A Timeline Extraction Approach to Derive Drug Usage Patterns in Pregnant Women Using Social Media

by

Pramod Bharadwaj Chandrashekar

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved April 2016 by the
Graduate Supervisory Committee:

Hasan Davulcu, Co-Chair
Graciela Gonzalez, Co-Chair
Sharon Hsiao

ARIZONA STATE UNIVERSITY

May 2016

ABSTRACT

Proliferation of social media websites and discussion forums in the last decade has resulted in social media mining emerging as an effective mechanism to extract consumer patterns. Most research on social media and pharmacovigilance have concentrated on Adverse Drug Reaction (ADR) identification. Such methods employ a step of drug search followed by classification of the associated text as consisting an ADR or not. Although this method works efficiently for ADR classifications, if ADR evidence is present in users posts over time, drug mentions fail to capture such ADRs. It also fails to record additional user information which may provide an opportunity to perform an in-depth analysis for lifestyle habits and possible reasons for any medical problems.

Pre-market clinical trials for drugs generally do not include pregnant women, and so their effects on pregnancy outcomes are not discovered early. This thesis presents a thorough, alternative strategy for assessing the safety profiles of drugs during pregnancy by utilizing user timelines from social media. I explore the use of a variety of state-of-the-art social media mining techniques, including rule-based and machine learning techniques, to identify pregnant women, monitor their drug usage patterns, categorize their birth outcomes, and attempt to discover associations between drugs and bad birth outcomes.

The technique used models user timelines as longitudinal patient networks, which provide us with a variety of key information about pregnancy, drug usage, and post-birth reactions. I evaluate the distinct parts of the pipeline separately, validating the usefulness of each step. The approach to use user timelines in this fashion has produced very encouraging results, and can be employed for a range of other important tasks where users/patients are required to be followed over time to derive population-based measures.

To my parents, family, and friends

ACKNOWLEDGMENTS

First and foremost, I would like to thank God who has made me who I am today and for his blessings throughout my research work and my entire life.

I would like to express my deep sense of gratitude to Dr. Graciela Gonzalez, for her continued support, encouragement, and allowing me carry out my thesis and research under her. I would also like to thank Dr. Hasan Davulcu for his valuable guidance and insight throughout my Masters degree. I would also like to thank Dr. Sharon Hsiao for being a part of my thesis committee and for taking time in reviewing my thesis and providing feedback on my work.

I would also like to thank Dr. Abeed Sarker. His helpful suggestions, encouragement, and comments helped me rethink in different angles in my research. I would also wish to thank my close friend, Arjun Magge. Without his help, this thesis would be incomplete. I would like to thank all my friends and colleagues at Cognitive Information Processing Systems (CIPS) lab for making my journey of research exciting and making it a great learning experience.

Life in Tempe wouldnt have been more memorable and fun without my friends. I am extremely grateful to my parents, grandparents, and family for their love, prayers, and sacrifices.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Motivation

According to research conducted in 2008, Every year about 7.9 million infants (6% of births worldwide) are afflicted by serious birth defects (Lobo and Zhaurova, 2008), and the causes for 50% of these birth defects are unknown. While the infant mortality rate and birth complications are higher in the third world countries, statistics from Centers for Disease Control and Prevention (CDC) show that in the year 2013 in the United States alone, infant mortality rate was 5.96 deaths per 1,000 live births (mar, 2013).

Pregnancy complications comprises of health issues which could affect the baby's health or the mothers health or it could involve both. Some of the common complications are blood pressure, anxiety, and headaches. The more severe complications include preterm labor, preeclampsia, and pregnancy loss. Pregnancy period in pregnant women is the time where they are more prone to vulnerabilities and proper care needs to be taken. drug intake is a commonplace.

