions varied, although in general

[1]Past participle combined with the deletion of a Wh-word plus a form of be, quite often *"is"*, thus called *"whiz"* as a monosyllabic variant of *"Wh-is deletion"*.

terms we were able to obtain useful information from our newly generated set of register features, both from the set of Politeness features and from our custom expansion of Biber original features. The bigger gains were seen when detecting first-hand experience reports on the drug use feeding the system with tweets, and also when classifying messages as containing any drug and a symptom or disease related to the drug intake. For finer grain tasks, detecting sentences where the symptoms were positive or negative, the classifiers were not able to get any useful information from our register-related set of features.

Also, as part of **Hypothesis 4** we built a NER system for detecting drugs, diseases and symptoms in our set of tweets and PubMed sentences. We observed that the use of a custom adaptation of Biber features was not enough to produce noticeable gains, although the use of Politeness features was able to provide gains that increased the top-performing systems when using Twitter and PubMed messages.

By taking into account the results presented in Chapter 5, and analysing the sentences and the tokens where the newly added set of features contribute towards the correct identification of the appropriate labels we observe that features used to capture the use of differences in the use of the linguistic register, and the formality of the texts can contribute to NLP systems, and in particular the set of best performing features for Classification and NER systems are playing a role in the identified elements, although not all the best-performing elements appear in all the elements that are correctly labelled when using our proposed features to detect the formality of the text. We have seen that the use of certain verbs (e.g. past or present forms), the use of personal pronouns (first, second or third person), the use of time and place adverbs, the detection of profanity words, and the use of different English prefixes and affixes, among others, contribute towards improving the performance of the system with significant gains in the detection of first-hand experience tweets and when labelling sentences in Twitter containing any type of report on the outcomes related to the drug intake.

We also observed a similar trend in the NER systems when using the set of Politeness features as these features proved to be useful towards improving the accuracy of the systems in both Twitter and PubMed, although those gains did not show to be significant in the used set of sentences. We observed that most of the best-performing politeness features we identified appeared in the tokens that the system was able to correctly recognize, which shows that those features contribute to NER systems.

A final thought on the features we assessed is that these are not particularly hand-crafted for a pharmacovigilance setting, and even if our experiments were performed on data from that domain we believe the set of features we assessed and adapted for NER and classification systems are pervasive in both formal and informal English texts, and

its use can be of help in different tasks that do not take account those utterances. We also observed that not all tasks benefit from their use, and our observation showed that classification systems using informal texts can benefit from the use of an adaptation of Biber features while NER systems using both formal and informal texts improve when using politeness features.

6.1 Future work

We found that drug related tweets are more informative than generic tweets, but these drug related tweets still have many of the noisy features that other researchers noticed in generic tweets. These noisy features could be cleaned in an automatic way to improve the characterization of these messages, which could dilute even more the differences found between drug-related tweets and PubMed sentences. That normalization could also prove useful to show other elements where formal and informal reports differ.

Given the drugs mentioned in the messages we gathered from Twitter and PubMed had no correlation another follow-up study could assess if the differences between formal and informal drug use reports that we found were affected by the sets of drugs we extracted. We aimed at a balanced data set but we understand that the different distribution of drugs, and the symptoms related to the intake of those drugs, may have introduced another element of variability. Ideally, a new study on a single drug would help in answering this question. Similarly a new study using messages on the same topic but from a larger number of sources (e.g. internet forums, academic books, blogs and leaflets) could help in understanding the way in which the reports vary in terms of the register.

In our study we pointed out that the set of features Biber proposed may not be enough for explaining all the variability we can find in messages on the same topic but written using different levels of formality, and other linguistic elements such as politeness features, rethorical expressions, or phrasal verbs would need their own assessment.

A follow up study more focused on the information being conveyed in those sentences, and to understand which are the type of pharmacological reports that are most dissimilar between formal and informal sources would be very interesting to discover if machine learning systems that only use one of these sources of information is missing some kind of pharmacological reports.

Lastly, we have seen that register-related linguistic features can contribute positively to a classifier, and it would be needed to check if our findings can be generalized to other systems. Additionally, different linguistic features such as stylistic features or other sociolinguistics features could be explored to confirm the extent to which linguistic

studies can contribute to the field of NLP as after seeing that Politeness features provide gains to NER classifiers it seems clear that the work in the area of linguistics has potential for contribution in pharmacovigilance tasks like the ones presented in this work.

Appendix A

Expert annotator guidelines for annotating first-hand experience tweets

A.1 Document information

DRAFT VERSION 1.0 (March 2014)

Authors: Nestor Alvaro, Mike Conway and Nigel Collier

A.2 Introduction

This report is intended to guide annotators in marking up the drugs as well as people's attitudes to the drugs for use in text processing systems.

A.2.1 Scope note

This research memo is a product of a research project intended to guide annotators contributing to this project in understanding symptoms and the drugs that treat or cause them as well as people's attitudes to the drugs for use in text processing systems.

A.2.2 Purpose

This document describes the criteria for annotating texts to understand symptoms and the drugs that treat or cause them as well as people's attitudes to the drugs such as opinions, requests for advice or news on these drugs. In this first stage of annotation we will look only at texts provided by Twitter after requesting a data sample complying where certain substances (mainly the drugs of interest) were mentioned. Most texts will report on drugs of interest, but not all texts have to report on these drugs of interest.

A.2.3 Focus

Our main purpose in annotation is to identify both people's attitudes to the drugs and also the symptoms and the drugs that treat or cause them. The focus will be on both the tweet contents (drug, symptoms, genre, sentiment... fields) and also on the demographic information from the user posting it (gender, adult, country... fields). Not all of this information may be present in each text. For example sometimes the symptoms before or after taking the drug will not be mentioned it will be left empty. In such cases we will record that there's information on the drug but some information will be missing from the record.

A.3 Flow Chart

This is the Flow Chart corresponding the actions the person tagging the texts would perform.

As you can see on the Figure A.1, if the tweet is not in English or it's not about the drug the next thing to do is to move to the next tweet as there's no need to continue tagging the current tweet.

A.4 How to perform the tagging.

A.4.1 Understand the document format

You'll be given a table. Each row has 6 fields already filled (See below the section called **"List of given fields"**) which contain information from a tweet. Following, there are the fields that have to be filled by you using the information contained on the first 6 fields from that same row.

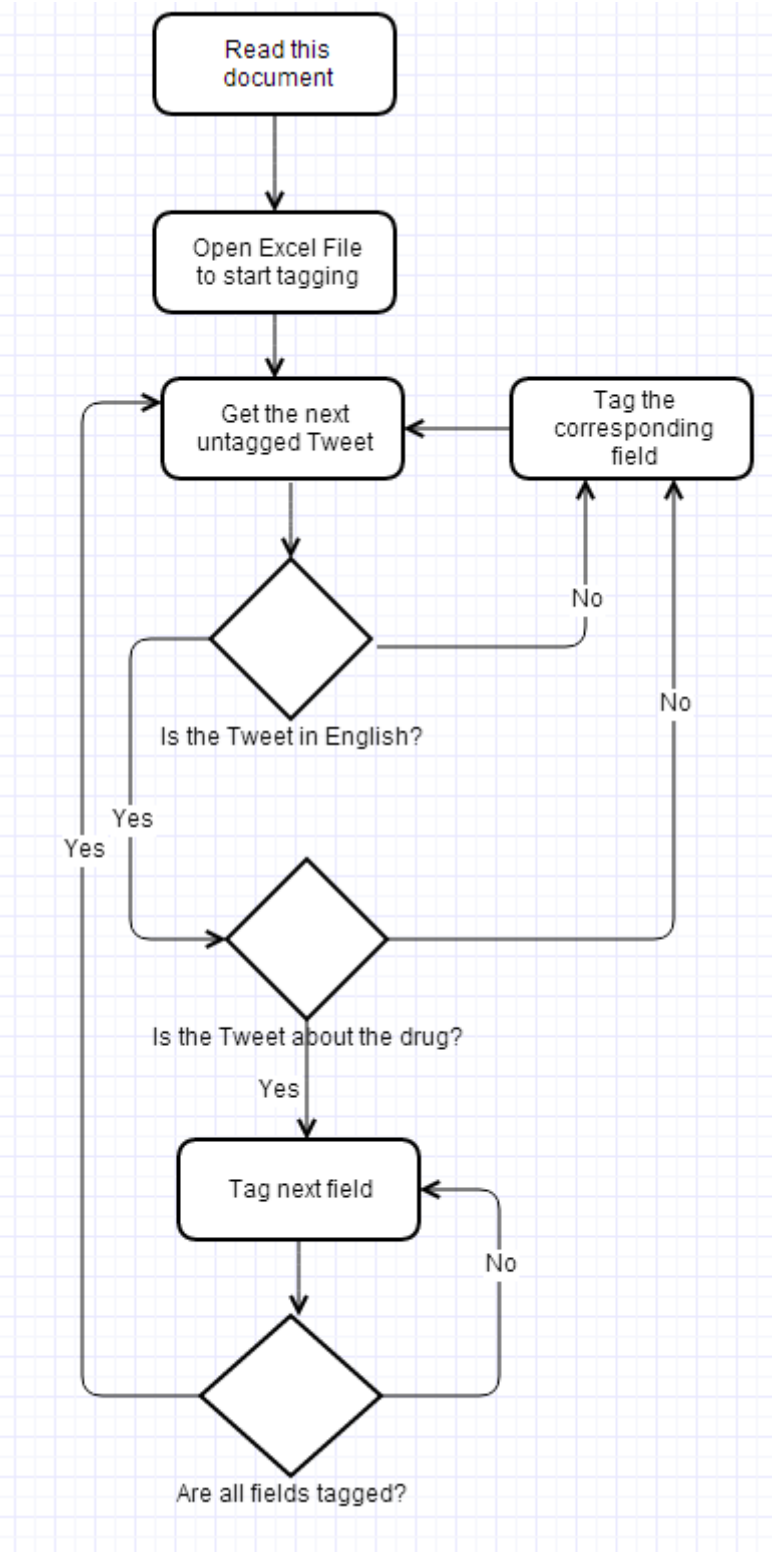

FIGURE A.1: Flowchart describing the annotation sequence used in first-hand experience tweets.

A.4.2 Understand each field

In order to fill the document we've created a table to explain how to fill each field. In Table 3 (by the end of this document) you will see the list of fields that you will be asked to annotate. This table has 3 columns. The first one corresponds to the name of the field as it's listed on the document that you have to fill. The second field within the table is a brief explanation of what the first field means. The third field contains the values that you can input when filling the document. Apart from the explanations within the table there are a few clarifications that have to be made:

- You shouldn't make any guess nor assumption. Just obtain the information by using the available data from the first 6 fields.
- In some fields there's the "empty" value listed among the available values that can be entered. If you are not sure about what to enter on that field, please leave it blank.
- Please note that in most of the cases there's a closed list of values, but for some fields there isn't a closed list of values. Also, more than one value can be entered within some fields.
- The fields "Symptoms causing the use", and "Symptoms after the use" have to use strings from a controlled vocabulary. To see how to fill these fields, please read next section.

A.4.3 Fill the fields

On each row, use all the given information (See below the section called **"List of given fields"**) to fill the fields. As stated above, the first fields contain the information that you will use to fill the rest of the fields (See below the section called **"List of fields to be filled"**) on the same row.

- Keep in mind that each row is independent, so you only have to make use of the information within the first fields. No information obtained from other rows should be used when filling a row. Just use the information from current row.
- The fields "Symptoms causing the use", "Symptoms after the use" and "Country" have to be filled in a particular way, explained next on the following section: "Special fields".

Excel sheet label	Synonyms
Adderall	Amphetamine mixed salts, amphetamine and dextroamphetamine, amphetamine salt
Ritalin	Concerta; Methylphenidate; Methylin; Metadate; Equasym XL; Daytrana; Phenida; Attenta; Hynidate; Focalin; Attenade; Quillivant; methyl phenyl(piperidin-2-yl)acetate
Modafinil	Modafinilo, Modafinilum, Moderateafinil, Modiodal, Provigil, Sparlon, Alertec, Modavigil, Modalert,()-2-(benzhydrylsulfinyl)acetamide
Adrafinil	CRL-40028, Olmifon, CRL 40028, (RS)-2-benzhydrylsulfinylethanehydroxamic acid
Armodafinil	Nuvigil
Citalopram	Celexa
Escitalopram	Lexapro, Cipralex
Paroxetine	Paxil, Seroxat
Fluoxetine	Prozac
Fluvoxamine	Luvox
Sertraline	Zoloft, Lustral

TABLE A.1: List of drug names along with the synonyms.

A.4.4 Special fields

About the drug

The excel sheets to be annotated are "Adderall", "Ritalin", "Modafinil", "Adrafinil", "Armodafinil", "Citalopram", "Escitalopram", "Paroxetine", "Fluoxetine", "Fluvoxamine" and "Sertraline".

On each excel sheet we only care about one drug and all the synonyms for such drug. This means that only in case the drug or a synonym for such drug are mentioned within the tweet text, the value of the field **"About the drug?"** would be **"1"**. Otherwise it would be left empty. Table A.1 shows the list of each drug along with its synonyms.

