Utilizing Text Mining on Online Medical Forums to Predict Label Change due to Adverse Drug Reactions

Ronen Feldman*
School of Business
Administration, Hebrew
University of Jerusalem
ronen.feldman@huji.ac.il

Oded Netzer
Graduate School of Business,
Columbia University
on2110@columbia.edu

Aviv Peretz*
School of Computer Science
and Engineering, Hebrew
University of Jerusalem
aviv.peretz@mail.huji.ac.il

Binyamin Rosenfeld
grurgrur@gmail.com

ABSTRACT

We present an end-to-end text mining methodology for relation extraction of adverse drug reactions (ADRs) from medical forums on the Web. Our methodology is novel in that it combines three major characteristics: (i) an underlying concept of using a head-driven phrase structure grammar (HPSG) based parser; (ii) domain-specific relation patterns, the acquisition of which is done primarily using unsupervised methods applied to a large, unlabeled text corpus; and (iii) automated post-processing algorithms for enhancing the set of extracted relations.

We empirically demonstrate the ability of our proposed approach to predict ADRs prior to their reporting by the Food and Drug Administration (FDA). Put differently, we put our approach to a predictive test by demonstrating that our methodology can credibly point to ADRs that were not uncovered in clinical trials for evaluating new drugs that come to market but were only reported later on by the FDA as a label change.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*

General Terms

Algorithms; Experimentation

Keywords

Adverse Drug Reactions; Pharmaceutical Drugs; Hpsg; Medical Forums; Text Mining

1. INTRODUCTION

The field of text mining deals with algorithms and techniques by which valuable information can be derived from textual data. Specifically, text mining has been found to be useful when the amount of text is extensive. One such corpus that is characterized by a large body of textual data is Internet forums. In this research we are particularly interested in mining *medical* online forums in an attempt to draw insights and make predictions with respect to pharmaceutical ADRs - medical damages caused by taking a given medication at a normal dose. Several papers have investigated whether mining social media can help detecting ADRs (e.g., [13] and [20]). However, these papers relate the set of reported ADRs to ADRs found in the forum. They do not attempt to predict unreported ADRs that are likely to be found at a later time. Clinical trials are often limited in terms of the number of participants and scope. Accordingly, these trials sometimes fail to indicate ADRs associated with a particular drug. ADRs that were not found in the clinical trials are often reported later on, sometimes even years after a drug has come to market as an FDA label change [20]. Because adverse drug events comprise the fourth leading cause of death in the US [3], the FDA has developed several monitoring channels for physicians, pharmacists, patients and drug companies to report ADR post-marketing of the drugs.

Now, suppose a new drug is introduced to the market. A natural task in terms of text mining would therefore be finding a way to efficiently and accurately utilize the wisdom of crowds, manifested in health related online forums, in order to identify potential ADRs that were not found in clinical trials. Netzer, Feldman, Goldenberg and Fresko [13] portray the curse underlying the abundant supply of user generated content. Practically, one needs to track and quantify an overwhelming amount of information. Furthermore, the data available in user-generated forums, while rich, is highly unstructured. Text in such forums often includes informal jargon, abbreviations, and lack of formal linguistic structure, making the task of parsing a sentence and identifying the role of each part of speech, burdensome.

The goal of this paper is to present a text mining methodology that will allow a rather unlaborious, yet reliable, detection of unreported ADRs that are likely to be identified in the future. Since we cannot prove our predictions on to-be-announced ADRs - not until the future becomes the present - we take a set of ADRs that have been reported by the FDA

in a label change post the marketing of the drug, and examine whether signals of these ADRs appear in user-generated health forum prior to the first announcement by the FDA about these ADRs. We find that our approach is able to detect such label changes sometime years prior to the FDA announcement.

2. RELATED WORK

There is a large body of literature on the information and insights that can be elicited from text mining consumer forums [8, 7, 5, 6, 11, 9]. Moreover, the specific task of mining ADRs from health-related social media, has also drawn the attention of researchers. Several papers have taken various approaches to tackle this problem. These papers vary with respect to the way Drug-ADR relations were perceived and, as a result, in the techniques that were used for extracting those Drug-ADR relations. We describe this literature next.

2.1 Co-occurrence Based Approach

Several papers have addressed the problem of relation extraction based on the notion of sheer co-occurrence analysis. These papers explicitly or implicitly assume that a Drug-ADR pair co-occurring within the same sentence, or within a window of n tokens, or even within an entire document, conceals a relation. Recently Segura-Bedmar, PePeña and Martìnez [17] presented a system aimed at detecting ADRs from Spanish medical forums based on co-occurrence of Drug-ADR pairs while considering different window sizes stretching from 10 up to 250 tokens. Similarly, Benton et al. [2] developed a system to extract Drug-ADR pairs from medical message boards using co-occurrence statistics, in which any two terms co-occurring within a window of 20 tokens apart were considered an association rule. In an early paper by Leaman at al. [10] co-occurrence analysis was implicit, as the mined information had been scraped from a website, in which users indicated a specific treatment when posting comments, thus treatments which were not drug-based, were filtered out. In order to mine the ADRs themselves, the terms mentioned in user comments in a bag-of-words sliding window were matched against terms in a lexicon. Yang, Yang, Jiang and Tang [21] (followed by [22]), proposed that co-occurrences of a drug and an ADR in online healthcare communities were to be regarded as an association rule, that is, a drug $\Rightarrow$ ADR association. An association rule mining algorithm was employed, in which ADRs were identified by comparing a sliding window of tokens against each term in an ADR lexicon.

The main limitation of using co-occurrence as a measure for Drug-ADR relation is that in many cases co-occurrence of two entities is loosely related to these two entities comprising a relation. Co-occurrence ignores the syntactic and semantic structures of the relation. Moreover, association rules that are derived from co-occurrence don't necessarily have causal direction, rather they simply imply a correlation between a drug and an ADR. The problem exacerbates as the allowed window size for co-occurrence increases. These problems lead to low precision in relation extraction.

2.2 Supervised Learning Based Approach

A few recent papers have tried to harness the power of supervised machine learning to extract Drug-ADR relations. Nikfarjam and Gonzalez [14] applied association rule to mine a set of annotated comments for generating underlying language patterns frequently used for ADR mentions in social media websites. Ensemble classification technique was used by Chee, Berlin and Shatz [3] to classify drugs into those that had undergone label changes or regulatory by the FDA, and those which had not. For their experiments they built hundreds of classifiers. Naïve Bayes and Support Vector Machine (SVM) with a RBF kernel were the two base classification algorithms used. They were able to successfully identify drugs that have been withdrawn from the market based on messages that had been posted before those withdrawals. Patki et al. [15] explored a supervised text classification approach for mining health-related forum. Applying a two-step approach, they first classified whether or not a user comment included a mention of an ADR, and then inferred whether the combined comments for the drug indicated an inordinate incidence of adverse reactions. Two supervised machine learning algorithms were utilized: Multinomial Naïve Bayes (MNB) and SVM. They demonstrate mixed results with respect to the ability of their method distinguishing true ADRs from noise.