Although consuming medications/drugs during pregnancy is not recommended by doctors worldwide, the usage of prescription drugs during pregnancy is commonplace for various reasons. For instance, during pregnancy, women continue taking prescription drugs for ailments which preceded the pregnancy. Women also tend to take over-the-counter drugs for common health problems (like heartburn, acidity, headache, common cold and body pains) which may cause harm to the fetus. Over-the counter(OTC) drugs are the medicines are taken by the people without prescription

which are considered to be safe. However, some of these over-the-counter drugs have shown to cause adverse fatal outcomes in the past.

Past research has also indicated that 50% of the pregnancies in the United States are unintended (Finer and Henshaw, 2006). In such cases, the fetus may be exposed to drugs without the mother's explicit knowledge. For these and other reasons, it is difficult to assess how intentional or unintentional usage of medications during pregnancy may adversely affect the outcomes of childbirth, despite the vital importance of this information. Hence, it is important to maintain a vigil on the drugs consumed by pregnant women and ensure that only safe drugs are being consumed.

Pharmacovigilance

WHO(World Health Organization) (Phase *et al.*, 2004) defines Pharmacovigilance as "*the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other medicine-related problem*". Pre-market clinical trials assess the safety of drugs in limited settings, and so the effects of those drugs on particular patient groups (*e.g.*, *pregnant women*) cannot be assessed. In addition, spontaneous reporting systems that are in place for post-market surveillance, suffer from problems such as under-reporting (Harpaz *et al.*, 2012).

Social Media

As such, social media sites and online health forums that serve as sources of patient reported data are gaining popularity in pharmacovigilance research. A study from 2012 has shown that 26% of online adults discuss health information using social media (BusinessWire, 2012), with approximately 90% women using online media for healthcare information, and 60% using pregnancy related apps for support. These statistics suggest that social media sources are likely to contain key information re-

garding pregnant women, and their drug usage habits.

A very popular social network, that is currently being extensively used for public health monitoring tasks, is twitter - a microblogging site which is actively used by over 320 million users [1] . The real-time tweets by users help health monitoring services and researchers in multiple ways. For example, by tracking the first-hand reports of disease outbreaks, interested agencies can observe patterns of their spread, and take appropriate actions to minimize the effects. One advantage of twitter over other social networks is the high frequency of tweets by users, which make it easier to find drug mentions and their reactions when compared to other social media venues. Hence, twitter has been a widely used source of social media data in pharmacovigilance research (Sarker *et al.*, 2015). However, it comes with its own challenges in information extraction due to the use of abbreviations, informal language and colloquial terms.

Within the social media domain, majority of the research in pharmacovigilance has been in the areas of identification, classification and extraction of adverse drug reactions (Sarker *et al.*, 2015). In addition to pharmacovigilance, social media data has been previously used in a other public health related research such as disease surveillance and behavioral medicine research (PAUL *et al.*, 2016).

1.2 Problem Statement

My objectives in this thesis are two-fold. Firstly, I propose a model using which we can identify users of a specific characteristic from social media (in this case, pregnant women), and follow their activities on twitter, using their timelines as longitudinal networks which reveal a wide range of information about them including their drug usage patterns. Secondly, I use the collected data to assess the prevalence of use of different drugs among pregnant women, and draw the risk factor associated with each

[1] https://about.twitter.com/company

drug. I hypothesize that this approach may eventually aid in categorizing given drugs into safety classes for pregnancy: safe and unsafe.

1.3 Document Outline

The rest of the document is organized as follows. Chapter 2 gives an overview of related work. Chapter 3 introduces the system architecture. Chapter 4 details the data collection step in the architecture. In chapter 5, I introduce a new algorithm to extract pregnancy period from the user's timeline. Chapter 6 describes the steps involved in the Drug Outcome association phase. Chapter 7 discusses results and evaluations followed by conclusion and future work in chapter 8.