As an example. If we are annotating tweets on the excel sheet named **"Adderall"**, a tweet that is considered to be "About the drug" ("About the drug?" field would be annotated as **"1"**) would be a tweet that mentions any of the following drugs on the tweet text:

- **"Adderall"** or
- **"Amphetamine mixed salts"** or
- **"amphetamine and dextroamphetamine"** or

- **"amphetamine salt"**

As you can see, these are the drugs on "Excel sheet label" and on "Synonyms" on the table above. If none of these drugs appears on the tweet text, the field **"About the drug?"** is left empty and the annotator doesn't have to continue annotating that tweet.

It's important to keep in mind what drugs are considered on each excel sheet. For example, if the annotator is annotating a tweet on **"Adderall"** sheet, in case such tweet just mentions another drug of this study such as **"Concerta"** (which is a synonym of **"Ritalin"**) and none of the drugs mentioned above ("Adderall", "Amphetamine mixed salts", "amphetamine and dextroamphetamine", "amphetamine salt") this tweet wouldn't be considered to be **"About the drug"**, so **"About the drug?"** field would be left blank in this case. On the other hand, if the annotator would be annotating the same tweet in the excel sheet named **"Ritalin"** this same tweet would have the field **"About the drug?"** annotated as **"1"** as in this case case **"Concerta"** is a synonym of **"Ritalin"**.

CUI Identifier

For filling "Symptoms causing the use" and "Symptoms after the use" you have to find all the terms related to symptoms after and before the drug usage. Changes on the behaviour or how the user feels before and after taking the drug will show this symptoms. Once this symptoms have been found they have to be entered in a normalized way. To do so we have to use **Consumer Health Vocabulary**'s website: http://consumerhealthvocab.chpc.utah.edu/CHVwiki/index.jsp?orgDitchnetTabPaneId=searchPane There, on the **"CHV Entry Search"** search box mark the option "Term" ("Search by" option) in case it's not selected. Then enter the term on the search box.[1] Finally, just click on the "Search" button.

In Figure A.2 there's the example where we entered the term *"fatigue"*:

Once you get the results you'll have to use the **"CUI"** value for such symptom.

Figure A.3 shows the **"CUI"** value for "fatigue", which is "C0015672".

In some cases the word may not appear if it's not searched in this way. In case of "fatigued" the search showed results (Also pointing out that the **Consumer Health Vocabulary** preferred name is the noun form of the word, "fatigue"). On the other

[1] The **Consumer Health Vocabulary** "Preferred Name" is the noun form of the word, so instead of using adjectives (eg: "motivated" or "fatigued") you'll have to enter the noun corresponding to such adjective ("motivation" or "fatigue").

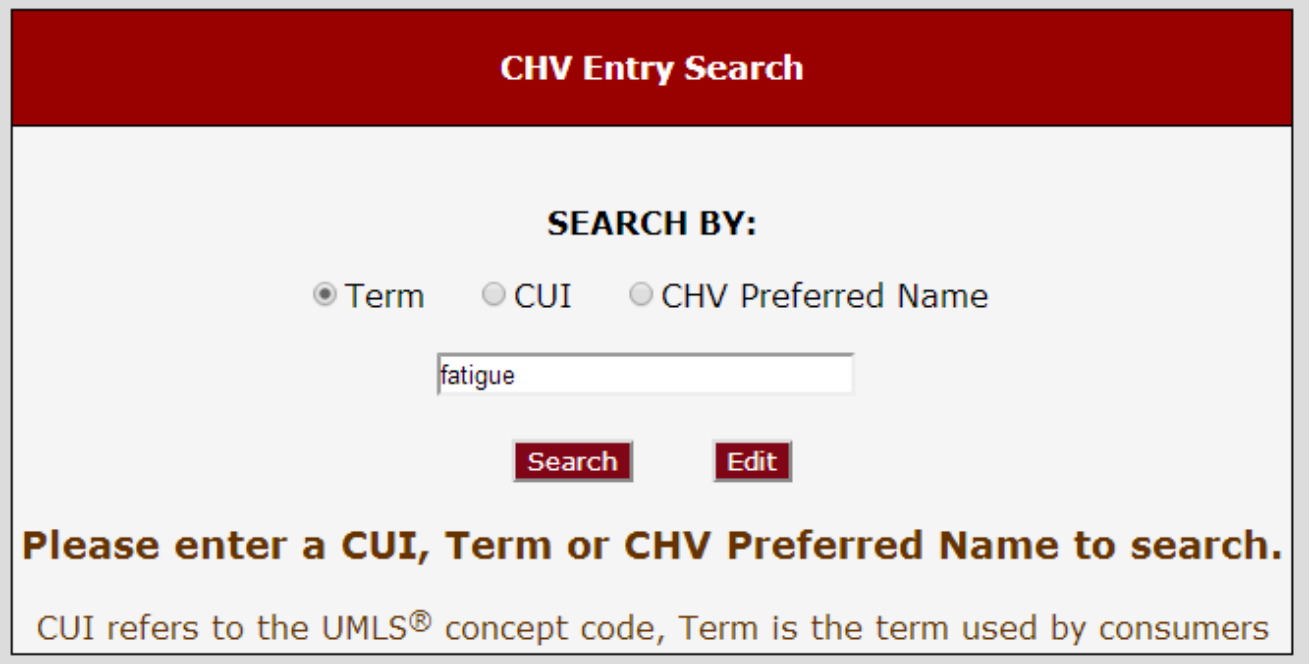

FIGURE A.2: CHV Entry Search Box.

CUI	Term	CHV Preferred Name	UMLS Preferred Name	Explanation of Consumer Term	Disparaged
C0015672	fatigue	fatigue	fatigue		no

FIGURE A.3: CHV results for the term fatigue.

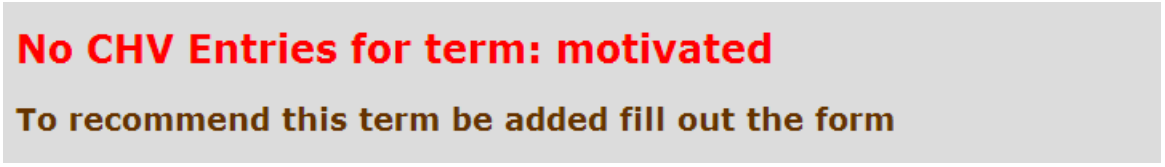

FIGURE A.4: No results messages in CHV.

hand, in case of using the term "motivation" Figure A.4 shows that such search doesn't provide any result:

Country Code

To get the country code the annotator will be asked to only use the information from the "location" field (3rd column on the excel sheet) to obtain the country information. The country codes that can be entered are the "ISO 3166-1 alpha-2" code, which are two-letter country codes to represent countries, dependent territories, and special areas of geographical interest.

UM	United States Minor Outlying Islands	1986	.um	ISO 3166-2:UM	Consist Navass
US	United States	1974	.us	ISO 3166-2:US	
UY	Uruguay	1974	.uy	ISO 3166-2:UY	

FIGURE A.5: Country code selection.

Tag	Explanation
Tweet date	Date when the tweet was published.
Username	Name of the user.
Location	Location where the user is based.
Tweet text	Text for the tweet.
Hashtags (#)	List of hashtags (if any) within the tweet.
Mentions (@)	List of mentions (if any) within the tweet.

TABLE A.2: List of fields that are provided with data to the annotators.

You can get a list of available "ISO 3166-1 alpha-2" code on this URL http://en.wikipedia.org/wiki/ISO_3166-1_alpha-2#Officially_assigned_code_elements. The "ISO 3166-1 alpha-2" code that has to be used on the column "Code". In order to get the code the annotator has to look for the country on the column named "Country name" and then get the corresponding two-letter "Code" for such country.

In the case of having a Tweet that is from United States of America, the country code that has to be entered on the "Country" cell would be the two-letter code "US", as shown in the Figure A.5. If the country is unknown, please leave this field empty.

A.5 Fields on the Excel Table

A.5.1 List of given fields

These are the fields that are completed and the annotator can use to fill the rest of the fields. The list of these fields is shown on Table A.2.

A.5.2 List of fields to be filled

The fields that can have more than one value are:

- Symptoms causing the use.
- Symptoms after the use.
- Multiple substances.

When you enter more than one value on any of these cells, please separate them by a semicolon (;).

Next, Table A.3 shows the available tags, followed by a brief explanation and the values that each tag can take:

Tag	Explanation	Values
English?	Is the tweet in English language? If it's not in English leave this field empty. Otherwise, "1" (e.g. a Tweet containing only mentions to other users would be considered to be in English). ***In case this field is "empty", the annotator should move to the next line.***	"1" or empty field.
About the drug?	Is the message about the drug or not about the drug? To see the drug of interest, please see **"About the drug"** section. ***In case this field is "empty', the annotator should move to the next line.***	"1" or empty field.
Symptoms causing the use	What categories of symptoms are causing the use of the drug? Use the CUI identifier, as explained on **"CUI Identifier"** section. *More than one value can be entered.*	*"CUI Identifier"* or empty field.
Symptoms after the use	What categories of symptoms are resulting from the use of the drug? Use the CUI identifier, as explained on **"CUI Identifier"** section. *More than one value can be entered.*	*"CUI Identifier"* or empty field.
Genre	Genre in which the drug is mentioned (choose among the following).	

Tag	Explanation	Values
	More than one value can be entered, so as many fields as possible options are present on the excel file: • Experience (first-hand): Reports a personal use of the drug • Experience (other): Reports someone elses use of the drug. • Activism: Alarm or call for change in the drug policy • Cultural reference: Song lyric, movie title, etc • Humor: Formulaic joke, bumper sticker, etc. • News: News item • Info/resource: Factoid or informational resource. • Marketing: Sale of the drug product/accessory. • Opinion: Personal opinion related to the drug. The fields that match will be marked with "1" to indicate that the option is selected. These fields have the headers in **blue** on the excel file.	"1" or empty field.
Multiple substances	List containing all mentioned substances (drugs or not) taken with the drug (Eg: Coffee, prozac, orange juice...). Each substance will be separated by a semicolon (";"). So in the previous example this field would have the following string: **coffee; prozac; orange juice** If no other drug is mentioned (just the drug of interest or any of its synonyms), the value of this field will be empty.	Name of the substances or empty field.

Tag	Explanation	Values
Adult?	Is the drug user an adult? "1" If the user is an adult "2" If the user is a child/adolescent or empty field if we aren't sure	"1", "2" or empty field.
Male?	Is the drug user male? (from the User name) "1" In case the use is male "2" In case the user is female or empty field if we aren't sure	"1", "2" or empty field.
Sentiment	Is the author positive about the drug, negative, or neither positive nor negative (neutral) in terms of sentiment? "1" If the author is positive about the drug. "2" If the author is negative about the drug. or empty field if the author is neither positive nor negative about the drug (neutral sentiment).	"1", "2" or empty field.
Theme	More than one from the following list can be used, so as many fields as possible options are present on the excel file: • Pleasure: Drug usage as a pleasurable activity. • Craving: Possibly related to stress relief. • Disgust: Drug users or usage as repulsive. The fields that match will be marked with "1" to indicate that the option is selected. These fields have the headers in green on the excel file.	"1" or empty field.

Tag	Explanation	Values
Country	Country where the user is based or empty field. Use the "ISO 3166-1 alpha-2" two-letter code, as explained on "Country code" section.	"ISO 3166-1 alpha-2" code or empty field

TABLE A.3: List of fields that are to be filled by the annotators

Appendix B

Laymen annotator guidelines for annotating first-hand experience tweets

B.1 Document information

DRAFT VERSION 1.0 (May 2014)

Authors: Nestor Alvaro

These guidelines are an adapted version of "*A Expert annotator guidelines for annotating first-hand experience tweets*" to be used in an on line questionnaire.

B.2 Instructions

Our main purpose in this annotation is to identify both peoples attitudes to the drugs and also the symptoms before and after taking the drugs. For this study, the DRUGS OF INTEREST are "Adderall", "Ritalin", "Modafinil", "Adrafinil", "Armodafinil", "Citalopram", "Escitalopram", "Paroxetine", "Fluoxetine", "Fluvoxamine", "Sertraline", and also the following synonyms of each drug:

- "Adderall": "Amphetamine mixed salts"; "amphetamine and dextroamphetamine"; "amphetamine salt"

- "Ritalin": "Concerta"; "Methylphenidate"; "Methylin"; "Metadate"; "Equasym XL"; "Daytrana"; "Phenida"; "Attenta"; "Hynidate"; "Focalin"; "Attenade"; "Quillivant"; "methyl phenyl(piperidin-2-yl)acetate"
- "Modafinil": "Modafinilo"; "Modafinilum"; "Moderateafinil"; "Modiodal"; "Provigil"; "Sparlon"; "Alertec"; "Modavigil"; "Modalert;()-2-(benzhydrylsulfinyl)acetamide"
- "Adrafinil": "CRL-40028"; "Olmifon"; "CRL 40028"; "(RS)-2-benzhydrylsulfinylethanehydroxamic acid"
- "Armodafinil": "Nuvigil"
- "Citalopram": "Celexa"
- "Escitalopram": "Lexapro"; "Cipralex"
- "Paroxetine": "Paxil"; "Seroxat"
- "Fluoxetine": "Prozac"
- "Fluvoxamine": "Luvox"
- "Sertraline": "Zoloft"; "Lustral"

Both the drugs and the stated synonyms for such drugs are what we call the "DRUGS OF INTEREST" for this study.