While text classification approaches gained popularity in analyzing unstructured, informal texts commonly found in online forums, due to their ability to filter out noise from noisy data, it is worth emphasizing that the supervised nature of the these approaches is quite laborious. Numerous user comments had to be annotated to create the training data, often annotation was performed by domain experts or under the guidance of a pharmacology expert. This effort might not be feasible when researching a large number of drugs and ADRs. Moreover, results are highly dependent on the richness and the representativeness of the training data.

2.3 Semi-Supervised and Unsupervised Learning Based Approach

In light of the above-mentioned limitations of the supervised-based approach, recent papers have tried to evade its chains by suggesting unsupervised or semi-supervised learning methodologies, thus eliminating the need for plentiful human annotations. Wang, Li, Ferguson and Zhai [19] proposed a completely unsupervised, probabilistic topic model to mine, for each drug, the related ADRs mentioned by online health forum users. The model was able to distinguish ADRs from just disease symptoms by explicitly modeling the two components separately, therefore, enabled the mining of meaningful ADRs, some of which have never been reported in the FDA database. Liu and Chen [12] developed a framework for Drug-ADR relation extraction from patient forums, which was based on statistical learning.

Lastly, in a study closest in its methodology to ours, Netzer et al. [13] have utilized consumer-generated data posted on consumer forums to allow commercial marketing analysis. One of the cases investigated Drug-ADR relationships extracted from diabetes online forums. They show firm consistency between ADRs mentioned in diabetes online forums and ADRs reported in formal medical venues. Our methodology, to be described hereinbelow, also resides in the unsupervised learning yard, and will exhibit the following qualities: (1) An end-to-end text mining solution for Drug-ADR detection from online forums, which is unsupervised or semi-supervised with high degree of precision; (2) A text mining technology that is based on a HPSG-based parser; (3) A post-processing algorithm that enhances the set of Drug-

ADR extracted relations in order increase recall and allow statistical testing of the results.

3. METHODOLOGY

At the core of our mechanism for mining the Drug-Symptom relations lies an Unsupervised Relation Extraction (URE) framework. The framework, as thoroughly illustrated by Rozenfeld and Feldman [16], is a domain independent set of tools and methods for building Information Extraction (IE) and Relation Extraction (RE) systems focusing on natural language sentences. The framework includes both supervised and unsupervised methods, as well as allowing arbitrary blends between them. Purely unsupervised methods can be utilized in Open IE task setting (Banko and Etzioni [1]), in which the RE system trains itself to extract all possible target relations without any human input. Among other tools, the URE framework includes a grammar description language, which is able to define arbitrary weighted typed-feature-structure context-free grammars (WTFSCFG-s), a parser for such grammars, and a specific grammar and lexicon for parsing general English sentences. This grammar is a generic, HPSG-based English grammar with an underspecified lexicon. In the lexicon, only the most frequent and functional words have full definitions, while the open classes of words are defined with underspecified lexical entries and tightly-integrated separate PoS and NER components. For general English sentences, URE produces parses of only mediocre quality. However, when its lexicon is augmented with focused domain-specific entries, the quality of parses for precisely the relevant (i.e. containing interesting relations) domain-specific sentences is dramatically improved. The domain-specific lexical entries can be added manually or learned in an unsupervised manner from an unlabeled text corpus.

We utilize the URE framework and augment its lexicon with respect to the medical domain, in an attempt to mine Drug-ADR relations. The URE framework serves as our main processing flow: relation pattern acquisition followed by relation extraction. The downside of using this method, as shown later in our empirical application, is that while being highly accurate, it can suffer from low recall. Only patterns that are both semantically and linguistically meaningful can be directly converted into lexical definitions, which severely restricts the possible pattern forms. To mitigate this problem and increase recall we employ an additional post-processing step.

Next, we outline the steps involved in our text mining methodology.

3.1 Pre-processing

Following are the steps that are taken as part of a pre-processing flow.

1. *Download and Parse*: Download relevant Web pages from various popular medical forums in HTML format. Then, parse each HTML-formatted page and extract vital information from it: thread link, thread ID, thread title. Additionally, for each message in a thread, i.e. the original message and its replies, extract its user name, date and its textual content. These operations are performed using Scrapy[1].

2. *Remove Duplicates*: Filter out identical messages by using the Python Standard Library's *md5* module, which implements RSA's MD5 message-digest algorithm for generating a 128-bit hash value that serves as a fingerprint for each message. The hash value is produced on the string resulting in concatenating a lowered-case message's text, its date and user name.

3. *Break Into Sentences*: Split each message into sentences. This is done using Natural Language Toolkit (NLTK)[2]. More specifically, using a NLTK's *tokenize* module, that uses an unsupervised learning approach to build a model for finding sentence boundaries.

3.2 Relation Pattern Acquisition

After obtaining the pertinent medical forums texts, and after breaking all unique message into sentences, and preparing them for the mining process, we continue to the learning phase, in which relevant Drug-Symptom relation patterns are generated by applying an acquisition process on those unlabeled texts. We follow the Drug and Symptoms entity identification procedure, described in detail in the Web Appendix in [13]. That is, the Drug entity is a dictionary-based entity type, while the Symptom entity is a combination of dictionary-based and pattern-based entity type. The acquisition process amounts to the following steps.

1. Run the Automatic Lexicon Acquisition mode of URE on a sample of the data in order to learn Drug-Symptom relation patterns between the following entity types: Person, Drug, Symptom and, when relevant, Disease.

2. Manually remove irrelevant relations that have been generated. In addition, cluster relation patterns if necessary and promote relevant relation patterns that were discovered but did not cross a predefined threshold for them to be considered sufficiently supported to comprise relation patterns.

3.3 Extraction

After learning the relevant patterns, the URE lexicon acquisition component converts them into HPSG lexical entries and adds them to the lexicon, which is then used to parse the entire input corpus. The output of this process includes all extracted relevant entities (Drug, Symptom, etc.), and relations (Drug-Symptom, Person-Drug, Person-Symptom, etc.) in the form of HPSG semantic structures, which then can be converted into a standard "relation frame" form. The extracted entities are further used during the post-processing stage.

3.4 Post-processing

When taking only straightforwardly extracted Drug-Symptom relations, URE often misses many valuable relations that are mentioned in indirect ways. For example, a relation slot may be missing due to unresolved co-references or because of the main relation verb being in a gerund form or in an actor-less passive form. There are also more elusive syntactic relation forms which cannot be easily learned, and also relations which are semantically implied but not stated directly. Therefore, a post-processing stage has been designed

---

[1]An open-source web crawling/web scraping framework, which is written in Python. Available at *scrapy.org.*

[2]A suite of libraries for Natural Language Processing (NLP), which is written in Python. Available at *www.nltk.org.*

to try and catch those missing relations, thus boosting the recall while keeping the precision at a high level.

In the following description of the post-processing algorithms, we will refer to Drug-Symptom relations as *full relation* (FE), and to relations that are not themselves Drug-Symptom relations, but contain either a Drug slot or a Symptom slot, as *partial relation*. For example, each of Rel_PERSON_take_DRUG and
Rel_PERSON_experience_SYMPTOM are partial relations.