Chapter 2

RELATED WORK

Most of the research in pharmacovigilance has focused on identifying adverse reactions associated with medications. Some past research has attempted to employ classification techniques to determine ADR assertive posts. For these tasks, two primary techniques have been attempted: lexicon-based classification or supervised classification. In lexicon-based classifications (Nikfarjam and Gonzalez, 2011; OConnor *et al.*, 2014; Leaman *et al.*, 2010; Benton *et al.*, 2011), a given text is classified as having an ADR if it meets a set of specified lexical rules which have been derived by analyzing pre-classified texts. supervised classification techniques, (Patki *et al.*, 2014; Sarker and Gonzalez, 2015; Bian *et al.*, 2012; Jiang and Zheng, 2013) involve training a classifier using features from annotated data (used as training data) to automatically make classification decisions on test data based on observed probabilities in the training data.

Due to the advances in NLP and data science techniques, social media has recently been used for a variety of public health monitoring tasks in addition to pharmacovigilance. These include monitoring the patterns of influenza (Culotta, 2010), (Aramaki *et al.*, 2011), tracking tropical diseases like dengue fever (Gomide *et al.*, 2011), and analyzing disease outbreaks such as E. coli (Diaz-Aviles and Stewart, 2012) and ebola (Odlum, 2015). In behavioral medicine research, social media has been used to study users lifestyle and analyzing the health related choices they make. Researchers have used social media to study nutrition (Sharma and De Choudhury, 2015), obesity patterns (Mejova *et al.*, 2015; Fried *et al.*, 2014), and effects of exercises on mental health (Dos Reis and Culotta, 2015). Applications also include analyzing alcohol use

(Aphinyanaphongs *et al.*, 2014), and prescription drug abuse (Hanson *et al.*, 2013; Genes, 2014).

Only a handful of studies have attempted to predict pregnancy outcomes. Banjari *et al.* (2015) uses clustering techniques in predicting pregnancy outcomes. Their main source of data was a collection of questionnaire results accompanied by blood samples of 222 pregnant women who were at the first trimester. The authors performed hierarchical clustering considering three main features namely pre-pregnancy BMI, their age, and haemoglobin content. Using cluster analysis, the authors found that women with higher pre-pregnancy BMI and age have higher risks of complications during pregnancy.

Laopaiboon *et al.* (2014) study the effect of maternal age and pregnancy outcome. They conclude that higher the maternal age, higher are the risks of adverse pregnancy outcomes. They used health records of 308,149 singleton pregnant women admitted to various health facilities across countries. They used a multilevel, multivariate logistic regression with clustering technique to perform the study and found that 12.3% of these women had advanced maternal age (AMA) which varied across countries. They also found AMA significantly had an effect on the pregnancy outcome and increased the risks of birth complications.

Wettach *et al.* (2013) studied 202 fetal disorders from Swiss ADR database to find drug safety profiles. Using records classified by regional pharmacovigilance centers (RPVCs) as having ADRs, they performed a likelihood ratio and t-test, and found that fetal disorders were closely associated with the ADRs of drugs they consumed. We notice that all pregnancy related research have involved data sources from clinical records, reports, hospital patient data which often is expensive to obtain.

De la Cruz-Mesía and Quintana (2007) studies the effect of different β-hcg levels in pregnancy outcome. Abnormal β-hcg levels causes ectopic pregnancy, miscarriage

or spontaneous abortion. They use a bayesian classification technique to predict predictive probability of pregnancy outcome. They use vectors of β-hcg levels measured at different time period during pregnancy for the analysis of 173 women.

While the nature of the data collected in the previous cases were reported in a clinical environment, little information is available on drug usage after the patients exit the medical facilities. Hence, social media and health-forum data appear to be the best sources for extracting drug usage patterns and their effects. However, there are challenges in extracting useful and relevant information from social media data due to its lack of structure and use of informal language.

Chapter 3

SYSTEM ARCHITECTURE

Figure 3.1 gives a detailed illustration of my proposed system, which is broadly divided into three main steps.