In all cases you are expected to ONLY use the information provided within this questionnaire, so please refrain from using external sources of information such as Wikipedia or Google.

All the texts have been obtained from Twitter.

ONE MORE ADVICE (AUTHOR of the tweet Vs USER OF THE DRUG) : Read the questions carefully as not all questions are asked about the AUTHOR of the tweet (some questions are asked about the USER OF THE DRUG). The user of the drug can be the author of the tweet too, but it doesnt have to be the same person. Finally, remember that in Twitter a mention to a user is done using the "@" mark before the users name, which means that some mentions will be addressed to specific users, but other mentions will be about the drug itself (in order to try to get this point right try to understand whether what is addressed is a "person" or a "substance").

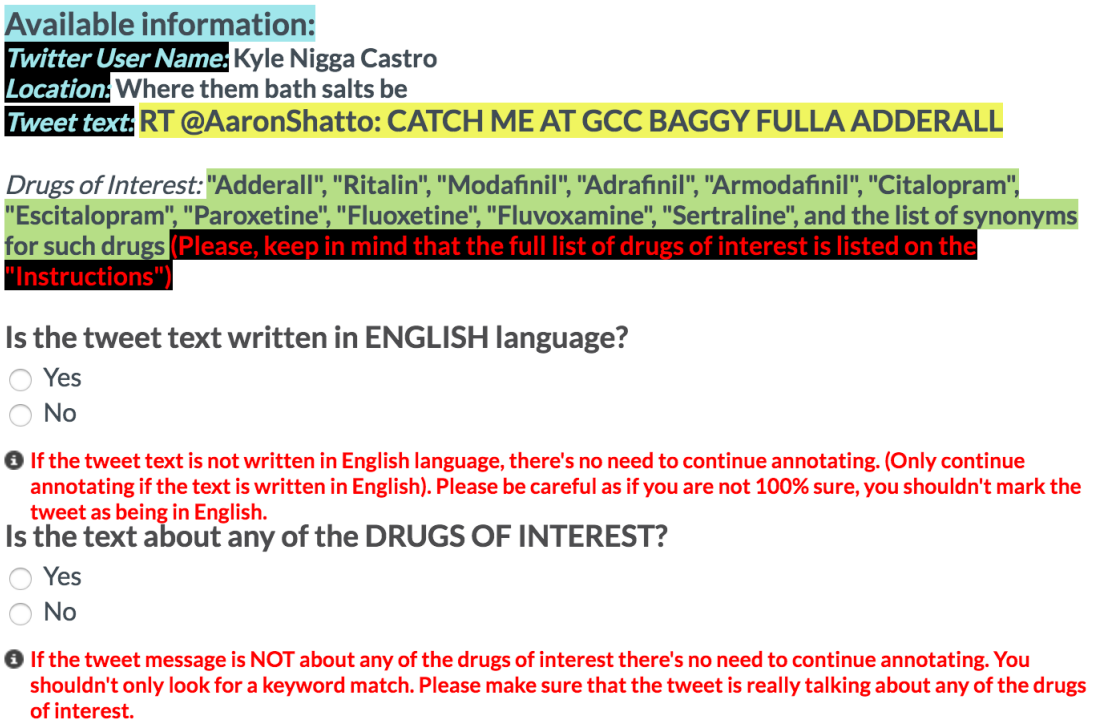

FIGURE B.1: Part 1 in the questionnaire to the laymen.

B.3 Questionnaire

This section contains a sample including all the fields we requested the annotators to complete for a given tweet as shown in the annotation tool. The following screen shots (Figures B.1 to B.11) cover all the fields that were available for annotation.

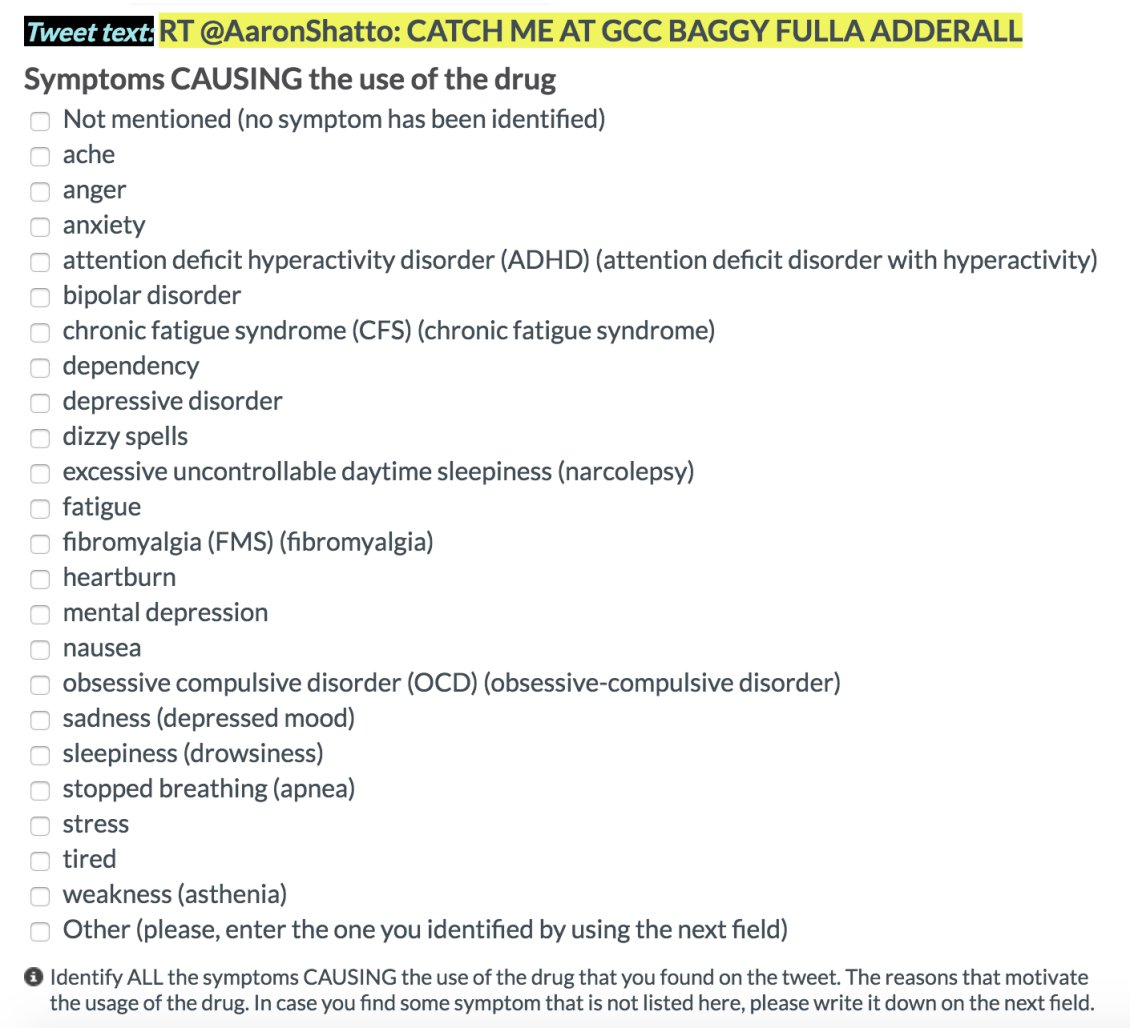

FIGURE B.2: Part 2 in the questionnaire to the laymen.

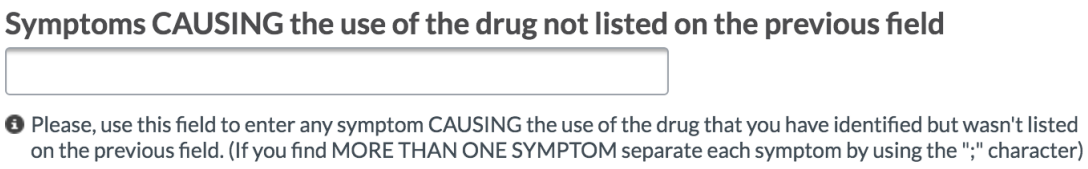

FIGURE B.3: Part 3 in the questionnaire to the laymen.

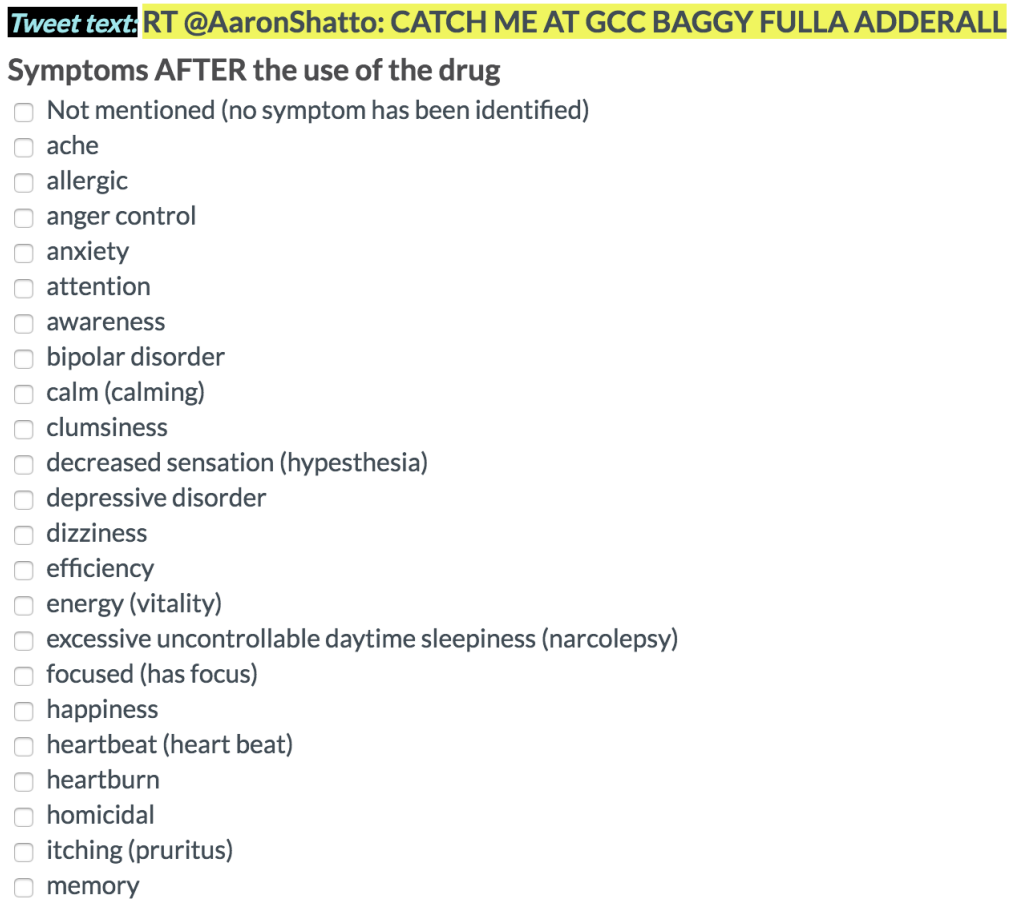

FIGURE B.4: Part 4 in the questionnaire to the laymen.

- [ ] memory impairment
- [ ] mental concentration (concentration)
- [ ] mental depression
- [ ] numbness
- [ ] poor concentration
- [ ] rapid heartbeat (tachycardia)
- [ ] sadness (depressed mood)
- [ ] seizure (seizures)
- [ ] sexual desire (libido)
- [ ] sleepiness (drowsiness)
- [ ] sleeplessness (insomnia)
- [ ] suicidal ideation (suicidal thoughts)
- [ ] suicide
- [ ] tired
- [ ] vigilance
- [ ] Other (please, enter the one you identified by using the next field)

Please indicate ALL the symptoms AFTER the use of the drug that you found on the tweet. In case you find some symptom that is not listed here, please write it down on the next field.

Symptoms AFTER the use of the drug not listed on the previous field

Please, use this field to enter any symptom AFTER the use of the drug that you have identified but wasn't listed on the previous field. (If you find MORE THAN ONE SYMPTOM separate each symptom by using the ";" character)

FIGURE B.5: Part 5 in the questionnaire to the laymen.

Tweet text: **RT @AaronShatto: CATCH ME AT GCC BAGGY FULLA ADDERALL**

GENRE in which the drug is mentioned.

- [ ] First-hand experience (The writer is taking the drug. He's telling a personal experience.)
- [ ] Other's Experience (The writer is reporting about some else's experience with the drug)
- [ ] Activism (The tweet invites to take part on some kind of movement or relates some act of activism)
- [ ] Cultural reference (The tweet mentions something specific to the popular culture.
- [ ] Humor (The tweet is generally funny)
- [ ] News (The writer is reporting on some recent news)
- [ ] Info/resource (The writer is providing useful information like advices, tips or tricks)
- [ ] Marketing (The writer is trying to sell some product)
- [ ] Opinion (The writer is giving his opinion,thoughts or point of view on something)

What's the type of tweet?. Check all that apply.

FIGURE B.6: Part 6 in the questionnaire to the laymen.