3.4.1 Types of Post-processed Relation Extractions

Apart from the full relation extractions that are straight-forwardly extracted by URE, we employ the following three post-processing methods to extract additional relations:

3.4.2 Merging Partial Relations (MPR)

Detect two partial relations that complete each other, either in the same sentence or in adjacent sentences. In our case, it means one that is missing a Drug entity in the relation and the other that is missing a Symptom entity in the relation. For example, from the sentence in Figure 1 two relations are extracted: Rel_PERSON_take_DRUG and Rel_PERSON_suffer_SYMPTOM. Thus, it is very natural and intuitive to deduce a Drug-Symptom relation in this sentence between Lipitor and memory loss.

I took Lipitor and {I} suffered muscle weakness and memory loss.

Figure 1: An example of a MPR candidate. (Curly brackets denote an implicit word in the sentence.)

3.4.3 Merging a Partial Relation with an Entity (MPRE)

Detect a partial relation along with an entity of the type that the partial relation lacks, either in the same sentence or in adjacent sentences. For example, in the sentences in Figure 2, the relation Rel_PERSON_take_DRUG is extracted from the first sentence and several Symptom entities, including memory loss, are extracted from the second sentence, allowing us to infer the statins-memory loss relation.

My husband took statins for 9 years, the last one was Lipitor. Side effects included severe neck and shoulder pain, muscle atrophe, loss of muscle strength and both short term and long term memory loss.

Figure 2: An example of a MPRE candidate.

3.4.4 Merging Same Sentence Entities (MSSE)

Detect two complementary entities in the same sentence. This heuristics is a special case of the co-occurrence approach. However, unlike the naïve co-occurrence approach, we severely restrict the extractions by distance between entities, and by requiring the absence of other entities of the same types in the sentence. Naturally, this heuristic is applied after all the other methods, and only if they cannot be applied. These restrictions limits false-positive cases, while serving as a safety net for catching unextracted relations.

This post-processing method also has a development-related aspect, as it can assist us in manually expanding the list of relation patterns by manually reviewing such MSSE extractions in the development set.

3.5 Post-Processing Algorithms

The two algorithms presented hereinafter together implement the post-processing flow. The first algorithm, a relation generation algorithm, is responsible for generating candidates for relations that have been missed by URE. The second algorithm filters possibly superfluous relations generated by the first algorithm. The reason for separating the two algorithms stems from the fact that the relation generation algorithm makes only local decisions, on whether or not to add the current extraction as a relation candidate. On the other hand, the filtering algorithm has a global scope, and checks all relations generated from the message when making its decisions.

Algorithm 1 Post-processing: Relation Generation

1: **In**: A list of extractions E, $EntTyp_1$, $EntTyp_2$
2: **Out**: A list of post-processed extractions PPE
3: $ME \leftarrow$ Group E by message
4: **for all** me in ME **do**
5:     $SenInd2PartExt \leftarrow$ Map me's partial ext. by sentence index

6:     $EntExt_1 \leftarrow$ Map me's $EntTyp_1$ ent. by sentence index
7:     $EntExt_2 \leftarrow$ Map me's $EntTyp_2$ ent. by sentence index
8:     **for all** e in me **do**
9:         **if** $|e.drugs| \geq 1$ **and** $|e.symptoms| \geq 1$ **then**
10:           $e.type \leftarrow$ FE
11:           $PPE \leftarrow PPE||\mathbf{addRel}(e)$
12:         **end if**
13:     $twoPartExtSentInd \leftarrow$ sentence indices with
14:     at least two partial extractions in SenInd2PartExt
15:     **for all** si in twoPartExtSentInd **do**
16:         $twoPartExt \leftarrow$ SenInd2PartExt[si]
17:         **for all** (peA, peB) in twoPartExt **do**
18:           **if** $|peA.EntTyp_1| = 0$ **and** $|peA.EntTyp_2| \geq 1$ **and** $|peB.EntTyp_1| \geq 1$ **and** $|peB.EntTyp_2| = 0$ **then**
19:             $e \leftarrow$ Merge peA and peB into relation e with e.type = MPR
20:             $PPE \leftarrow PPE||\mathbf{addRel}(e)$
21:           **end if**
22:         **end for**
23:     **end for**
24:     Repeat adding MPR's logic with 1 sentence-distance.
25:     **for all** (si, pe) in SenInd2PartExt **do**
26:         $sentDist \leftarrow 0$
27:         $missEntTyp \leftarrow pe.missEntTyp$
28:         $adjSent \leftarrow$ Using relevant $EntExt_i$, 0 sentDist sentences in relation to si having missEntTyp (excluding anaphora)

29:         **for all** s in adjSent **do**
30:           $e \leftarrow$ Merge pe and entity of type missEntTyp from s to relation e with e.type = MPRE
31:           $PPE \leftarrow PPE||\mathbf{addRel}(e)$
32:         **end for**
33:     **end for**
34:     $twoEntSentInd \leftarrow$ sentence indices with (at least) two entities of different types, based on $EntExt_1$ and $EntExt_2$

35:     **for all** si in twoEntSentInd **do**
36:         $e \leftarrow$ Merge each possible combination of entities into relation e with e.type = MSSE
37:         $PPE \leftarrow PPE||\mathbf{addRel}(e)$
38:     **end for**
39:     Repeat adding MPRE's logic with 1 sentDist.
40:     Repeat adding MPRE's logic with 2 sentDist.
41:   **end for**
42: **end for**

43: $\mathbf{addRel}(e)$: Break extraction having more than one drug and/or symptoms into multiple extractions.

Algorithm 2 Post-processing: Relation Filtering

1: **In**: A list of post-processed extractions PPE, $EntTyp_1$, $EntTyp_2$, A predefined list of weak anaphoras WA
2: **Out**: A list of filtered post-processed extractions FPPE
3: $ME \leftarrow$ Group PPE by message
4: **for all** me in ME **do**
5:   $SME \leftarrow$ Sort by sentence distance then by extraction type

6:   $PriorE \leftarrow []$
7:   **for all** e in SME **do**
8:     $localSkipCond1 \leftarrow$ e.ent_1 and e.ent_2 are anaphoras
9:     $localSkipCond2 \leftarrow$ e.ent_1 in WA or e.ent_2 in WA
10:     **if** localSkipCond1 **or** localSkipCond2 **then**
11:       continue
12:     **end if**
13:     $skipExtraction \leftarrow$ **false**
14:     **for all** pe in PriorE **do**
15:       $skipCond_1 \leftarrow$ e.senDist >0 **and** e.supSenInd = pe.senInd

16:       $skipCond_2 \leftarrow$ e.senDist >0 **and** e.supSenInd = pe.supSenInd **and** e.senInd $<>$pe.senInd
17:       $skipCond_3 \leftarrow$ e.senDist = 0 **and** e.senInd = pe.senInd **and** pe.type = FE **or** pe.type = MPR
18:       $skipCond_4 \leftarrow$ e.senDist = 0 **and** e.senInd = pe.senInd **and** e.ent_1 is anaphora **or** e.ent_2 is anaphora
19:       **if** any $skipCond_i$ holds **then**
20:         skipExtraction = **true**
21:         break
22:       **end if**
23:       **if not** skipExtraction **then**
24:         $PriorE \leftarrow PriorE || e$
25:         $FPPE \leftarrow FPPE || e$
26:       **end if**
27:     **end for**
28:   **end for**
29: **end for**

3.6 Analysis

We compute a lift measure to evaluate the likelihood of a particular Drug-ADR relationship to occur over and beyond chance. We then use a chi-square statistic to evaluate the statistical significance of the lift measure.