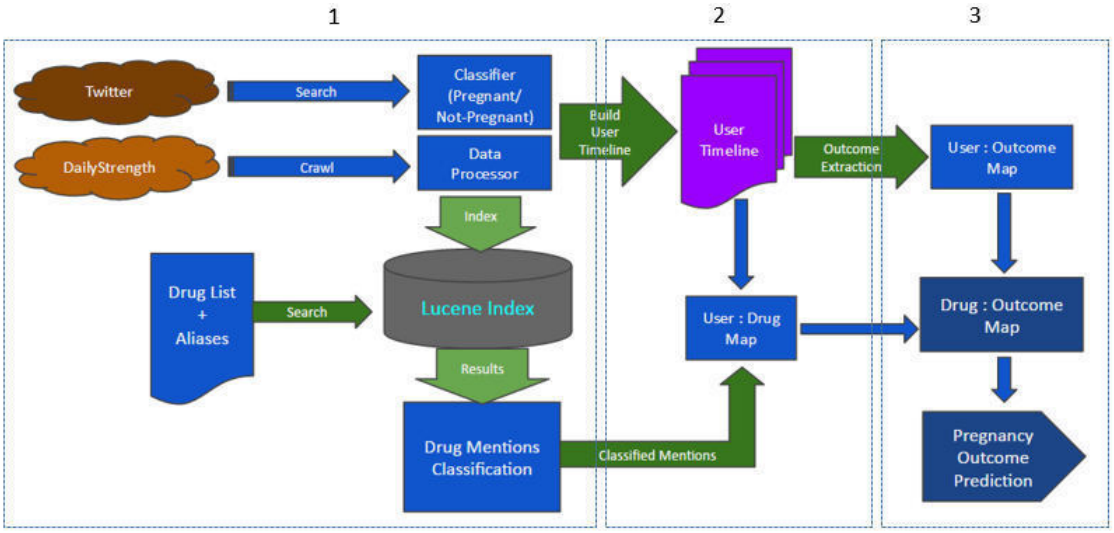


Figure 3.1: System Architecture

We categorize the pipeline into three major steps. The three major steps are as follows

1. Data Collection and Preprocessing
2. User Timeline Extraction
3. Drug Outcome Association

3.1 Data Collection and Preprocessing

In this step, tweets mentioning pregnancy announcements are first collected using twitter stream API. It is then given to a supervised classifier to extract the users mentioning legitimate pregnancy announcements. The tweets from these users are extracted using twitter search API.

3.2 User Timeline Extraction

In this step, the tweets are read to derive the pregnancy time period within the timeline and tag each tweet into individual trimesters. It is then indexed into Lucene for a faster search retrieval.

3.3 Drug Outcome Association

Here, I map each user to the pregnancy outcome. I then collect all the drugs taken by each user which helps me in associating the risk factor of each drug. I associate the drug with the outcome based on the risk factor.

Each of these steps are detailed in the subsequent chapters.

Chapter 4

DATA COLLECTION

4.1 Twitter

twitter being a popular site for sharing personal views and information is the main source of this study. The tweets which contain pregnancy announcement were collected. A total of 35,355 tweets during a one-year time-period starting from Jan 2014 to Jan 2015 were collected using twitter search API using a list of search queries. Some of the search queries used for collecting these tweets are "i am weeks pregnant lang:en since:2015-01-01 until:2015-07-31", "i am months pregnant lang:en since:2014-01-01 until:2015-01-01". For a full list of search terms, please refer APPENDIX A.