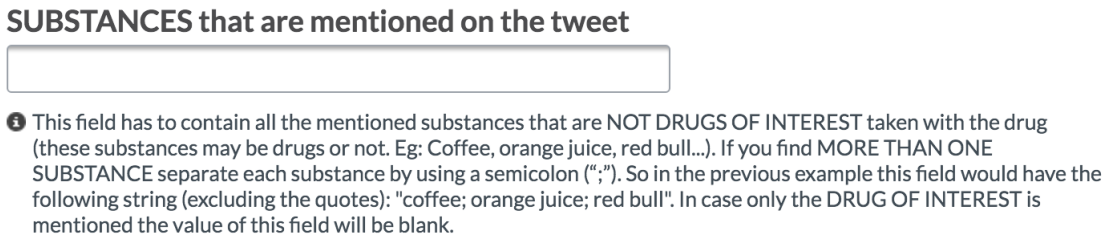

FIGURE B.7: Part 7 in the questionnaire to the laymen.

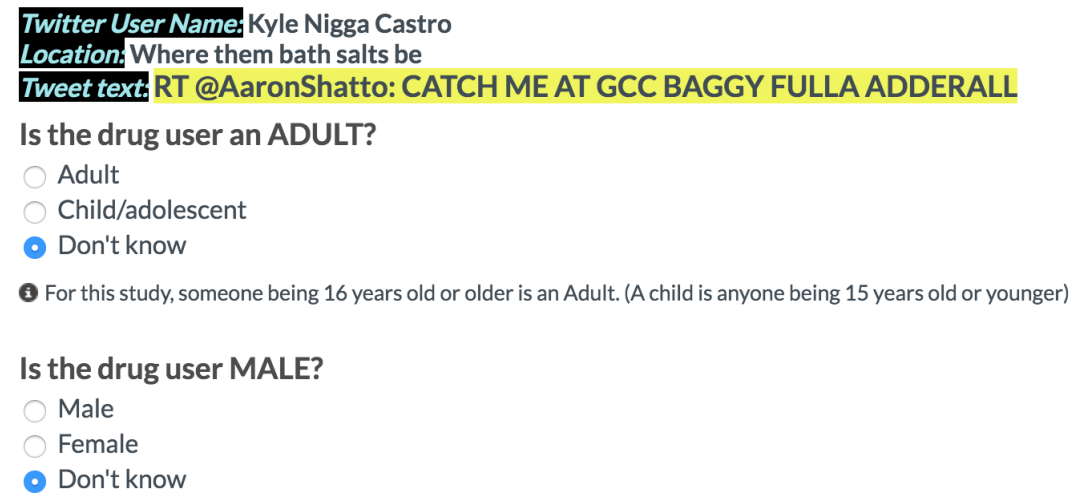

FIGURE B.8: Part 8 in the questionnaire to the laymen.

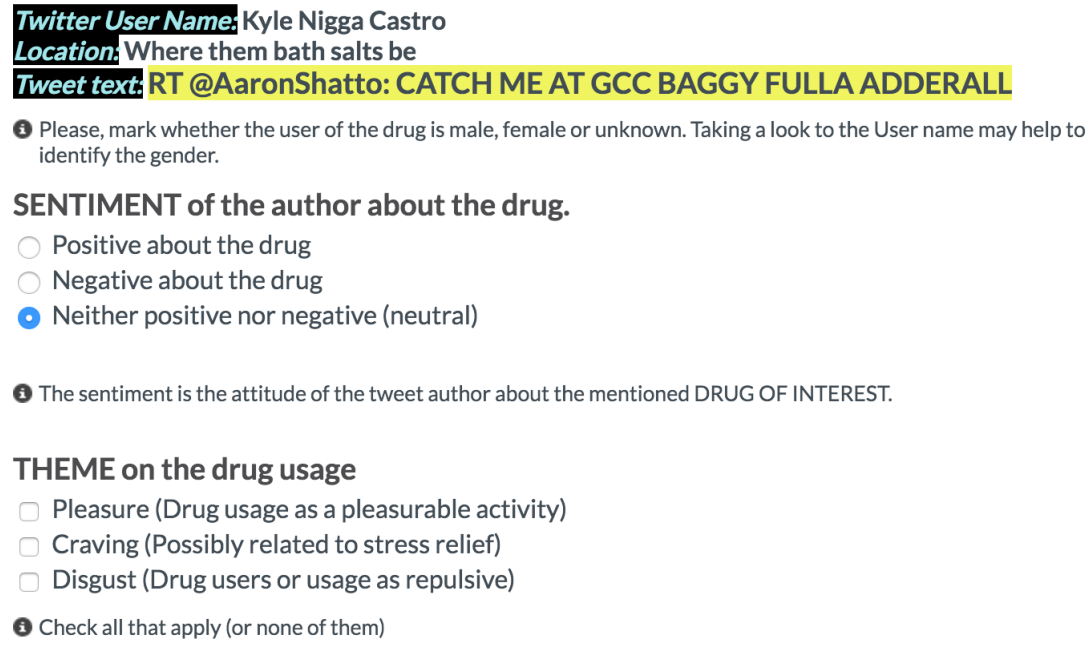

FIGURE B.9: Part 9 in the questionnaire to the laymen.

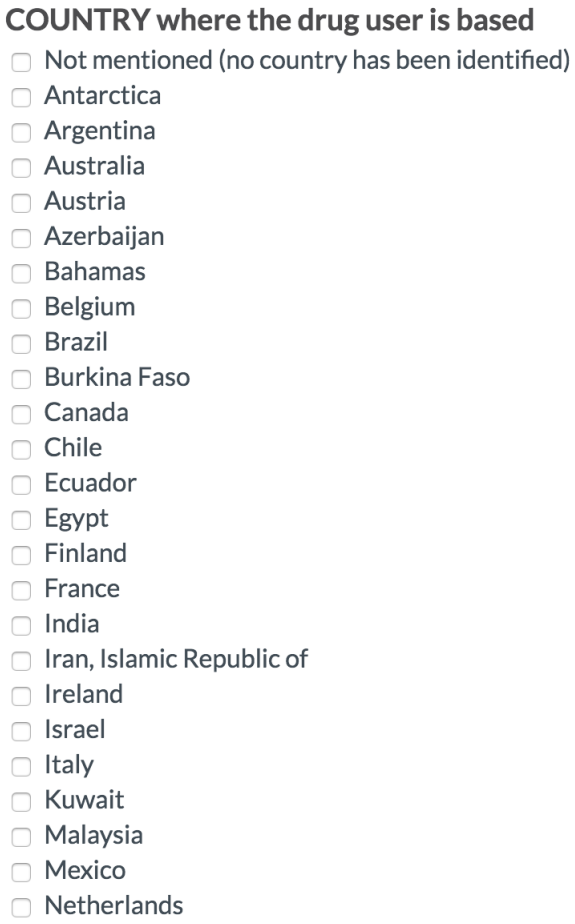

FIGURE B.10: Part 10 in the questionnaire to the laymen.

FIGURE B.11: Part 11 in the questionnaire to the laymen.

Appendix C

Guidelines to classify sentences of interest in Twitter and PubMed texts

C.1 Document information

DRAFT VERSION 0.6 (9th-Dec-2015)

Author: Nestor Alvaro

C.2 Introduction

The goal of this project is to annotate a set of sentences to indicate which are the sentences including drug mentions, and symptoms and diseases related to the drug effects in humans.

During the second phase of the annotation process the annotators will annotate the entities and the relations between entities present in the sentences identified during this first phase.

The final goal of this two-phases project is to create an annotated corpus of symptoms, diseases, and drugs mentions in sentences taken from PubMed articles and from tweets.

C.3 Structure of the document

The document has 2 columns ("A" and "B"):

- **A**: Contains the "Sentence" that is to be classified.
- **B**: Column where the annotator's classification will be entered.

C.4 Annotation guidelines

- The cells in column "A" ("Sentence") should not be modified.
- All the annotations will be entered in column "B" ("Target").
- The annotator will enter a "one" (the number 1, without the quotes) in the cells in column "Target" if the sentence mentions a drug and any symptom related to the drug intake.
  For example: "I take an aspirin when I have a headache" 1
- In case the sentence does not mention any drug, or there is no mention to any symptom the cell "Target" will be left blank.
- In case there is any doubt about which are the drugs that are of interest for this study, please check Table C.1 where we have listed the names of the drugs of interest.
- All the sentences in the spreadsheet file contain the name of some drug although in some cases the name of the drug may match the name of some other entity and the sentence may not be about the drug. If the sentence is not including the name of the drug, the annotation for that sentence will be left blank.
- The symptoms can appear before (usually causing the intake of the drug) or after the intake of the drug.
- The symptoms that appear after the intake of the drug can be symptoms causing a improvement on patient's condition or causing a worse condition (such as side effects of any kind, e.g. allergic reactions).
- In case the annotator is **not sure** if the concept is a symptom the annotator has to check it and make sure that it is listed in this website http://purl.bioontology.org/ontology/MEDDRA (Search for it by using the "Jump to" search box). In case the concept appears in the list that means that such concept is a valid symptom.

- Some concepts can appear in an "explanatory manner" in the sentences that have to be annotated. For example, the sentence "I could not close my eyes during the whole night" would be a reference to the concept "insomnia".

- It is very important to **reduce the guesses to a minimum**. Also, the annotator should reduce the number of assumptions to a minimum.

- The symptoms should not be induced, and all symptoms should be "identifiable" in the text of the sentence.
  This is, taking for example the sentence "I took an antidepressant" the word "antidepressant" **should not** be considered to be a symptom because it is referring to the drug, not to a symptom.
  In the same way, the word "antidepressant" does not unequivocally indicate that the patient is suffering from depression.
  On the other hand, the sentence "I took an antidepressant because I was depressed" does include a symptom that can be referenced (the words "be depressed").
  Back to the first example, the sentence "I took an antidepressant and I am cured now" does contain a symptom ("to be cured", which could be a reference to "not to be depressed").
  It is **very important** to understand the difference between these examples.

- This study targets the effects of the drugs in **humans**. If the sentence is not talking about the intake of the drug by humans, or in case the symptoms are not reported for humans the annotation for that sentence will be left blank. It is particularly important to be very careful with the examples where the sentences mention studies using animals such as "rats", "mouse", "mice" or "murine".

- In case the annotator is not clear on how to annotate the sentence, enter the letter "ex" ("x", without the quotes) in the corresponding "Target" cell for that sentence.

Drug Name
Adderall
Alergoliber
Alertec
Alkeran
Attenta
Avalox
Avastin
Avelon

Avelox

Bevacizumab

Bunavail

Buprenex

Buprenorphine

Butrans

Carbamazepine

celexa

cipramil

Ciprofloxacin

Citalopram

Cizdol

Concerta

Cortisone

Cymbalta

Daytrana

Depyrel

Destroamphetamine sulphate

Desyrel

Dextroamphetamine sulphate

Docetaxel

Duloxetine

Effexor

Elvanse

Equasym

Faverin

Fevarin

Floxyfral

Fluoxetine

Fluvoxamine maleate

Focalin

Genox

Geodon

Hynidate

Istubal

Kenalog

Lamictal

Lamotrigine
Lanzek
Levaquin
Levofloxacin
Lisdexamfetamine
Lisinopril
Lovetas
Lustral
Luvox
Medikinet
Melphalan
Mesyrel
Methylin
Methylphenidate hydrochloride
Modafinil
Modavigil
Molipaxin
Montecarloflo
Montecarolo-10
Montelukast
Moxeza
Moxifloxacin hydrochloride
Nolvadex
Norspan
Olanzapine
Oleptro
Pafinur
Paroxetine
paxil
Phenida
Pluralair
Prednisone
Prinivil
Prinzide
Provigil
prozac
Quetiapine

Ralif
Rinialer
Ritalin
Rupafin
Rupatadine
Rupax
Sarcolysin
Seroquel
Sertraline
Singulair
Suboxone
Subutex
Tamoxifen
Tavanic
Taxotere
Tegretol
Temgesic
Tensopril
Topamax
Topiramate
Trazodil
Trazodone
Trazorel
Trevilor
Trialodine
Triamcinolone acetonide
Trittico
Tyvense
Valodex
Venlafaxine
Venvanse
Vigamox
Volon A
Vyvanse
Zeldox
Zestril

Zestroretic
Ziprasidone
Zipwell
Zoloft
Zubsolv
Zypadhera
Zyprexa

TABLE C.1: Drug names used in the study.

Appendix D

Guidelines for Drug-Disease-Symptom annotation in Twitter and PubMed texts

D.1 Document information

DRAFT VERSION 0.9 (4th-Sept-2015)

Authors: Nestor Alvaro, Nut Limsopatham, and Nigel Collier

D.2 Introduction

The goal of this project is to create an annotated corpus of symptoms, diseases, and drugs mentions in sentences taken from PubMed articles and from tweets. In the rest of the document we will refer to the symptoms, diseases, and drugs using their capitalized form (SYMPTOMS, DISEASES, and DRUGS) when we talk about the generic entities that are to be annotated. For this study we focus on a closed set of DRUGS (see Table D.4), although the study is not limited to them, and any mention of any DRUG in our closed list of DRUGS should be annotated. This document provides a definition of SYMPTOM, DRUG, and DISEASE, the relations between these entities, and the guidelines to be followed during the annotation. In case there is some question not covered in this document, please send an e-mail to: nestoralvaro@nii.ac.jp . In these

Entity	Definition	Example
DRUG	Any of the marketed medicines that appears in the SIDER database (http://sideeffects.embl.de/), which is also listed in our closed set of drugs (See Table D.4).	The prescription included Lexapro.
SYMPTOMS	Any sign or SYMPTOM contained within the MedDRA ontology (http://bioportal.bioontology.org/ontologies/MEDDRA).	Adderall kept me focused.
DISEASE	Any DISEASE contained within the MedDRA ontology. (http://bioportal.bioontology.org/ontologies/MEDDRA).	The patient suffered from sleep deprivation without trazodone.