One important indication that sheds light on a Drug-ADR association is the *lift* metrics. The concept of lift, also known as *association ratio* and *point-wise mutual information* ([4, 18]), serves as a method for measuring the strength of the semantic association between words, by calculating the degree of statistical dependence between them. Our methodology aims at obtaining data regarding Drug-Symptom relations rather than mere co-occurrence of Drug and Symptom entities. Thus, given a drug D and a symptom S, we define:

$$\text{lift}(D, S) = \frac{\Pr(D, S)}{\Pr(D) \times \Pr(S)}$$

In this extended definition $\Pr(D, S)$ relates to the probability of a relation extraction, while $\Pr(D)$ and $\Pr(S)$ relates to the probability of an entity extraction.

In calculating lifts in our case the researcher has a decision to make with respect to what is the population of messages to be included in the denominator of the lift probabilities. On the one hand, one can include the entire population of messages whether or not the messages includes a Drug-Symptom relation. This is likely to lead a set of lift measures that are greater than 1, as many of messages include neither drug nor a symptom. An alternative is to include in the denominator only messages that include at least one Drug-Symptom relation. In our empirical application we include both measures of lift.

- *classic-induced lift*:

$$\frac{\Pr(\text{message has } D - S \text{ relation})}{\Pr(\text{message has D entity}) \times \Pr(\text{message has S entity})}$$

In this formula the population refers to *all* of the messages that were downloaded from the medical forums.

- *relation-driven lift*:

$$\frac{\Pr(D_0 - S_0 \text{ relation})}{\sum_i \Pr(D_i - S_0 \text{ relations}) \times \sum_i \Pr(D_0 - S_i \text{ relations})}$$

In this calculation the population is the sum of Drug-Symptom relation extractions. Note that this means a message is bound to have a Drug-Symptom relation extraction in order to be taken into account, specifically it is not enough for it to include just Drug and Symptom entity extractions, which don't relate one another. Put in other words, this lift calculation factors only relation extractions and ignores plain entity extractions.

We report these two lift metrics in our empirical applications. In addition to lift measures we use a chi-square test to measure the statistical significance of the difference between the calculated lift measure and a lift value of 1, which represents the lift expected by chance.

4. EMPIRICAL APPLICATIONS

One could use our proposed approach to extract Drug-ADR relations for each drug and compare them to ADRs that are "formally" reported on the drug label and pharmaceutical drug website. Several studies have taken that approach ([13], [21], [22]). However, that approach is descriptive rather than predictive. Specifically, it is not clear whether the mention of the Drug-ADR relation in social media is a reflection of the medical relation between the two or simply recitation of information posted by the FDA and/or pharmaceutical firm about the relation. In this study we wish to put our proposed approach under a tougher test of *predicting* Drug-ADR relations from their mention in social media prior to their reporting by the FDA. We present our above-described methodology applied to two case studies from different sub-domains. In addition to the extracted Drug-ADR mentions in online forums, we collected safety communications, reported by the FDA regarding one or more ADRs that were associated with one or more drugs post-marketing of the drug. Our goal is to test for substantial, statistically significant evidence for Drug-ADR relations mentions in forum messages, that were posted prior to the FDA communications. Such analysis could provide a retrospective proof that those Drug-ADR relations could have been detected beforehand. Before focusing on each case study, here are a few implementation details:

Data: We mined 4 online medical forums and monitored texts spreading over 15 years, from 1999 to 2013.

Extracted Drugs: Apart from a chemical name, each drug has a brand name (trade name). Moreover, often a drug is marketed by more than just one brand name. For example, *Bupropion* is the chemical name for the antidepressant medication *Wellbutrin*, however it is also marketed by

other brand names, such as *Aplenzin* and *Zyban*. Further-more, drugs are occasionally mentioned in online forums by an acronym or with common typos. In keeping with the last example, in several messages *Wellbutrin* was referred to as *WB* or as *Wellbutrine*. Thus, when mining for Drug-ADR relations we searched for all of those variations for the drug name, based on the Drugs.com[3] website, a popular on-line source of drug information. In the few cases when we stumbled upon a drug name that was missing in our com-prehensive Drug dictionary, consisting of nearly 10K entries (see the Web Appendix in [13] for details about forming this dictionary), it was manually added to it. Lastly, for the sake of our analysis we aggregated the results for each drug variation by referencing them all to one *head drug*.

Extracted Symptoms: When reviewing the set of extracted symptoms in our extracted Drug-ADR relations, we aggre-gated them into a set of disjoint clusters. Each subset in the partition was referenced by a *head symptom*. For in-stance, *cognitive impairment* is the head symptom of symp-toms such as *amnesia*, *blackouts*, *brain fog*, *cognitive prob-lems*, *confusion*, *feeling confused*, *forgetfulness*, etc. Those mappings were done manually, however note that it is a one-time chore for each symptom. Because symptoms are often common across therapeutic areas, as we progressed from one case study to another, less such mappings were left to be done. In the few cases when we came across a variation of a symptom that was missing in our comprehen-sive Symptom dictionary, consisting of nearly 15K entries (see the Web Appendix in [13] for details about forming this dictionary), it was manually added to it. In general, when deciding on a subset, as well as on its head symptom, we tried to be consistent with symptoms as presented on the WebMD[4] website, so that we will able to later on examine the correlation between our findings and the documented symptoms appearing on this website.

Evaluation: The collection of messages for each case study, was split into development set (25%) and test set (75%). This split was performed for each medical forum that con-tributed to the collection of messages, as well as for each year out of the 15 years that were surveyed in our case studies, resulting in equivalent sample sets.

Learning: We ran the Automatic Lexicon Acquisition mode of URE while setting the learning thresholds needed for pro-moting a pattern as follows: A frequency of 2 for a rela-tion (binary or trinary) or modifier and a frequency of 10 for an unary relation. During this process, 194 patterns were learned. Additional 46 patterns were manually added by reviewing the results of the development set, most of them were detected in post-processed results of the above-described *merged same sentence entities* type. All in all, the set was comprised of 246 patterns.

Acquisition: The unsupervised acquisition step was per-formed on a training set, which was comprised of 150K sen-tences (from each of the sub-domains). Note that this means training was performed based on a sample that, though equally representing all medical forums and all drugs and symptoms (not just the ones we will focus on in our experi-ments), comprised only a minority of the entire data.