4.1.1 Tweet Preprocessing

One of the main problems in twitter data when it comes to information retrieval on twitter is that it contains an equal amount of useful information and noise in them. We minimize the noise by preprocessing the data. The steps involved in preprocessing are explained in Algorithm 1

Algorithm 1 Preprocessing Tweets

```
1: procedure PREPROCESS_TWEETS(tweet)
2:     Remove non-ASCII characters
3:     Remove URLs and user handles
4:     Perform POS Tagging and remove punctuations
5: return processed_tweet
6: end procedure
```

Many of the tweets involved include URLs for embedded content, and user handles of the people to which the tweet is directed to. Since we are more interested in the text conveyed in the message and not who it is directed to, it is not useful for this study and hence we remove them. So the first step is to get rid of all the URLs and user handles. The tweet is first subjected to tokenization. The tokenization involves both sentence and word tokenizers. Each token is then given to a regular expression matcher which checks for the existence URLs in the form of "http" and "ftp" and user handles which starts with "@" and removes them.

Part-Of-Speech(POS) tagging is a very popular technique which involves associating each word in the given sentence to its corresponding part of speech. POS tags can be very hepful in extracting features which is used for classificaiton. However due to the informal language, POS tagging is a very big challenge in twitter. I use GATE POS-Tagging model (Derczynski *et al.*, 2013) in conjunction with Stanford POS Tagger (Toutanova *et al.*, 2003) for this purpose.

4.1.2 Pregnancy Tweet classification

On observing the tweets about pregnancy announcements, we can see that not all tweets mentioned were legitimate pregnancy announcement even though they had all the keywords from the search terms. The tweets that did not have a legitimate pregnancy mention talked about the users mentioning their friend or family member being pregnant, a character from a show being pregnant, or how they looked like a pregnant women. Due to the informal language used in twitter, rule-based and lexicon-based approach have been found to not work as efficiently as supervised classification. Hence, I use a supervised learning method to eliminate the tweets which are not legitimate pregnancy announcements.

Annotation

The main requirement for a supervised classification is the training data. I along with another human annotator annotated 1200 randomly selected tweets (approximately 3% of the total tweets) mentioning pregnancy announcements into isPreg (legitimate) and notPreg (not legitimate) classes. The inter-annotator agreement(IAA) for agreement between the annotators was calculated to obtain a kappa score of 0.79 which is regarded as a substantial score as per studies by Landis and Koch (1977) and an excellent score in statistical methods proposed by Fleiss *et al.* (2013). From the manually annotated 1200 tweets, 753 tweets were classified as isPreg and 447 were classified as notPreg. Table 4.1 illustrates some of the announcement examples and its annotation.

Table 4.1: Sample Pregnancy Announcement Tweets with Annotation

Tweet	Annotation
"*I honestly still cant believe Im almost 5 months pregnant. Like wut.*"	isPreg
"*Im 18 weeks pregnant today and my 21st birthday is tomorrow. Its a good day*"	isPreg
"*I hate how bloated I get when Im on my period, like I look like Im 3 months pregnant*"	notPreg
"*I hate that I look at least 4 months pregnant every time I eat something wtf*"	notPreg

Feature Extraction

Recent advances in social media text classification show that because of the short nature of twitter posts, and the added limitations of social media text, text classification benefits from the generation of large numbers of semantically rich features. As such, my classification approach focused on generating a set of lexical, semantic, and distributional features from the training data. The features used for training the classifier are:

N-grams My first feature set consists of word n-grams of the tweets. A word n-gram is a sequence of contiguous n words in a text segment, and this feature enables us to represent a document using the union of its terms. I use 1-,2-, and 3-grams as features.

Negation pregnancy phrases One of the main disadvantages of negative word identification is it fails to identify the context. For example, consider the tweet "havent been able to eat without being nauseous for two weeks now (im not pregnant) and have had headaches regularly". In this the user is actually trying to say she is not pregnant. To avoid these cases, I used a modified version of Fang and Zhan (2015)'s algorithm to identify all negation phrases of length 2, 3, and 4 and use it as a binary feature. The algorithm is modified to cover cases like "not pregnant", "not been pregnant", and "not 3 weeks pregnant" which was not covered in the original algorithm. Negation phrases contains negation of Verbs and negation of Adjectives.