TABLE D.1: Entities to be annotated.

guidelines we describe each one of the 3 types of entities (DRUGS, SYMPTOMS, and DISEASES), the attributes of the entities, and the 3 types of relations (reason to use, outcome-positive, outcome-negative) that are to be annotated.

D.3 Annotation of entities

An entity can be a single word such as "tiredness" as it appears in the sentence "the patient was experiencing tiredness", or a span of text such as "could not move from the couch" obtained from the sentence "I worked out so hard that when I got back home I could not move from the couch". Both entities refer to the concept "tiredness symptom" (MEDDRA Code 10043890). We provide a list of entities, the definition of each one and an example in Table D.1. The only DRUGS to be annotated are those appearing in Table D.4. In the following examples DRUGS are highlighted in green, SYMPTOMS in blue, and DISEASES in red.

D.4 Annotation of attributes

The entities (DRUGS, SYMPTOMS and DISEASES) have some attributes that will be annotated to clarify some concepts. We provide a list of attributes for the entities, the definition of each one, the values each attribute can take, and an example in Table D.2. Some attributes have default values (**in bold and highlighted in the table**) which will be used when no attribute is chosen. In the following examples DRUGS are highlighted in green, SYMPTOMS in blue, and DISEASES in red.

Attribute	Definition	Values	Examples
Polarity	Indicates whether the entity is negated or not. The negation has to be a linguistic negation ("not", "don't"...).	• **Positive**: The entity is not negated. Default value. • **Negative**: The mention of the entity is negated.	"I took prozac and now I don't have a headache" prozac: polarity=positive (left blank) headache: polarity=negative
Person	Indicates whether the entity is affecting the "1st", "2nd", "3rd" person, or whether there is no information. This attribute is based on the original sender.	• **1st**: The entity is described from a "first person" point of view. The entity is directly impacting the author of the text. Relates a first hand experience. • **2nd**: The entity is described from a "second person" point of view. The entity is impacting another person whom the author knows. • **3rd**: The entity is described from a "third person" point of view. The entity is impacting someone not directly related with the author of the text. • **Not available**: There is no clear reference to whom the entity is impacting. Default value.	"I took prozac and now I don't have a headache" prozac: Person=1st The entity is described in first person. headache: Person=1st "Hate prozac" prozac: Person=not available (value left blank).

Attribute	Definition	Values	Examples
Modality	Indicates whether the entity is stated in an "actual", "hedged", "hypothetical" or "generic" way.	• **Actual**: These mentions have already happened or are being scheduled (without hedging) to happen. Default value • **Hedged**: These mentions include lexical ("seems", "likely", "suspicious", "possible", "consistent with"), or phrasal ("I suspect that...", "It would seem likely that") hedging. These entities are strongly implied, but, for safety, liability, or due to lack of comprehensive evidence, are not stated as a fact. • **Hypothetical**: Will often follow "if" statements ("If X happens, then we'll use Y to treat Z") or other sorts of conditionals ("Depending on the patient's response, we might treat A with B or with C"). • **Generic**: When the mention is done in a general sense. These usually occur when putting justifications of decisions, or rationales for changing care.	"The patient did not report nausea". nausea: Modality=Actual "The patient may have undergone a mild Stroke" Stroke: Modality=hedged "We suspect either achalasia or pseudoachalasia here" achalasia: Modality= Hypothetical pseudoachalasia: Modality= Hypothetical "Adderall should not be taken with other medications." Adderall: Modality= Generic

Attribute	Definition	Values	Examples
Exemplification	Indicates whether the entity is presented using an example or a description. Only to be used when the entity is presented through an exemplification.	• **Positive**: When an exemplification is used to present the entity. • **Negative**: The entity is not presented through an example. Default value.	"I will not be able to get up unless I take my Adderall" I will not be able to get up: Exemplification=True Indicates "lack of energy" (SNOMED ID: 248274002) Adderall: Exemplification=Negative (value left blank).
Duration	Indicates whether the entity's lasting span is "Intermittent", "Regular", "Irregular". If the duration is not indicated the attribute is left empty. In the case of DRUGS this attribute refers to the time span when the DRUG has been taken.	• **Regular**: The entity has a continued lasting span. • **Intermittent**: The lasting span of the entity has been recurring. • **Irregular**: There is indicated that there is no pattern in the lasting span of the entity. • **Not available**: When the duration is not indicated. Default value.	"I had a strong headache last night, so I took prozac." prozac: Duration=not available (the value will be left empty) headache: Duration="Irregular" "I have been on prozac for 5 years now" prozac: Duration="Regular"

Attribute	Definition	Values	Examples
Severity	Indicates whether the seriousness of an entity is "Mild", or "Severe". If the severity is not indicated the attribute is left empty. This attribute does not apply to DRUGS.	• **Mild**: There is gentle (not acute, nor serious) severity of the entity. • **Severe**: There is a grave or critical seriousness of the entity. • **Not available**: When the severity of the entity is not indicated. Default value.	"I had a strong headache last night, so I took prozac." prozac: Severity=not available (the value will be left empty) headache: Severity="Severe"
Status	Indicates whether the duration of the entity is "Complete", or "Continuing". If the duration is not indicated the attribute is left empty. In the case of DRUGS this attribute refers to the time span when the DRUG is perceived as having effect.	• **Complete**: If the entity is already not showing evidence of its effects. • **Continuing**: If the entity is still showing evidence of its effects. • **Not available**: When the status is not indicated. Default value.	"I had a strong headache last night, so I took prozac." prozac: Status=not available (the value will be left empty) headache: Status="Completed" "I took prozac 2 hours ago, but it already wore off." prozac: Status ="Complete"

Attribute	Definition	Values	Examples
Sentiment	Indicates whether the entity is perceived as "positive", "negative" or "neutral". If the entity is perceived as "neutral" this attribute is left empty.	• **Positive**: The entity is referenced as something good. • **Negative**: The entity is referenced as something bad. • **Neutral**: There is no clear point of view towards the referenced entity. Default value	"I had a strong headache last night, so I took prozac." prozac: Sentiment=neutral (the value will be left empty) headache: Sentiment="Negative"
Entity identifier	Indicates the identifier for that entity.	• **XXXXXX**: The concept identifier. The database contains a set of concepts obtained as follows – For SYMPTOMS and DISEASES the concept identifiers represent the UMLS concept ID for the MedDRA term. – For DRUGS the concept identifiers represent the PubChem concept ID referenced in SIDER database for that concept. • **-1**: If there is no concept identifier for an entity this value will be "-1". This value can not be used for drugs (if the drug is not in the list it should NOT be annotated)	"I had a strong headache last night, so I took prozac." prozac: ID="3386" headache: ID="10019211"

TABLE D.2: Attributes of the entities

Relation	Definition	Example
Reason to use	Represents the relation appearing when a SYMPTOM or DISEASE leads to the use of some DRUG.	Prozac is indicated for patients with major depressive disorder.
Outcome-positive	Represents the relation between a DRUG, and an expected or unexpected SYMPTOM or DISEASE appearing after the DRUG consumption. The outcome has to be positive.	I wish I was prescribed adderall, I'd lose weight.
Outcome-negative	Represents the relation between a DRUG, and an expected or unexpected SYMPTOM or DISEASE appearing after the DRUG consumption. The outcome has to be negative.	The most common adverse events reported for fluoxetine were impulsivity and poor concentration.

TABLE D.3: Entities to be annotated.

D.5 Annotation of relations

A relation represents the existing connection between two entities. In our annotations we allow 4 types of relations. DISEASES and SYMPTOMS are not related. We provide a list of relations, the definition of each one and an example in Table D.3. In the following examples DRUGS are highlighted in green, SYMPTOMS in blue, and DISEASES in red.

It is important to notice that the annotation tool validates the origin-entity and the end-entity of each relation. This means that:

- "Reason to use" relation: Has to start on a "SYMPTOM" or a "DISEASE" and be directed towards a "DRUG".
- "Outcome-positive" relation: Has to start on a "DRUG", and be directed towards a "SYMPTOM" or "DISEASE".
- "Outcome-negative" relation: Has to start on a "DRUG", and be directed towards a "SYMPTOM" or "DISEASE".

D.6 Practical issues

In the following examples DRUGS are highlighted in green, SYMPTOMS in blue, and DISEASES in red.

D.6.1 What to annotate?

Entities

- Each mention of an entity should be annotated exactly once. Each annotation should refer to exactly one mention of the entity. All the entities should be annotated each time they are mentioned.
- Annotate mentions with morphological variations such as adjectives.
  - For instance, "hypertensive" is annotated as "hypertension.".
  - Hashtags, whenever present, will be included in the annotation span.
    - In the sentence "I had a terrible #headache" the concept to be annotated is #headache (including the hashtag)
- Synonyms or descriptions for SYMPTOMS and DISEASES should be annotated.
  - Example: "I Took Adderall and now I'm gonna be up for hours"
    - "up for hours" should be annotated as a synonym of "Sleeplessness" (notation "10041017" in MEDDRA)
- The annotations should only include the entity mention, keeping it as specific as possible, and annotate the most specific entity mentions and select the best-matching Concept ID from SIDER database (for DRUGS) or MedDRA ontology (for SYMPTOMS and DISEASES) .
  - For instance, the complete phrase "partial seizures" (ID: 10061334) should be preferred over "seizures" (ID: 10039910) as it is more specific.
  - If present, the mention span should include terms such as disease, syndrome, disorder, infection.
- Mentions of cancer, tumour, neoplasm, or infection, and other generic mentions to DISEASES/SYMPTOMS additional information, can be annotated, although it may happen that the identifier for that concept is not contained in the list of concepts.
  - In this case the ID for the concept would be "-1"
- An entity could be an acronym.
  - A long form, short form pair should be annotated as two mentions. Example: "Attention deficit hyperactivity disorder (ADHD)". In this case "Attention deficit hyperactivity disorder" and "ADHD" should be annotated separately.

- This study is focused in a closed set of DRUGS (Table D.4).
  - That list of DRUGS also includes the brand names for these DRUGS.
    - Any mention of any of this DRUGS (including the brand names) has to be always annotated.
  - Those drugs have different brand names and trade names. These variants have to be annotated too.
    - For example, the table contains "Adderall", but "Adderall XR" and it should be annotated (using the DRUG identifier for "Adderall", 3007)
- Lists and co-ordinations are phrases which mention multiple entities in a complex way. A simple illustrative example is "breast and ovarian cancer", which refers to the entities "breast cancer" and "ovarian cancer".
  - These constructs often overlap or do not explicitly mention some terms.
  - As the tool allows discontinuous annotations each entity should be annotated one time. One annotation would be "breast cancer" and the second annotation would be "ovarian cancer".
- A retweet is a re-posting of someone else's Tweet. In this case the tweet will be considered as if the user re-posting it would be author of the tweet. Retweets are indicated by the string "RT" at the beginning of the message.
  - Example: "RT I took prozac and now I don't have a headache"
    - This example is a retweet of "I took prozac and now I don't!!! have a headache", so it would be annotated as if it were "I took prozac and now I don't have a headache"
      - Prozac: Person=1st
      - The entity is described in first person.
      - Headache: Person=1st
      - The entity is described in first person.
- There are some cases when DRUGS/SYMPTOMS/DISEASES are used as an indicator of other entity. In those cases the entity used for the reference should be annotated.
  - Example: "The patient took ADHD prescription stimulants"
    - ADHD should be annotated as a SYMPTOM
    - "ADHD prescription stimulants" should not be annotated as there is no drug in the list that could be found by looking for that concept.

  - Example: "The patient received fatigue treatment"
    - "fatigue" should be annotated as a symptom.
    - "fatigue treatment" should not be annotated as there is no drug in the list that could be found by looking for that concept.

Attributes

When an entity cannot be found in the list of concepts, "-1" will be used as the corresponding Entity Identifier.

- All the annotations should have a value for the attribute Entity Identifier.
  - The "-1" value can not be use for DRUGS (All annotated DRUGS have to be in Table D.4).

Relations

It is allowed to annotate relations between entities even if the related entities are not in the same sentence.

- Example: "The patient took Adderall during the day. As a result the patient's concentration improved'.
  - The entities to be annotated are Adderall (DRUG), and concentration (SYMPTOM). There will be a relation "outcome-positive" between these two entities even if each entity belong to a different sentence.

D.6.2 What NOT to annotate?

Entities

Entities should not both start and end with parenthesis.

- In case this happens only the entity within the parenthesis will be annotated.

DRUGS that are not listed in the Table D.4 should not be annotated. In our annotations we don't allow co-reference nor anaphoric references.