Analysis: We computed the two measures of lift, the classic-induced and the relation-driven. Each lift with value

#	Class Name	Head Drugs
1	Bile Acid Sequestrants	Welchol, Questran, Colestid
2	Cholesterol Absorp-tion Inhibitors	Zetia
3	Fibric Acid Deriva-tives (Fibrates)	Tricor, Lopid, Trilipix, Atromid-S
4	Misc. Antihyperlipi-demic Agents	Niacin, Vascepa, Choloxin, Jux-tapid, Kynamro
5	Antihyperlipidemic Combinations	Vytorin, Advicor, Simcor, Caduet, Pravigard Pac, Juvisync, Liptruzet
6	Statins	Altoprev, Crestor, Lescol, Lipitor, Mevacor, Pravachol, Zocor, Livalo, Baycol

Table 1: Class names and their belonging head drugs.

Forum	No. of unique messages	No. of sentences	No. of unique usernames
medhelp.org	1,887	14,276	647
exchanges.webmd.com	5,492	32,693	854
healthboards.com	32,665	207,765	3,250
ehealthforum.com	1,042	7,150	562

Table 2: The data collected for the analysis of the Choles-terol Lowering sub-domain.

greater than 1 was considered a potential indicator for a Drug-ADR association. The chi-square test, statistically ex-amined the difference between the lift value and 1.

4.1 The Cholesterol Lowering Sub-Domain

In this experiment, we aimed to follow monitored ADRs, which are related to statins, a class of cholesterol lowering drugs. The cholesterol lowering sub-domain includes several *classes* of drugs. Statin drugs, for example, are often re-ferred to, by patients and doctors as well as in the medical literature, by the use of the collective name: *statins*. This aggregation might affect the output of a text mining task: On the one hand, it naturally thickens the number of extrac-tions for each class of drugs in question. On the other hand it might result in noisy output as not all drugs of a class share the same ADRs. Because our FDA label change was at the level of a class of drugs (statins) we mine Drug-ADR relations at the drug-class level.

We therefore aggregated the cholesterol lowering drugs un-der their drug classes. We followed the Drugs.com website, a popular online source of drug information, and categorized those drugs into 6 main classes alongside monitoring the class names themselves as drugs when seeking Drug-ADR relations. Table 1 presents the classes and the head drug that belongs to each class. All in all, 119 variations of drugs were monitored for this sub-domain. The collected data for this case study included $41K$ messages from 4 medical fo-rums as detailed in table 2.

4.1.1 Statins and Cognitive Impairment

In February 2012 the FDA approved important safety label changes for statins[5]. Among others, those changes included the addition of label information with regard to

---

Top Extracted ADRs	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Total
pain	12	4	12	20	11	453	512
muscle pain	9	0	14	37	9	374	443
flushing	2	0	0	2	180	16	200
heart attack	1	0	4	4	13	172	194
muscle damage	1	2	6	24	7	141	181
feeling weak	1	0	7	20	4	147	179
allergic reaction	3	2	2	8	21	101	137
liver failure	4	0	10	0	42	63	119
diabetes	6	2	2	5	17	78	110
cognitive impairment	1	0	4	2	2	95	104
leg pain	5	0	5	7	7	77	101
muscle problems	2	1	2	8	2	57	72
infection	1	0	2	1	9	59	72
leg cramps	3	2	2	4	2	56	69
muscle weakness	0	2	6	3	0	56	67
cancer	2	0	0	4	4	54	64
head pain	5	2	4	4	10	30	55
heart problems	0	0	1	0	2	51	54
stroke	2	0	2	3	1	42	50
burning sensation	1	0	0	0	4	38	43
Total	61	17	85	156	347	2,160	2,826

Figure 3: Leading extracted relations for the Cholesterol Lowering sub-domain, 1999-2011 (Class 6 is Statins).

the potential for non-serious and reversible cognitive side effects, such as memory loss or confusion. This announcement stemmed from post-marketing reports and data that the FDA had reviewed. For instance, the prescribing information for Lipitor, a member of the statins class, was revised to include the following note under post-marketing information clause: *There have been rare post-marketing reports of cognitive impairment (e.g., memory loss, forgetfulness, amnesia, memory impairment, confusion) associated with statin use.*[6]

We implemented our above-described apparatus to mine for Drug-Symptom relations between statins, as well as the other drug classes, and cognitive impairment, as well as other symptoms. Figure 3 presents the number of Drug-Symptom relationships between each drug class and main side effects mined from the test set data from 1999 to 2011 (the year that preceded the FDA's relevant communication). The relationship between statins and cognitive impairment was identified 95 times by our algorithm between 1999 and 2011. It was the 7th most common symptom to be mentioned with statins. Furthermore, the tables in Figure 4, demonstrate that the lift associated with the relationship between statins and cognitive impairment is above 1 and is statistically significant at the 0.01 level. One caveat of that analysis is that it is possible that most of mentions of statins and cognitive impairment appeared in the later years close to the label change announcement when the information about this relation became public. Table 3 presents analysis of the lift between statins and cognitive impairment by year from 2003 and 2011. A strong and significant relationship between statin and cognitive impairment is identified as early as 9 years prior to the FDA label change. Note also that we do not find similar relationship between cognitive impairment and other classes of cholesterol lowering drugs. Indeed, no such association has been reported.

Relation-Driven Lift	1	2	3	4	5	6	1	2	3	4	5	6
pain	1.1	1.3	0.8	0.7	0.2	1.2	0.1	0.3	0.0	0.0	0.0	50.3
muscle pain	0.9	0.0	1.1	1.5	0.2	1.1	0.0	0.0	0.0	8.1	0.0	18.6
flushing	0.5	0.0	0.0	0.2	7.3	0.1	0.0	0.0	0.0	0.0	1207.0	0.0
heart attack	0.2	0.0	0.7	0.4	0.5	1.2	0.0	0.0	0.0	0.0	0.0	17.3
muscle damage	0.3	1.8	1.1	2.4	0.3	1.0	0.0	0.8	0.1	22.2	0.0	0.2
feeling weak	0.3	0.0	1.3	2.0	0.2	1.1	0.0	0.0	0.5	11.7	0.0	3.4
allergic reaction	1.0	2.4	0.5	1.1	1.2	1.0	0.0	1.8	0.0	0.0	1.2	0.0
liver failure	1.6	0.0	2.8	0.0	2.9	0.7	0.9	0.0	12.4	0.0	61.1	0.0
diabetes	2.5	3.0	0.6	0.8	1.3	0.9	5.9	2.8	0.0	0.0	1.1	0.0
cognitive impairment	0.4	0.0	1.3	0.3	0.2	1.2	0.0	0.0	0.3	0.0	0.0	13.3
leg pain	2.3	0.0	1.6	1.3	0.6	1.0	3.9	0.0	1.4	0.4	0.0	0.0
muscle problems	1.3	2.3	0.9	2.0	0.2	1.0	0.1	0.8	0.0	4.4	0.0	0.3
infection	0.6	0.0	0.9	0.3	1.0	1.1	0.0	0.0	0.0	0.0	0.0	1.2
leg cramps	2.0	4.8	1.0	1.1	0.2	1.1	1.6	6.2	0.0	0.0	0.0	0.9
muscle weakness	0.0	5.0	3.0	0.8	0.0	1.1	0.0	6.5	8.3	0.0	0.0	1.9