Consider the following tagged tweet for example - "havent_VBP been_VBN able_JJ to_TO eat_VB without_IN being_VBG nauseous_JJ for_IN two_CD weeks_NNS now_RB im_PRP not_RB pregnant_VB and_CC have_VBP had_VBN headaches_NNS regularly._NN ". Here, I first identify the negation word and then check if the next word

Require: Tagged Sentences, Negative Prefixes
Ensure: NOA Phrases, NOV Phrases
1: **for** every Tagged Sentences **do**
2: **for** $i/i+1$ as every word/tag pair **do**
3: **if** $i+1$ is a Negative Prefix **then**
4: **if** there is an adjective tag or a verb tag in next pair **then**
5: NOA Phrases $\leftarrow (i, i+2)$
6: NOV Phrases $\leftarrow (i, i+2)$
7: **else**
8: **if** there is an adjective tag or a verb tag in the pair after next **then**
9: NOA Phrases $\leftarrow (i, i+2, i+4)$
10: NOV Phrases $\leftarrow (i, i+2, i+4)$
11: **end if**
12: **end if**
13: **end if**
14: **end for**
15: **end for**
16: **return** NOA Phrases, NOV Phrases

Figure 4.1: Fang and Zhan (2015)'s Algorithm

is an adjective or verb. If not, I check the next word as well. In this case, since I identified NOT, I then check pregnant_VB which is a verb. So I add "not pregnant" to the negation phrase list. A binary feature of "hasNeg" and "noNeg" is calculated for each tweet.

Bots, Blogs, and Forums Noise in twitter also includes tweets which are about promotional purposes, ads from forum or blog posts, or tweets coming from bots. A bot is a computer program which is written to post automated tweets which occurs in different forms like spams, promotional links, or even in the from of a tweet by an actual user. This feature detects if the tweets come from blogs, forums, or bots using lexicon match. Some of the lexicons are "question", "forum", "inbox", "asks", and

"fan q" which resulting in a binary feature of "isBbm" and "notBbm" for each tweet.

Point of View In many cases of pregnancy announcement tweets, it is either the family member or a friend mentioning about pregnancy of his friend and not themselves. This feature extracts whether the tweet is about the first person's point of view or from someone else.

Classification

Classification is a machine learning task of assigning right labels to a given input. It is the problem of predicting a discrete random variable from another random variable. Examples of classification problems ranges from text categorization(e.g., spam filtering) to bioinformatics(e.g., classify proteins according to their function). Classification can be of two types: Supervised and Unsupervised classification. Supervised classification is one where the classifier is built on some training data. Unsupervised classification involves grouping the given data points into separate categories based on some distance or similarity measures. In other words, unsupervised classification is popularly known as clustering. In this thesis, I am using a supervised classifier to classify the tweets into "isPreg" and "notPreg" categories.

Support Vector Machine (Vapnik and Cortes, 1995) is a supervised classification technique used for both regression and classification problems. Given the training data, SVM finds an optimal hyperplane based on which it classifies new data. SVM works best for two-label(binary) classification.

It treats the training data as points in a 2D space. It tries to split the space into regions where each separate region has maximum density of points belonging to a particular label which means there is a plane which separates the space into regions and this plane is known as linearly separable planes. In fact, there are many such

planes, but SVM tries to find one such hyperplane which has the largest minimum distance from these training points. SVM can be linear and Non-linear. Non-linear SVM implies that the boundary need not be a straight line which helps in capturing more complex relations among data points.

Performance Evaluation

Since the technique used is a supervised learning, it involves both training and testing phase. In the training phase, 1200 annotated tweets are first given to the feature extraction system where the features are extracted and a feature matrix is built. SVM classifier is trained on these input feature matrix.