- Example: "Geodon used to make me sleep...now with Adderall and Ritalin at night? Nope"
  - In that tweet "Nope" could be understood as "No sleep", but we don't annotate that concept because we don't annotate anaphoric mentions.
- Example: "Respondents used stimulants mostly for wakefulness and performance enhancement"
  - In this example "stimulants" is not listed among our drugs, so it should not be annotated.
- Example: "I took Geodon yesterday. It doesn't work anymore"
  - In this sentence "It" could be understood as "Geodon", but as we don't allow anaphora "It" will not be annotated.
  - EXCEPTION: When an entity that has to be annotated contains an anaphoric mention to another entity to be annotated, the entity containing the anaphora should be annotated using that context information.
    - In the sentence "the patient experienced Severe imbecility, and that imbecility was intensified with the presence of [...]" the second occurrence of "imbecility" refers to "Severe imbecility", and should be annotated as such (Severe imbecility, with ID=10040442).

Attributes

If the DRUG is negated the relation will not be annotated.

- Example: "I did not take prozac and now I don't have a headache"
  - The relation between "prozac" and "headache" should NOT to be annotated.
- If it is just the SYMPTOM/DISEASE what is negated we annotate the relation.
  - Example: "I took prozac and now I don't have a headache"
    - The relation between "prozac" and "headache" has to be annotated.

The attributes of the entities should not be included in the annotation span unless required by the tokenisation, or in case the entity is a concept per se.

- Example: "nondiabetics" (annotate the entire word)
- Example: "no pain" (annotate only "pain")

- Example: "probable chronic fatigue syndrome" (only annotate "chronic fatigue syndrome").
- Example: "Severe dengue" (annotate the 2 words as "severe dengue" is a concept recognized by MEDDRA)

Determiners and quantifiers are never included in concept annotation unless that represents a different concept.

- Example: "I took prozac and adderall and now I'm very tired", the DISEASE is "tired", not "very tired".
  - In this case "very" will be encoded using the attribute "Severity", setting it to "severe"
- Example: "The patient has Severe imbecility"
  - In this case "Imbecility" is a concept (ID=10021409), but "Severe imbecility" is a concept too (ID=10040442), so we would annotate "Severe imbecility"
    - In this case too "Severity" attribute will be "severe".

D.7 Drugs of interest

Drug Name	**Brand name(s)**
Lisinopril	Zestril, Zestroretic, Prinzide, Prinivil, Tensopril
Prednisone	
Montelukast	Singulair, Pluralair, Montecarolo-10, Montecarloflo, Lovetas
Triamcinolone acetonide	Kenalog, Volon A
Topiramate	Topamax
Destroamphetamine sulphate	Adderall
Cortisone	Cortisone
Venlafaxine	Effexor, Trevilor
Buprenorphine	Suboxone, Cizdol, Subutex, Zubsolv, Bunavail, Temgesic, Buprenex, Norspan, Butrans
Sertraline	Zoloft, Lustral

Dextroamphetamine sulphate	Adderall
Methylphenidate hydrochloride	Ritalin, Concerta, Methylin, Medikinet, Equasym, Daytrana, Phenida, Attenta, Hynidate, Focalin
Modafinil	Modafinil, Alertec, Modavigil, Provigil
Citalopram	Citalopram, celexa, cipramil
Paroxetine	Paroxetine, paxil
Fluoxetine	Fluoxetine, prozac
Fluvoxamine maleate	Faverin, Fevarin, Floxyfral, Luvox
Carbamazepine	Tegretol
Olanzapine	Zyprexa, Zypadhera, Lanzek
Trazodone	Depyrel, Desyrel, Mesyrel, Molipaxin, Oleptro, Trazodil, Trazorel, Trialodine, Trittico
Ziprasidone	Geodon, Zeldox, Zipwell
Ciprofloxacin	Ciprofloxacin
Levofloxacin	Levaquin, Tavanic
Moxifloxacin hydrochloride	Avelox, Avalox, Avelon, Vigamox, Moxeza
Quetiapine	Seroquel
Bevacizumab	Avastin
Melphalan	Alkeran, Sarcolysin
Rupatadine	Rupafin, Alergoliber, Rinialer, Pafinur, Rupax, Ralif
Tamoxifen	Nolvadex, Istubal, Valodex, Genox
Docetaxel	Taxotere
Seroquel	Quetiapine
Lamotrigine	Lamictal
Duloxetine	Cymbalta
Lisdexamfetamine	Vyvanse, Venvanse, Elvanse, Tyvense

Table D.4: Drug names and brand names of the targeted DRUGS.

Bibliography

[1] WHO. The importance of pharmacovigilance. 2002.

[2] I Ralph Edwards and Marie Lindquist. Social media and networks in pharmacovigilance. *Drug Saf*, 34(4):267–71, 2011.

[3] Peder Larsen and Markus Von Ins. The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics*, 84 (3):575–603, 2010.

[4] Robert Leaman, Graciela Gonzalez, et al. Banner: an executable survey of advances in biomedical named entity recognition. In *Pacific Symposium on Biocomputing*, volume 13, pages 652–663. Citeseer, 2008.

[5] Marie Dupuch, Latitia Dupuch, Thierry Hamon, and Natalia Grabar. Inferring semantic relations between pharmacovigilance terms with the nlp methods. In *Computational Methods in Pharmacovigilance Workshop*, pages 27–31. Citeseer, 2012.

[6] Azadeh Nikfarjam, Abeed Sarker, Karen OConnor, Rachel Ginn, and Graciela Gonzalez. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, page ocu041, 2015.

[7] William P Isham. The role of message analysis in interpretation. In *Interpreting: The art of cross-cultural mediation. Proceedings of the RID Convention*, 1985.

[8] Douglas Biber. *Variation across speech and writing*. Cambridge University Press, 1991.

[9] Haiying Li, Zhiqiang Cai, and Arthur C Graesser. Comparing two measures for formality. In *The Twenty-Sixth International FLAIRS Conference*, 2013.

[10] Alejandro Mosquera and Paloma Moreda. A qualitative analysis of informality levels in web 2.0 texts: The facebook case study. In *Proceedings of the LREC workshop:@ NLP can u tag# user generated content*, pages 23–29, 2012.

[11] Costanza Asnaghi, Dirk Speelman, and Dirk Geeraerts. Geographical patterns of formality variation in written standard california english. *Digital Scholarship in the Humanities*, page fqu060, 2014.

[12] Suman Kalyan Maity, Bhadreswar Ghuku, Abhishek Upmanyu, and Animesh Mukherjee. Out of vocabulary words decrease, running texts prevail and hashtags coalesce: Twitter as an evolving sociolinguistic system. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pages 1681–1690. IEEE, 2016.

[13] Douglas Biber and Susan Conrad. *Register, genre, and style*. Cambridge University Press, 2009.

[14] Peter Trudgill. *Sociolinguistics: An introduction to language and society*. Penguin UK, 2000.

[15] Peter Trudgill. *Introducing language and society*. Penguin English, 1992.

[16] Suzanne Romaine. *Language in society: An introduction to sociolinguistics*. OUP Oxford, 2000.

[17] David Crystal. *Dictionary of linguistics and phonetics*, volume 30. John Wiley & Sons, 2011.

[18] David Crystal. *The Cambridge encyclopedia of language*, volume 2. Cambridge Univ Press.

[19] Michael Alexander Kirkwood Halliday. *Language as social semiotic*. London Arnold, 1978.

[20] Martin Joos. The five clocks–a linguistic excursion into the five styles of english usage. 1967.

[21] Marco Baroni and Silvia Bernardini. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274, 2006.

[22] Sudha Verma, Sarah Vieweg, William J Corvey, Leysia Palen, James H Martin, Martha Palmer, Aaron Schram, and Kenneth Mark Anderson. Natural language processing to the rescue? extracting" situational awareness" tweets during mass emergency. In *ICWSM*. Citeseer, 2011.

[23] Douglas Biber. Corpus linguistics and the study of english grammar. *Indonesian JELT*, 1(1):1–21, 2005.

[24] Lou Burnard. *British National Corpus: Users Reference Guide British National Corpus Version 1.0*. Oxford Univ. Computing Service, 1995.

[25] Hal Daumé III. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*, 2009.

[26] John Blitzer, Mark Dredze, Fernando Pereira, et al. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, volume 7, pages 440–447, 2007.

[27] Wei Wang, Krystl Haerian, Hojjat Salmasian, Rave Harpaz, Herbert Chase, and Carol Friedman. A drug-adverse event extraction algorithm to support pharmacovigilance knowledge mining from pubmed citations. In *AMIA Annu Symp Proc*, volume 2011, pages 1464–1470, 2011.

[28] Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of biomedical informatics*, 45(5):885–892, 2012.

[29] Paul Avillach, Jean-Charles Dufour, Gayo Diallo, Francesco Salvo, Michel Joubert, Frantz Thiessard, Fleur Mougin, Gianluca Trifirò, Annie Fourrier-Réglat, Antoine Pariente, et al. Design and validation of an automated method to detect known adverse drug reactions in medline: a contribution from the eu–adr project. *Journal of the American Medical Informatics Association*, 20(3):446–452, 2013.

[30] Jeana Frost, Sally Okun, Timothy Vaughan, James Heywood, and Paul Wicks. Patient-reported outcomes as a source of evidence in off-label prescribing: analysis of data from patientslikeme. *Journal of medical Internet research*, 13(1):e6, 2011.

[31] Jiang Bian, Umit Topaloglu, and Fan Yu. Towards large-scale twitter mining for drug-related adverse events. In *Proceedings of the 2012 international workshop on Smart health and wellbeing*, pages 25–32. ACM, 2012.

[32] Rachel Ginn, Pranoti Pimpalkhute, Azadeh Nikfarjam, Apurv Patki, Karen OConnor, Abeed Sarker, Karen Smith, and Graciela Gonzalez. Mining twitter for adverse drug reaction mentions: a corpus and classification benchmark. In *Proceedings of the fourth workshop on building and evaluating resources for health and biomedical text processing*. Citeseer, 2014.

[33] Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310, 2001.

[34] Henk Harkema, John N Dowling, Tyler Thornblade, and Wendy W Chapman. Context: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of biomedical informatics*, 42(5):839–851, 2009.

[35] Academy of Medical Sciences. Stratified, personalised or p4 medicine: a new direction for placing the patient at the centre of healthcare and health education. 2015. URL http://www.acmedsci.ac.uk/download.php?f=file&i=32644.

[36] K Haerian, D Varn, S Vaidya, L Ena, HS Chase, and C Friedman. Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. *Clinical Pharmacology & Therapeutics*, 92(2):228–234, 2012.

[37] Pernille Warrer, Ebba Holme Hansen, Lars Juhl-Jensen, and Lise Aagaard. Using text-mining techniques in electronic patient records to identify adrs from medicine use. *British journal of clinical pharmacology*, 73(5):674–684, 2012.

[38] Srinivasan V Iyer, Rave Harpaz, Paea LePendu, Anna Bauer-Mehren, and Nigam H Shah. Mining clinical text for signals of adverse drug-drug interactions. *Journal of the American Medical Informatics Association*, 21(2):353–362, 2014.

[39] Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. How noisy social media text, how diffrnt social media sources? In *IJCNLP*, pages 356–364, 2013.

[40] Penelope Brown and Stephen C Levinson. *Politeness: Some universals in language usage*, volume 4. Cambridge university press, 1987.

[41] Deborah Tannen. *Conversational style: Analyzing talk among friends*. Oxford University Press, 2005.

[42] Vijai K Bhatia. A generic view of academic discourse. *Academic discourse*, pages 21–39, 2002.

[43] Betty Samraj. Introductions in research articles: Variations across disciplines. *English for specific purposes*, 21(1):1–17, 2002.

[44] Betty Samraj. Disciplinary variation in abstracts: The case of wildlife behaviour and conservation biology. *Academic discourse*, pages 40–56, 2002.

[45] David Bunton. Generic moves in phd thesis introductions. *Academic discourse*, pages 57–75, 2002.

[46] Alison Love. Introductory concepts and cutting edgetheories: Can the genre of the textbook accommodate both. *Academic discourse*, pages 76–92, 2002.

[47] John Swales. *Genre analysis: English in academic and research settings*. Cambridge University Press, 1990.

[48] John Swales. *Research genres: Explorations and applications*. Ernst Klett Sprachen, 2004.

[49] Jean Ure. Introduction: approaches to the study of register range. *International Journal of the Sociology of Language*, 1982(35):5–24, 1982.

[50] Charles A Ferguson. Sports announcer talk: Syntactic aspects of register variation. *Language in society*, 12(02):153–172, 1983.

[51] Dell Hymes. Sociolinguistics: stability and consolidation. *International Journal of the Sociology of Language*, 1984(45):39–46, 1984.

[52] SB Heath and J Langman. Shared thinking and the register of coaching'in d. biber and e. finegan (eds.): Sociolinguistic perspectives on register, 1994.

[53] Paul Bruthiaux. Me tarzan, you jane: linguistic simplification in" personal ads" register. *Sociolinguistic perspectives on register*, pages 136–154, 1994.

[54] Paul Bruthiaux. *The discourse of classified advertising: Exploring the nature of linguistic simplicity*. Oxford University Press on Demand, 1996.