Classic-Induced Lift	1	2	3	4	5	6	1	2	3	4	5	6
pain	0.8	0.8	0.6	1.0	0.1	1.7	0.0	0.0	0.0	0.0	0.0	144.1
muscle pain	1.0	0.0	1.0	2.3	0.2	2.2	0.0	0.0	0.0	23.4	0.0	298.0
flushing	0.5	0.0	0.0	0.3	7.6	0.2	0.0	0.0	0.0	0.0	1017.8	0.0
heart attack	0.1	0.0	0.3	0.3	0.2	0.9	0.0	0.0	0.0	0.0	0.0	0.0
muscle damage	0.3	1.6	1.2	3.4	0.4	2.0	0.0	0.5	0.3	29.8	0.0	76.5
feeling weak	0.2	0.0	0.8	2.4	0.1	1.5	0.0	0.0	0.0	15.8	0.0	29.8
allergic reaction	0.8	1.3	0.3	1.3	1.0	1.4	0.0	0.2	0.0	0.5	0.0	12.6
liver failure	1.6	0.0	2.2	0.0	2.9	1.3	0.9	0.0	6.4	0.0	56.3	4.9
diabetes	0.8	0.8	0.2	0.5	0.5	0.7	0.0	0.0	0.0	0.0	0.0	0.0
cognitive impairment	0.5	0.0	1.3	0.6	0.2	2.0	0.0	0.0	0.2	0.0	0.0	49.3
leg pain	2.2	0.0	1.7	2.0	0.6	2.0	2.5	0.0	1.5	3.1	0.0	49.9
muscle problems	1.6	2.0	1.0	3.8	0.3	2.3	0.4	0.5	0.0	17.3	0.0	58.5
infection	0.2	0.0	0.2	0.1	0.3	0.6	0.0	0.0	0.0	0.0	0.0	0.0
leg cramps	1.7	2.9	0.7	1.4	0.2	1.7	1.0	2.6	0.0	0.5	0.0	18.5
muscle weakness	0.0	3.5	2.6	1.3	0.0	1.9	0.0	3.5	5.9	0.2	0.0	30.6

Figure 4: Lifts and respective chi-square values for leading extracted relations for the Cholesterol Lowering sub-domain, 1999-2011. Each lift cell with value that is greater than 1 is filled with color green, and its chi-square value is calculated. Each such equivalent chi-square value corresponding p-value ≤ 0.05 and p-value ≤ 0.01 is filled with pale blue and blue, respectively.

Year	Relation-driven lift	Chi-square value	Classic-induced lift	Chi-square value
2011	1.20	**13.33**	1.99	**49.28**
2010	1.21	**13.24**	1.94	**42.21**
2009	1.22	**13.35**	1.97	**40.03**
2008	1.21	**10.70**	1.89	**31.42**
2007	1.20	**9.95**	2.00	**36.46**
2006	1.21	**10.30**	1.89	**28.20**
2005	1.20	**6.63**	2.04	**25.12**
2004	1.25	3.46	2.18	**12.93**
2003	1.27	1.55	2.16	**5.79**

Table 3: Lifts and respective chi-square values by year for the 9 years that preceded the relevant FDA label change for statin and cognitive impairment. All values in bold are chi-square values ≥ 3.85 or ≥ 6.64, corresponding p-value ≤ 0.05 or ≤ 0.01.

4.2 The Anti-Depression Sub-Domain

This sub-domain presented us with an additional text mining challenge. The ADR in question, agitation, was referred to by user of the medical forums not only as a symptom of drug but as a symptom of depression, or even as a

Forum	No. of unique messages	No. of sentences	No. of unique usernames
medhelp.org	86,715	729,888	25,049
exchanges.webmd.com	17,403	138,936	1,310
healthboards.com	149,238	1,129,341	16,381
ehealthforum.com	20,634	196,765	9,688

Table 4: The data collected for the analysis of the Anti-Depression sub-domain.

disease of its own. To address this issue we added a new entity type to the framework: DISEASE along with a very basic lexicon of common diseases. After doing so we repeated the acquisition process, this time enabling a richer set of patterns, in which 4 entity types participate: Person, Drug, Symptom and Disease. During this process, we enriched the set of patterns from 246 to 256 patterns.

For this case study we monitored a list of 6 popular anti-depression drugs, whose head drug are the following: *Celexa*, *Effexor*, *Pristiq*, *Wellbutrin*, *Xanax*, and *Zoloft*. All in all, 32 variations of drugs were monitored for this sub-domain. The collected data included $274K$ messages from 4 medical forums as detailed in Table 4.

4.2.1 Wellbutrin and Agitation

In an FDA alert from July 2009[7] the manufacturers of Wellbutrin were required to add new box warnings highlighting the risk of serious neuropsychiatric symptoms in patients using the drug. Among those symptoms were agitation and hostility. Healthcare professionals were also instructed to monitor patients taking Wellbutrin for those symptoms. In addition, FDA also required those manufacturers to describe those risks in the Medication Guides for these products.

The table in Figure 5 presents the accumulated Drug-Symptom relations mined from the test set data from 1999 to 2008, the year preceding the FDA's alert discussed above. Wellbutrin was mentioned 77 times in our forums with agitation. Agitation is the 6th most frequently mentioned symptom with Wellbutrin. Looking at the tables in Figure 6, we see that the relationship between Wellbutrin and agitation is statistically significant. Moreover Wellbutrin is the only anti-depression drug that demonstrates that relationship. As can be seen in Table 5, as early as 7 year prior to the FDA communication we find strong and significant relationship between Wellbutrin and agitation.

5. DISCUSSION

We have presented two out of several case studies that we carried out on different medical sub-domains, involving various drugs and ADRs. We used lift measures and corresponding statistical chi-square test primarily to evaluate the quality of our overall approach to predict the eventual outcome of a medical drug-ADR relation. All of those cases demonstrated strong and significant relationship between the drug and the ADR reported by the FDA *prior* to the FDA communication and sometimes years prior to the communication. However, the results are only as good as

[7] http://www.fda.gov/Drugs/DrugSafety/
PostmarketDrugSafetyInformationforPatientsandProviders/
DrugSafetyInformationforHeathcareProfessionals/
ucm169986.htm

Top Extracted ADRs	celexa	effexor	pristiq	wellbutrin	xanax	zoloft	Total
anxiety	217	359	3	343	423	492	1,837
weight gain	71	144	1	130	11	169	526
head pain	56	127	1	100	21	102	407
panic state	44	70	0	19	133	96	362
sleep disorder	43	78	1	86	49	104	361
allergic reaction	46	69	0	67	23	98	303
feeling weak	40	74	3	48	15	60	240
pain	21	74	0	38	30	60	223
tremors	29	59	0	30	43	57	218
agitation	24	46	0	77	11	54	212
nausea	22	89	2	33	7	51	204
sweating	22	103	1	21	5	42	194
seizure	5	16	0	113	40	16	190
dizziness	18	66	1	25	9	50	169
suicidality	15	67	0	18	9	36	145
sexual dysfunction	21	37	0	46	2	28	134
cognitive impairment	10	41	1	17	23	21	113
weight loss	5	23	0	61	2	18	109
mood swings	18	36	0	21	3	27	105
sleepiness	26	32	0	17	5	17	97
Total	753	1,610	14	1,310	864	1,598	6,149

Figure 5: Leading extracted relations for the Anti-Depression sub-domain, 1999-2008.