For the testing phase, I use a popular mechanism called Cross-Validation. Cross validation is basically a model evaluation technique where the training data is divided into k-sized equal subsets. In one iteration of cross validation, out of these k subsets, k-1 subsets are given to the classifier for training and the remaining subset is used as a testing set where the classifier assigns the label for each data in the testing set. Similarly k different iterations run with different testing set each time and the overall performance is based on the mean performance of each iteration.

Usually the performance of a classifier is mainly based on confusion matrix. It has the number of correctly and incorrectly classified entries for each binary class (in my case, "isPreg" and "notPreg"). Three main metrics Precision, Recall, and F-measure are calculated based on the confusion matrix for each iteration and the mean of all the iterations gives the effectiveness and the performance of the classifier.

I employed SVM classifier with a large set of rich features extracted from the feature extraction step and trained it on the manually annotated tweets about pregnancy announcement. Out of the 35,355 user heandles, 15130 users were classified as legitimate pregnant women.

I then collect all the available timelines of the users classified to be legitimately pregnant using the twitter streaming API.

4.2 Drug List

FDA (Food and Drug Administration)[Food *et al.* (2008)] groups pregnancy related drugs into five categories, namely Category A, B, C, D, and X.

Category A

Studies have shown that the drugs in this category are safe during the first trimester. However, there is no evidence about its safety and risks in second and third trimesters.

Category B

Studies among animals have shown that these drugs have not created any harm to the fetus but not many studies and evidence is present for pregnant women.

Category C

Studies among animals have proven that these drugs have shown adverse effect to the fetus but not many studies and evidence is present for pregnant women. However, these drugs can be consumed by pregnant women due to its potential benefits even though it has displayed adverse effects.

Category D

Studies have shown some positive evidence of adverse effects in pregnant women but due to its potential benefits, it can be used despite its harmfulness.

Category X

Studies have proven fetal abnormalities in animals and positive evidence of adverse

effects in pregnant women. The adverse effects outweigh benefits.

A total of 7396 drugs were collected across these 5 categories from three different sources [1] , [2] , [3] . I then expand this drug list by retrieving the brand names and constituent drugs of the original using RxNorm. RxNorm (Liu *et al.*, 2005) is a drug database containing drugs from various sources which has list of all available drugs as well as the relationship between drugs. Some of the relationships are "ingredient of", "has dose form", and "contains". For this study, I took the above retrieved initial set of drugs and extracted a list of other drugs which were related using the relation "has tradename" or "tradename of". For Example, "bisacodyl" is a drug in "Category B". This when searched in RxNorm had trade names "dulcolax", "bisa-plex", and "bisolax".

Due to the multiplicity of sources and further expansion of the drug list, I observed that there were few drugs that appeared across categories. To resolve the conflicts, the drugs for all pairs of drug categories were compared: AB, AC, AD, AX, BC, BD, BX, CD, CX, DX in the mentioned order. During comparison, if a duplicate is found among two categories, i remove the drug from the category with the lower severity. this technique gives the benefit of doubt to a category with a higher risk. For instance, the drug "amturnide" was present in B, C, D, and X categories. During comparison between category B and C this drug was removed from B. Similarly, CD, and DX were compared subsequently and "amturnide" was removed from C and D categories and finally placed in category X. With the above procedure a total of 7387 drugs across the five categories were obtained.

[1] http://www.tga.gov.au/prescribing-medicines-pregnancy-database

[2] http://www.empr.com/clinical-charts/drugs-used-in-pregnancy/article/125912/

[3] http://www.just.edu.jo/DIC/Manuals/Drugs contrandicated in pregnancy.pdf

4.3 Outcome List

Pregnancy outcomes fall into two categories: Good and Bad. Due to the 140 character constraint in twitter, it is difficult to identify tweets mentioning good outcomes in comparison to bad outcome mentions. After reading through various timelines of the users and the internet, separate lists of search terms for good and bad outcomes were extracted. Few examples of search terms for good outcomes are "baby healthy", "beautiful daughter born", "boy active", and "was born our baby". For bad outcomes, I extracted the list from CDC pages [4] . Some of the examples for search terms for extracting bad outcomes are "miscarriage", "stillbirth", "down syndrome", "Anotia", "Spina Bifida", and "almond shaped nose". Please refer to Appendix B for full list of good and bad outcomes search terms.