[55] Douglas Biber. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press, 1995.

[56] Susan Conrad. Variation among disciplinary texts: A comparison of textbooks and journal articles in biology and history. *Variation in English: Multi-dimensional studies*, pages 94–107, 2001.

[57] Douglas Biber. A register perspective on grammar and discourse: variability in the form and use of english complement clauses. *Discourse studies*, 1(2):131–150, 1999.

[58] Francesca Bargiela and Sandra J Harris. *Managing language: The discourse of corporate meetings*, volume 44. John Benjamins Publishing, 1997.

[59] CourtneyB Cazden. Classroomdiscourse: Thelanguageofteachingandlearning, 1988.

[60] Frances Christie. *Classroom discourse analysis: A functional perspective*. Bloomsbury Publishing, 2005.

[61] Douglas Biber, Ulla Connor, and Thomas A Upton. *Discourse on the move: Using corpus analysis to describe discourse structure*, volume 28. John Benjamins Publishing, 2007.

[62] Susan M Conrad. Investigating academic texts with corpus-based techniques: An example from biology. *Linguistics and education*, 8(3):299–326, 1996.

[63] John Flowerdew. *Academic discourse*. Routledge, 2014.

[64] Ken Hyland. *Hedging in scientific research articles*, volume 54. John Benjamins Publishing, 1998.

[65] Ken Hyland. Talking to students: Metadiscourse in introductorycoursebooks. *English for Specific Purposes*, 18(1):3–26, 1999.

[66] Ken Hyland and Polly Tse. Hooking the reader: A corpus study of evaluative that in abstracts. *English for specific purposes*, 24(2):123–139, 2005.

[67] Susan Peck MacDonald. The language of journalism in treatments of hormone replacement news. *Written Communication*, 22(3):275–297, 2005.

[68] Hilkka Stotesbury. Evaluation in research article abstracts in the narrative and hard sciences. *Journal of English for Academic Purposes*, 2(4):327–341, 2003.

[69] Thomas A Upton and Ulla Connor. Using computerized corpus analysis to investigate the textlinguistic discourse moves of a genre. *English for Specific Purposes*, 20(4):313–329, 2001.

[70] Minna Vihla. *Medical writing: Modality in focus*. Number 28. Rodopi, 1999.

[71] Françoise Salager-Meyer. Referential behavior in scientific writing: adiachronic study (1810–1995). *English for Specific Purposes*, 18(3):279–305, 1999.

[72] David Crystal. *Language and the Internet*. Cambridge University Press, 2001.

[73] Crispin Thurlow and Alex Brown. Generation txt? the sociolinguistics of young peoples text-messaging. *Discourse analysis online*, 1(1):30, 2003.

[74] Łukasz Grabowski. Register variation across english pharmaceutical texts: A corpus-driven study of keywords, lexical bundles and phrase frames in patient information leaflets and summaries of product characteristics. *Procedia-Social and Behavioral Sciences*, 95:391–401, 2013.

[75] Łukasz Grabowski. Keywords and lexical bundles within english pharmaceutical discourse: A corpus-driven description. *English for Specific Purposes*, 38:23–33, 2015.

[76] Tony Berber Sardinha. 25 years later. *Multi-Dimensional Analysis, 25 years on: A tribute to Douglas Biber*, 60:81, 2014.

[77] Eugene Charniak. A maximum-entropy-inspired parser. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 132–139. Association for Computational Linguistics, 2000.

[78] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 42–47. Association for Computational Linguistics, 2011.

[79] Helen Spencer-Oatey and Wenying Jiang. Explaining cross-cultural pragmatic findings: moving from politeness maxims to sociopragmatic interactional principles (sips). *Journal of Pragmatics*, 35(10):1633–1650, 2003.

[80] Instructor Rufaidah Kamal Abdul-Majeed. The realization of positive politeness strategies in language. 2009.

[81] Bo Han and Timothy Baldwin. Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 368–378. Association for Computational Linguistics, 2011.

[82] Bo Han, Paul Cook, and Timothy Baldwin. Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 421–432. Association for Computational Linguistics, 2012.

[83] Linda T Kohn, Janet M Corrigan, Molla S Donaldson, et al. *To err is human:: building a Safer Health System*, volume 6. National Academies Press, 2000.

[84] Jason Lazarou, Bruce H Pomeranz, and Paul N Corey. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *Jama*, 279(15):1200–1205, 1998.

[85] Sally Giles. Yellow card scheme reaches 50. *Prescriber*, 26(18):6–6, 2015.

[86] Lorna Hazell and Saad AW Shakir. Under-reporting of adverse drug reactions. *Drug Safety*, 29(5):385–396, 2006.

[87] Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen OConnor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. Utilizing social media data for pharmacovigilance: a review. *Journal of biomedical informatics*, 54:202–212, 2015.

[88] Adrian Benton, Lyle Ungar, Shawndra Hill, Sean Hennessy, Jun Mao, Annie Chung, Charles E Leonard, and John H Holmes. Identifying potential adverse effects using the web: A new approach to medical hypothesis generation. *Journal of biomedical informatics*, 44(6):989–996, 2011.

[89] Jelena Hadzi-Puric and Jeca Grmusa. Automatic drug adverse reaction discovery from parenting websites using disproportionality methods. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 792–797. IEEE Computer Society, 2012.

[90] Apurv Patki, Abeed Sarker, Pranoti Pimpalkhute, Azadeh Nikfarjam, Rachel Ginn, Karen OConnor, Karen Smith, and Graciela Gonzalez. Mining adverse drug reaction signals from social media: going beyond extraction. *Proceedings of BioLinkSig*, 2014:1–8, 2014.

[91] Abeed Sarker and Graciela Gonzalez. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of biomedical informatics*, 53:196–207, 2015.

[92] Brant W Chee, Richard Berlin, and Bruce Schatz. Predicting adverse drug events from personal health messages. In *AMIA Annu Symp Proc*, volume 2011, pages 217–26, 2011.

[93] Ming Yang, Xiaodi Wang, and Melody Y Kiang. Identification of consumer adverse drug reaction messages on social media. In *PACIS*, page 193, 2013.

[94] Clark C Freifeld, John S Brownstein, Christopher M Menone, Wenjie Bao, Ross Filice, Taha Kass-Hout, and Nabarun Dasgupta. Digital drug safety surveillance: Monitoring pharmaceutical products in Twitter. *Drug Safety*, 37(5):343–350, 2014.

[95] Robert Leaman, Laura Wojtulewicz, Ryan Sullivan, Annie Skariah, Jian Yang, and Graciela Gonzalez. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. In *Proceedings of the 2010 workshop on biomedical natural language processing*, pages 117–125. Association for Computational Linguistics, 2010.

[96] Azadeh Nikfarjam and Graciela H Gonzalez. Pattern mining for extraction of mentions of adverse drug reactions from user comments. In *AMIA Annual Symposium Proceedings*, volume 2011, page 1019. American Medical Informatics Association, 2011.

[97] Christopher C Yang, Haodong Yang, Ling Jiang, and Mi Zhang. Social media mining for drug safety signal detection. In *Proceedings of the 2012 international workshop on smart health and wellbeing*, pages 33–40. ACM, 2012.

[98] Christopher C Yang, Haodong Yang, and Ling Jiang. Postmarketing drug safety surveillance using publicly available health-consumer-contributed content in social media. *ACM Transactions on Management Information Systems (TMIS)*, 5(1):2, 2014.

[99] Hariprasad Sampathkumar, Xue-wen Chen, and Bo Luo. Mining adverse drug reactions from online healthcare forums using hidden markov model. *BMC medical informatics and decision making*, 14(1):91, 2014.

[100] SriJyothsna Yeleswarapu, Aditya Rao, Thomas Joseph, Vangala Govindakrishnan Saipradeep, and Rajgopal Srinivasan. A pipeline to extract drug-adverse event pairs from multiple data sources. *BMC medical informatics and decision making*, 14(1):1, 2014.

[101] Andrew Yates and Nazli Goharian. Adrtrace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites. In *Advances in Information Retrieval*, pages 816–819. Springer, 2013.

[102] Isabel Segura-Bedmar, Ricardo Revert, and Paloma Martínez. Detecting drugs and adverse events from spanish health social media streams. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)@ EACL*, pages 106–115, 2014.

[103] Vassilis Koutkias and Marie-Christine Jaulent. A multiagent system for integrated detection of pharmacovigilance signals. *Journal of medical systems*, 40(2):1–14, 2016.

[104] SAFAA ELTYEB and NAOMIE SALIM. Pattern-based system to detect the adverse drug effect sentences in medical case reports. *Journal of Theoretical and Applied Information Technology*, 71(1), 2015.

[105] Ning Kang, Bharat Singh, Chinh Bui, Zubair Afzal, Erik M van Mulligen, and Jan A Kors. Knowledge-based extraction of adverse drug events from biomedical text. *BMC bioinformatics*, 15(1):1, 2014.

[106] María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5):914–920, 2013.

[107] Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55:73–81, 2015.

[108] Erik M Van Mulligen, Annie Fourrier-Reglat, David Gurwitz, Mariam Molokhia, Ainhoa Nieto, Gianluca Trifiro, Jan A Kors, and Laura I Furlong. The eu-adr corpus: annotated drugs, diseases, targets, and their relationships. *Journal of biomedical informatics*, 45(5):879–884, 2012.

[109] Dimah Sweis and Ian CK Wong. A survey on factors that could affect adverse drug reaction reporting according to hospital pharmacists in great britain. *Drug Safety*, 23(2):165–172, 2000.

[110] Hale Zerrin Toklu and Meral Keyer Uysal. The knowledge and attitude of the turkish community pharmacists toward pharmacovigilance in the kadikoy district of istanbul. *Pharmacy world & science*, 30(5):556–562, 2008.

[111] Susan J Semple, Elizabeth Hotham, Deepa Rao, Karen Martin, Caroline A Smith, and Geraldine F Bloustien. Community pharmacists in australia: barriers to information provision on complementary and alternative medicines. *Pharmacy World and Science*, 28(6):366–373, 2006.

[112] M Tamuz, EJ Thomas, and KE Franchois. Defining and classifying medical error: lessons for patient safety reporting systems. *Quality and Safety in Health Care*, 13 (1):13–20, 2004.

[113] Daniel R. Longo, John E. Hewett, Bin Ge, and Shari Schubert. The long road to patient safety. *jama*, 2005.

[114] Lucian L Leape and Donald M Berwick. Five years after to err is human: what have we learned? *Jama*, 293(19):2384–2390, 2005.

[115] Chung-Chih Lin, Chung-Liang Shih, Hsun-Hsiang Liao, and Cathy HY Wung. Learning from taiwan patient-safety reporting system. *International journal of medical informatics*, 81(12):834–841, 2012.

[116] Paea LePendu, Srinivasan V Iyer, Anna Bauer-Mehren, Rave Harpaz, Jonathan M Mortensen, Tanya Podchiyska, Todd A Ferris, and Nigam H Shah. Pharmacovigilance using clinical notes. *Clinical pharmacology & therapeutics*, 93(6):547–555, 2013.

[117] Özlem Uzuner, Imre Solti, and Eithon Cadag. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17 (5):514–518, 2010.

[118] Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. Dnorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917, 2013.

[119] Carl Lee Hanson, Ben Cannon, Scott Burton, and Christophe Giraud-Carrier. An exploration of social circles and prescription drug abuse through Twitter. *Journal of medical Internet research*, 15(9), 2013.

[120] FDA. Guidance for industry internet/social media platforms with character space limitations presenting risk and benefit information for prescription drugs and medical devices, June 2014. URL http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm401087.pdf.

[121] Medicines and healthcare products regulatory agency (mhra). press release: Uk regulator leads innovative eu project on the use of smartphones and social media for drug safety information, November 2015. URL http://www.imi.europa.eu/content/web-radr.

[122] European Medicines Agency. European medicines agency. guideline on good pharmacovigilance practices (gvp), 2013. URL http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/document_listing/document_listing_000345.jsp.

[123] European Medicines Agency. Guideline on good pharmacovigilance practices (gvp) module vi management and reporting of adverse reactions to medicinal products (rev 1), 2014. URL http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/09/WC500172402.pdf.

[124] Guodong Zhou, Jie Zhang, Jian Su, Dan Shen, and Chewlim Tan. Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, 20(7): 1178–1190, 2004.

[125] Jari Björne, Juho Heimonen, Filip Ginter, Antti Airola, Tapio Pahikkala, and Tapio Salakoski. Extracting complex biological events with rich graph-based feature sets. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pages 10–18. Association for Computational Linguistics, 2009.

[126] Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Michael Hess, Christos Andronis, Ourania Konstandi, and Andreas Persidis. Mining of relations between proteins over biomedical scientific literature using a deep-linguistic approach. *Artificial intelligence in medicine*, 39(2):127–136, 2007.

[127] Buzhou Tang, Hongxin Cao, Xiaolong Wang, Qingcai Chen, and Hua Xu. Evaluating word representation features in biomedical named entity recognition tasks. *BioMed research international*, 2014, 2014.

[128] M Sebastian A Wolfram. Modelling the stock market using twitter. *School of Informatics*, page 74, 2010.