Relation-Driven Lift	C	E	P	W	X	Z	C	E	P	W	X	Z
anxiety	1.0	0.7	0.7	0.9	1.6	1.0	0.0	0.0	0.0	0.0	174.7	0.9
weight gain	1.1	1.0	0.8	1.2	0.1	1.2	0.8	0.4	0.0	4.0	0.0	11.3
head pain	1.1	1.2	1.1	1.2	0.4	1.0	0.9	5.7	0.0	2.8	0.0	0.0
panic state	1.0	0.7	0.0	0.2	2.6	1.0	0.0	0.0	0.0	0.0	164.0	0.1
sleep disorder	1.0	0.8	1.2	1.1	1.0	1.1	0.0	0.0	0.0	1.5	0.0	1.6
allergic reaction	1.2	0.9	0.0	1.0	0.5	1.2	2.6	0.0	0.0	0.1	0.0	6.7
feeling weak	1.4	1.2	5.5	0.9	0.4	1.0	4.5	2.8	11.5	0.0	0.0	0.0
pain	0.8	1.3	0.0	0.8	1.0	1.0	0.0	5.9	0.0	0.0	0.0	0.1
tremors	1.1	1.0	0.0	0.6	1.4	1.0	0.2	0.1	0.0	0.0	6.0	0.0
agitation	0.9	0.8	0.0	1.7	0.4	1.0	0.0	0.0	0.0	29.5	0.0	0.0
nausea	0.9	1.7	4.3	0.8	0.2	1.0	0.0	33.2	5.3	0.0	0.0	0.0
sweating	0.9	2.0	2.3	0.5	0.2	0.8	0.0	75.1	0.7	0.0	0.0	0.0
seizure	0.2	0.3	0.0	2.8	1.5	0.3	0.0	0.0	0.0	170.4	8.0	0.0
dizziness	0.9	1.5	2.6	0.7	0.4	1.1	0.0	14.9	1.0	0.0	0.0	1.2
suicidality	0.8	1.8	0.0	0.6	0.4	1.0	0.0	30.8	0.0	0.0	0.0	0.0

Classic-Induced Lift	C	E	P	W	X	Z	C	E	P	W	X	Z
anxiety	0.8	0.6	0.3	0.7	2.0	0.9	0.0	0.0	0.0	0.0	211.9	0.0
weight gain	1.4	1.4	0.6	1.6	0.3	1.7	8.8	19.2	0.0	25.9	0.0	51.8
head pain	0.9	1.0	0.5	1.0	0.5	0.9	0.0	0.1	0.0	0.0	0.0	0.0
panic state	1.0	0.8	0.0	0.3	4.2	1.2	0.0	0.0	0.0	0.0	335.2	2.9
sleep disorder	0.9	0.8	0.7	1.1	1.4	1.0	0.0	0.0	0.0	0.6	6.8	2.3
allergic reaction	1.1	0.8	0.0	0.9	0.8	1.2	0.7	0.0	0.0	0.0	0.0	3.8
feeling weak	0.8	0.7	2.0	0.6	0.4	0.7	0.0	0.0	1.6	0.0	0.0	0.0
pain	0.2	0.3	0.0	0.2	0.4	0.3	0.0	0.0	0.0	0.0	0.0	0.0
tremors	1.2	1.1	0.0	0.7	2.2	1.1	0.7	0.5	0.0	0.0	27.2	0.4
agitation	0.9	0.9	0.0	1.8	0.6	1.1	0.0	0.0	0.0	28.8	0.0	0.5
nausea	0.8	1.7	2.4	0.7	0.4	1.0	0.0	24.1	1.6	0.0	0.0	0.0
sweating	1.1	2.4	1.7	0.7	0.4	1.1	0.1	74.2	0.3	0.0	0.0	0.2
seizure	0.3	0.4	0.0	3.5	3.0	0.5	0.0	0.0	0.0	187.4	50.3	0.0
dizziness	0.8	1.4	1.4	0.7	0.6	1.2	0.0	6.6	0.1	0.0	0.0	1.5
suicidality	0.3	0.7	0.0	0.3	0.3	0.5	0.0	0.0	0.0	0.0	0.0	0.0

Figure 6: Lifts and respective chi-square values for leading extracted relations for the Anti-Depression sub-domain, 1999-2008.

our ability to mine the Drug-ADR relations. Therefore, we now discuss the quality of our text mining approach. Given the manageable number of extracted relations we annotated each and every extracted relation of focus for its correct-

Annotated Extracted ADRs (Development Set)	Sub-domain	Precision	True Positive Extractions	False Positive Extractions	Total Extractions	Distribution by Impression		Distribution by Extraction			
						Subjective Impressions	Objective Impressions	FE Type	MPR Type	MPRE Type	MSSE Type
cognitive impairment	Cholesterol Lowering	100.0%	38	0	38	76%	24%	32%	8%	29%	31%
agitation	Anti-Depression	89.2%	58	7	65	84%	16%	6%	8%	57%	29%
Total		93.2%	96	7	103	81%	19%	16%	8%	47%	30%

Annotated Extracted ADRs (Test Set)	Sub-domain	Precision	True Positive Extractions	False Positive Extractions	Total Extractions	Distribution by Impression		Distribution by Extraction				Recall	True Positive Extractions	False Negative Extractions	Total Extractions
						Subjective Impressions	Objective Impressions	FE Type	MPR Type	MPRE Type	MSSE Type				
cognitive impairment	Cholesterol Lowering	96.9%	125	4	129	55%	45%	32%	5%	39%	24%	71.7%	104	41	145
agitation	Anti-Depression	83.4%	231	46	277	84%	16%	9%	7%	51%	33%	69.1%	105	47	152
Total		87.7%	356	50	406	75%	25%	17%	6%	47%	30%	70.4%	209	88	297

Figure 7: Annotated Drug-ADR extractions for both the development and test sets.

Year	Relation-driven lift	Chi-square value	Classic-induced lift	Chi-square value
2008	1.70	**29.53**	1.81	**28.79**
2007	1.75	**28.70**	1.83	**26.97**
2006	1.72	**21.13**	1.61	**12.91**
2005	1.69	**16.54**	1.78	**15.90**
2004	1.46	**5.13**	1.51	**5.13**
2003	1.64	**8.03**	2.05	**13.99**
2002	1.78	**6.41**	2.36	**12.25**

Table 5: Lifts and respective chi-square values by year, for the 7 years that preceded the relevant FDA communication for Wellbutrin and agitation. All values in bold are chi-square values ≥ 3.85 or ≥ 6.64, corresponding p-value ≤ 0.05 or ≤ 0.01.

Impression Type	Definition
Subjective	Extractions, in which a patient was confessing about a personal experience or an experience of a close family member or a friend, e.g., *I had **memory problems** with **Simvastatin** also to the point that I forgot where I was while driving.*
Objective	Extractions, in which the writer either quotes facts from scientific articles or information gained from the media, or states a general saying (e.g. hearsay) about a Drug-ADR connection, e.g., *There are more people out here having **memory loss problems** from **statin** drug that anyone can count.*

Table 6: Different types of annotated impressions.

ness. The aggregated results of our annotations are detailed in Figure 7.