[129] Wen Hao Chen, Yi Cai, and Kin Keung Lai. Weibo mood towards stock market. In *Database Systems for Advanced Applications*, pages 3–14. Springer, 2016.

[130] Akira Yoshihara, Kazuhiro Seki, and Kuniaki Uehara. Leveraging temporal properties of news events for stock market prediction. *Artificial Intelligence Research*, 5(1):p103, 2015.

[131] Rachel Lynn Kendra, Suman Karki, Jesse Lee Eickholt, and Lisa Gandy. Characterizing the discussion of antibiotics in the twittersphere: what is the bigger picture? *Journal of medical Internet research*, 17(6), 2015.

[132] A Mishra, A Malviya, and S Aggarwal. Towards automatic pharmacovigilance: Analysing patient reviews and sentiment on oncological drugs. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 1402–1409. IEEE, 2015.

[133] Yutaka Sasaki, Paul Thompson, John McNaught, and Sophia Ananiadou. Three bionlp tools powered by a biological lexicon. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*, pages 61–64. Association for Computational Linguistics, 2009.

[134] Kirill Degtyarenko, Paula De Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. Chebi: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, 36(suppl 1):D344–D350, 2008.

[135] Nigel Collier, Tudor Groza, Damian Smedley, Peter N Robinson, Anika Oellrich, and Dietrich Rebholz-Schuhmann. Phenominer: from text to a database of phenotypes associated with omim diseases. *Database*, 2015:bav104, 2015.

[136] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[137] Isabel Segura-Bedmar, Vıctor Suárez-Paniagua, and Paloma Martınez. Exploring word embedding for drug name recognition. In *SIXTH INTERNATIONAL WORKSHOP ON HEALTH TEXT MINING AND INFORMATION ANALYSIS (LOUHI)*, page 64, 2015.

[138] Karen OConnor, Pranoti Pimpalkhute, Azadeh Nikfarjam, Rachel Ginn, Karen L Smith, and Graciela Gonzalez. Pharmacovigilance on twitter? mining tweets for adverse drug reactions. In *AMIA Annual Symposium Proceedings*, volume 2014, page 924. American Medical Informatics Association, 2014.

[139] Nestor Alvaro, Mike Conway, Son Doan, Christoph Lofi, John Overington, and Nigel Collier. Crowdsourcing twitter annotations to identify first-hand experiences of prescription drug use. *Journal of biomedical informatics*, 58:280–287, 2015.

[140] United States Securities and Exchange Commission. Form s-1. registration statement. Twitter, inc., October 2013. URL http://www.sec.gov/Archives/edgar/data/1418091/000119312513390321/d564001ds1.htm.

[141] Francisca González-Rubio, Amaia Calderón-Larrañaga, Beatriz Poblador-Plou, Cristina Navarro-Pemán, Anselmo López-Cabañas, and Alexandra Prados-Torres. Underreporting of recognized adverse drug reactions by primary care physicians: an exploratory study. *Pharmacoepidemiology and drug safety*, 20(12):1287–1294, 2011.

[142] Revealed: The demographic trends for every social network, October 2014. URL http://www.businessinsider.com/2014-social-media-demographics-update-2014-9.

[143] U.S. Food and Drug Administration. Fda adverse event reporting system (FAERS), 2015. URL http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/.

[144] Sebastian Schneeweiss, Amanda R Patrick, Daniel H Solomon, Colin R Dormuth, Matt Miller, Jyotsna Mehta, Jennifer C Lee, and Philip S Wang. Comparative safety of antidepressant agents for children and adolescents regarding suicidal acts. *Pediatrics*, pages peds–2009, 2010.

[145] Carl L Hanson, Scott H Burton, Christophe Giraud-Carrier, Josh H West, Michael D Barnes, and Bret Hansen. Tweaking and tweeting: exploring Twitter for nonmedical use of a psychostimulant drug (adderall) among college students. *Journal of medical Internet research*, 15(4), 2013.

[146] C Ian Ragan, Imre Bard, and Ilina Singh. What should we do about student use of cognitive enhancers? an analysis of current evidence. *Neuropharmacology*, 64: 588–595, 2013.

[147] Nicole C White, Toby Litovitz, and Cathleen Clancy. Suicidal antidepressant overdoses: a comparative analysis by antidepressant type. *Journal of medical toxicology*, 4(4):238–250, 2008.

[148] Twitter Inc. Get statuses/sample, September 2014. URL https://dev.twitter.com/streaming/reference/get/statuses/sample.

[149] David Sell. Fda warns of counterfeit adderall, Oct 2012. URL http://articles.philly.com/2012-05-31/business/31900817_1_rogue-websites-and-distributors-generic-versions-adderall.

[150] Danny O'Neil. John moffitt on adderall: 'it was a total mistake', Nov 2012. URL http://seattletimes.com/html/seahawksblog/2019783660_adderall28.html.

[151] Aurobindo pharma gets usfda nod for modafinil tablets, Sep 2012. URL http://articles.economictimes.indiatimes.com/2012-09-28/news/34148290_1_aurolife-pharma-llc-usfda-nod-aurobindo-pharma.

[152] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, April 1960. ISSN 0013-1644, 1552-3888.

[153] Joseph L Fleiss and Jacob Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 1973.

[154] Matthias Gamer. Package 'irr', 2012. URL http://cran.r-project.org/web/packages/irr/irr.pdf.

[155] Mark Myslín, Shu-Hong Zhu, Wendy Chapman, and Mike Conway. Using Twitter to examine smoking behavior and perceptions of emerging tobacco products. *Journal of medical Internet research*, 15(8), 2013.

[156] Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.

[157] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, pages 81–93, 1938.

[158] CP Dancy and J Reidy. Statistics without maths for psychology. *IEEE Statistics without maths for psychology*, 2004.

[159] Stefanie Nowak and Stefan Rüger. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, pages 557–566. ACM, 2010.

[160] Lawrence D Brown, T Tony Cai, and Anirban DasGupta. Interval estimation for a binomial proportion. *Statistical science*, pages 101–117, 2001.

[161] How to calculate a confidence score, July 2014. URL http://success.crowdflower.com/customer/portal/articles/1295977-how-to-calculate-a-confidence-score.

[162] CrowdFlower. Get results — how to calculate a confidence score, 2015. URL https://success.crowdflower.com/hc/en-us/articles/201855939-Get-Results-How-to-Calculate-a-Confidence-Score.

[163] Cheryl A Kerfeld. Introduction: Sequences and consequences. *Biochemistry and Molecular Biology Education*, 41(1):12–15, 2013.

[164] Thomas Lippincott, Diarmuid Ó Séaghdha, and Anna Korhonen. Exploring subdomain variation in biomedical language. *BMC bioinformatics*, 12(1):212, 2011.

[165] J-D Kim, Tomoko Ohta, Yuka Tateisi, and Junichi Tsujii. Genia corpusa semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl 1):i180–i182, 2003.

[166] Lorraine Goeuriot, Liadh Kelly, Gareth JF Jones, Guido Zuccon, Hanna Suominen, Allan Hanbury, Henning Müller, and Johannes Leveling. Creation of a new evaluation benchmark for information retrieval targeting patient information needs. 2013.

[167] Raheel Nawaz, Paul Thompson, John McNaught, and Sophia Ananiadou. Meta-knowledge annotation of bio-events. In *LREC*, 2010.

[168] Michael Bada, Lawrence E Hunter, Miriam Eckert, and Martha Palmer. An overview of the craft concept annotation guidelines. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 207–211. Association for Computational Linguistics, 2010.

[169] Jin-Dong Kim, Tomoko Ohta, Y Tateisi, and Junichi Tsujii. Genia corpus manual-encoding schemes for the corpus and annotation. Technical report, Technical report TR-NLP-UT-2006-1, School of Information Science, University of Tokyo, 2006.

[170] Steven Belknap, Elaine Freund, Nadya Frid, Zuofeng Li, Rashmi Prasad, Balaji Ramesh, and Hong Yu. The annotation guideline manual: Extracting adverse drug event information from clinical narratives in electrical medical records, 2015. URL http://www.bio-nlp.org/papers/THE%20ANNOTATION%20GUIDELINE%20MANUAL%202015MAY18%20%20final.pdf.

[171] William F Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, et al. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154, 2014.

[172] Angus Roberts, Robert Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Andrea Setzer, and Ian Roberts. Semantic annotation of clinical text: The clef corpus. In *Proceedings of the LREC 2008 workshop on building and evaluating resources for biomedical text mining*, pages 19–26, 2008.

[173] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[174] Oliver Freudenreich. Differential diagnosis of psychotic symptoms: medical mimics. *Psychiatric Times*, 27(12):56–61, 2010.

[175] L Khaodhiar, KC McCowen, and GL Blackburn. Obesity and its comorbid conditions. 1999.

[176] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[177] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

[178] Robert L Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, 1953.

[179] Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. Automatic labelling of topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1536–1545. Association for Computational Linguistics, 2011.

[180] George Miller and Christiane Fellbaum. Wordnet: An electronic lexical database, 1998.

[181] Sharon A Caraballo. Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 120–126. Association for Computational Linguistics, 1999.

[182] Emmanuelle Dietz, Damir Vandic, and Flavius Frasincar. Taxolearn: A semantic approach to domain taxonomy learning. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on*, volume 1, pages 58–65. IEEE, 2012.

[183] Erik Smitterberg. The progressive and phrasal verbs evidence of colloquialization. *The Dynamics of Linguistic Variation: Corpus evidence on English past and present*, 2:269, 2008.

[184] Kyle B Dempsey, Philip M McCarthy, and Danielle S McNamara. Using phrasal verbs as an index to distinguish text genres. In *FLAIRS Conference*, pages 217–222, 2007.

[185] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–768. ACM, 2010.

[186] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python.* " O'Reilly Media, Inc.", 2009.

[187] Thomas Hoffmann. *Preposition placement in English: A usage-based approach.* Cambridge University Press, 2011.

[188] Tom Lippincott, Diarmuid O Séaghdha, Lin Sun, and Anna Korhonen. Exploring variations across biomedical subdomains. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 689–697. Association for Computational Linguistics, 2010.

[189] Yuheng Hu, Kartik Talamadupula, Subbarao Kambhampati, et al. Dude, srsly?: The surprisingly formal nature of twitter's language. In *ICWSM*, 2013.

[190] Maurizio Gotti. Specialized discourse. *Linguistic Features and Changing Conventions. Peter Lang: Bern*, 2003.

[191] Geoff Thompson and Ye Yiyun. Evaluation in the reporting verbs used in academic papers. *Applied linguistics*, 12(4):365–382, 1991.

[192] Douglas Biber. On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes*, 15(2):133–163, 1992.

[193] Mike Conway, Son Doan, Ai Kawazoe, and Nigel Collier. Classifying disease outbreak reports using n-grams and semantic features. *International journal of medical informatics*, 78(12):e47–e58, 2009.

[194] Yogesh Tewari and Rajesh Kawad. Real-time topic modeling of microblogs, March 2013. URL http://www.oracle.com/technetwork/articles/java/micro-1925135.html.

[195] Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. Improving lda topic models for microblogs via tweet pooling and automatic labeling, 2013.

[196] Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. Automatic labelling of topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1536–1545. Association for Computational Linguistics, 2011.

[197] Yiming Yang and Jan O Pedersen. A comparative study on feature selection in text categorization. In *ICML*, volume 97, pages 412–420, 1997.

[198] Liangjie Hong and Brian D Davison. Empirical study of topic modeling in Twitter. In *Proceedings of the First Workshop on Social Media Analytics*, pages 80–88. ACM, 2010.

[199] Piotr Romanski. Package "FSelector", February 2013. URL http://cran.r-project.org/web/packages/FSelector/FSelector.pdf.

[200] Max Kuhn. Package 'caret', August 2014. URL http://cran.r-project.org/web/packages/caret/caret.pdf.

[201] David Martin Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. 2011.

[202] Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. Sentiment of emojis. *PLoS ONE*, 10(12):e0144296, 2015. URL http://dx.doi.org/10.1371/journal.pone.0144296.

[203] Saif M Mohammad and Peter D Turney. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[204] Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*, 2013.

[205] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[206] Bing Liu, Minqing Hu, and Junsheng Cheng. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351. ACM, 2005.

[207] Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA, June 2013.

[208] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422. Citeseer, 2006.

[209] Sebastian Köhler, Sandra C Doelken, Christopher J Mungall, Sebastian Bauer, Helen V Firth, Isabelle Bailleul-Forestier, Graeme CM Black, Danielle L Brown, Michael Brudno, Jennifer Campbell, et al. The human phenotype ontology project: linking molecular biology and disease through phenotype data. *Nucleic acids research*, page gkt1026, 2013.

[210] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[211] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[212] Thorsten Joachims. *Text categorization with support vector machines: Learning with many relevant features*. Springer, 1998.

[213] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. A practical guide to support vector classification. 2003.

[214] Robert Leaman, Ritu Khare, and Zhiyong Lu. Challenges in clinical natural language processing for automated disorder normalization. *Journal of biomedical informatics*, 57:28–37, 2015.

[215] Naoaki Okazaki. Crfsuite: a fast implementation of conditional random fields (crfs), 2007. URL http://www.chokkan.org/software/crfsuite/.

[216] WENTING Wang. Mining adverse drug reaction mentions in twitter with word embeddings. In *Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing*, 2016.