One can see that precision rates are indeed high: 88% overall precision for the test set. Precision for the development set is only moderately higher at 93%, suggesting that the manually added rules, which were affected by the development set, were not over-fitted much by the data. By adding the Disease entity type into the set of extracted entities and relations, our results for the test set have been moderately improved from a 80.7% precision to 83.4% precision (the results for the development set have been upgraded from 80.3% to 89.2%). Although the core of our research is in the design of a text mining methodology capable of predicting Drug-ADR relations, we also wanted to have a recall estimation of our relation extractions. Based on the test set messages between 1999-2011 (2012 being the pivotal year), we measured a 71.7% recall for relations involving the cognitive impairment Symptom entity. As for Drug-ADR relations with respect to the agitation Symptom entity, we measured a 69.1% recall on test set messages between 1999-2008 (2009 being the pivotal year). When combining those scores for recall and precision we get 82.4% and 75.6% F1-Scores, respectively.

Next, the distribution by extraction type indicates that without the post-processing step, our analysis would have reached a substantially lower recall. Thus, post-processing comprises a crucial step for establishing our findings.

When annotating the extractions, we also annotated them for their impressions (this concept is elaborated on Table 6). Note that the majority of the relation extractions are characterized as expressing subjective impressions. Patients, as

it seems, do tend to share their user experience with respect to drugs. Notwithstanding, the table also suggests that the dominance of the subjective impression dramatically varies from one ADR to another. It is considerably stronger for the agitation ADR compared to the cognitive impairment ADR. One might suggest it could stem from the characteristics of the ADR, mainly its manifestation in our lives. In this respect, this result coincides with a previous finding by Yang, Yang, Jiang and Tang [21], who show that their proposed technique was promising for detecting ADRs such as heart condition and diarrhea, however not effective for ADRs such as cancer.

6. CONCLUSION AND FUTURE WORK

In this paper we have presented an end-to-end, step-by-step text mining methodology for the process of extracting Drug-ADR relations from online message boards. Our methodology combined several advantages that have made it effective as a tool kit for practical use. Despite being unlaborious, our approach has shown to be precise and more importantly, able to predict Drug-ADR relations that have not been announced at the time of their mention in the medical forums. Using several case studies, we demonstrated that one could find significant traces in medical forums for ADRs related to different drugs years before these ADRs are reported by the FDA.

One can suggest a few research directions, which may enhance our current methodology and its outcomes, including:

Augmenting the pre-processing step by amending irregularities such as corrupt grammar, informal syntax, typos and extensive use of acronyms, aiming at reducing the amount of sentences for which the parse failed; Extending the corpus by mining more online message boards or other social media sources (e.g., Twitter feed), as obtaining more raw data may be crucial when mining less popular sub-domains, when looking into less common drugs, or when addressing less discussed ADRs; Lastly, perfecting our text mining algorithm, among others: Exploring improved approaches to distinguish Drug-Disease relations from Drug-ADR relations as well as bringing to a minimum cases of Drug-ADR false alarms.

7. REFERENCES

[1] M. Banko, O. Etzioni, and T. Center. The tradeoffs between open and traditional relation extraction. In *ACL*, volume 8, pages 28–36. Citeseer, 2008.

[2] A. Benton, L. Ungar, S. Hill, S. Hennessy, J. Mao, A. Chung, C. E. Leonard, and J. 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.

[3] B. W. Chee, R. Berlin, and B. Schatz. Predicting adverse drug events from personal health messages. 2011:217–226, 2011.

[4] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

[5] R. Feldman, M. Fresko, J. Goldenberg, O. Netzer, and L. H. Ungar. Extracting product comparisons from discussion boards. In *International Conference on Data Mining (ICDM)*, pages 469–474, 2007.

[6] R. Feldman, M. Fresko, J. Goldenberg, O. Netzer, and L. H. Ungar. Using text mining to analyze user forums. In *International Conference on Service Systems and Service Management (ICSSSM)*, pages 1–5, 2008.

[7] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton, and T. Tomokiyo. Analyzing online discussion for marketing intelligence. In *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, pages 1172–1173, 2005.

[8] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton, and T. Tomokiyo. Deriving marketing intelligence from online discussion. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 419–428, 2005.

[9] S. Hariharan and T. Ramkumar. Mining product reviews in web forums. In *Information Retrieval Methods for Multidisciplinary Applications*, pages 78–94. IGI Global, 2013.

[10] R. Leaman, L. Wojtulewicz, R. Sullivan, A. Skariah, J. Yang, and G. 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, 2010.

[11] T. Y. Lee and E. T. Bradlow. Automated marketing research using online customer reviews. *Journal of Marketing Research*, 48(5):881–894, 2011.

[12] X. Liu and H. Chen. Azdrugminer: An information extraction system for mining patient-reported adverse drug events in online patient forums. In *Smart Health*, pages 134–150. Springer, 2013.

[13] O. Netzer, R. Feldman, J. Goldenberg, and M. Fresko. Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, 31(3):521–543, 2012.

[14] A. Nikfarjam and G. H. Gonzalez. Pattern mining for extraction of mentions of adverse drug reactions from user comments. In *American Medical Informatics Association (AMIA) Annual Symposium Proceedings 2011*, pages 1019–1026. American Medical Informatics Association, 2011.

[15] A. Patki, A. Sarker, P. Pimpalkhute, A. Nikfarjam, R. Ginn, K. O'Connor, K. Smith, and G. Gonzalez. Mining adverse drug reaction signals from social media: Going beyond extraction. In *Proceedings of the 22nd Annual International Conference on Intelligent Systems for Molecular Biology (ISBM)*, pages 9–16, 2014.

[16] B. Rozenfeld and R. Feldman. Unsupervised lexicon acquisition for hpsg-based relation extraction. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1890–1895, 2011.

[17] I. Segura-Bedmar, S. de la Pena, and P. Martınez. Extracting drug indications and adverse drug reactions from spanish health social media. In *Proceedings of the 2014 Workshop on Biomedical Natural Language Processing*, pages 98–106, 2014.

[18] P. D. Turney and M. L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, 2003.

[19] S. Wang, H. Lin, D. Ferguson, and C. Zhai. Sideeffectptm: An unsupervised topic model to mine adverse drug reactions from health forums. 2014.

[20] R. W. White, N. P. Tatonetti, N. H. Shah, R. B. Altman, and E. Horvitz. Web-scale pharmacovigilance: Listening to signals from the crowd. *Journal of the American Medical Informatics Association (JAMIA)*, 20(3):404–408, 2013.

[21] C. C. Yang, L. Jiang, H. Yang, and X. Tang. Detecting signals of adverse drug reactions from health consumer contributed content in social media. In *Proceedings of ACM SIGKDD Workshop on Health Informatics*, 2012.

[22] C. C. Yang, H. Yang, L. Jiang, and M. Zhang. Social media mining for drug safety signal detection. In *Proceedings of the 2012 International Workshop on Smart Health and Wellbeing (SHB)*, pages 33–40, 2012.