[4]http://www.cdc.gov/ncbddd/birthdefects/types.html

Chapter 5

TIMELINE EXTRACTION

From the previous step, I have the timeline of each user who have mentioned about about their pregnancy. Due to the 3200 tweet limitation from twitter, not all timelines extracted have the tweets during the pregnancy time period. So, I perform two studies, one involving the entire timeline and another which involve only the pregnancy period in the timeline in this thesis.

From the timeline tweets, I first search for the word p̈regnant:̇ For that tweet, I then get a list of n-grams (n = 3, 4, 5, and 6). I then look for that phrase in the list of n-grams for the presence of the word "pregnant" which is the last word of the string. I further look for the word "week" or "month" in that string and extract the number from the string which has the previous two indexes to the words "week" and "month".

Some of the example tweets are "oh well managed 8 out of 10 combat tracks not bad at 28 weeks pregnant with the flu but still disappointing #frustrated" which was tagged as third trimester, "im officially 20 weeks pregnant and ive also never felt more sick in my life" was tagged as second trimester. However, some of the tweets that were not handled by this algorithm are "I b getting so much pressure next week is gone b my last week pregnant who want to make a bet lol" which should be tagged as third trimester, "it is crazy to me that i am only 3 days past 13 week pregnant" which was tagged as first instead of first.

Alongside the accuracy, the main factor in testing the performance of the system is speed. Apache Lucene(Jakarta, 2004) is a very powerful and extremely fast information retrieval tool. It first creates an index of the data which we require for searching

Algorithm 2 Pregnancy Timeline Extraction

```
1: procedure EXTRACT_TIMELINE(tweet)
2:     ngamList ← ngrams with n being 3, 4, 5, and 6
3:     for str ← ngram do
4:         Words ← str.split( )
5:         if words[length-1].contains(pregnant) then
6:             if words.length == 3 then
7:                 if words[length-2].contains("week") then
8:                     daysPregnant ← getNumFromWordsOrNum(words[0]) * 7
9:                 else if words[length-2].contains("month") then
10:                    daysPregnant ← getNumFromWordsOrNum(words[0]) * 30
11:                end if
12:            else if words.length == 4 then
13:                if words[length-2].contains("week") then
14:                    daysPregnant ← getNumFromWordsOrNum(words[0]+" "+words[1]) * 7
15:                else if words[length-2].contains("month") then
16:                    daysPregnant ← getNumFromWordsOrNum(words[0]+" "+words[1]) * 30
17:                end if
18:            else if words.length == 5 then
19:                if words[length-2].contains("month") then
20:                    daysPregnant ← getNumFromWordsOrNum(words[1]+" "+words[2]) * 30
21:                else if words[length-2].contains("week") then
22:                    daysPregnant ← getNumFromWordsOrNum(words[1]+" "+words[2]) * 7
23:                    if words[1].contains("month") then
24:                        daysPregnant ← daysPregnant + getNumFromWordsOrNum(words[0]) * 30
25:                    end if
26:                end if
27:            else if words.length == 6 then
28:                if words[length-2].contains("weeks") then
29:                    daysPregnant ← getNumFromWordsOrNum(words[2]+" "+words[3]) * 7
30:                    if words[2].contains("months") then
31:                        daysPregnant ← daysPregnant + getNumFromWordsOrNum(words[0]+" "+words[1]) * 30
32:                    else if words[1].contains("months") then
33:                        daysPregnant ← daysPregnant + getNumFromWordsOrNum(words[0]) * 30
34:                    end if
35:                end if
36:            end if
37:        end if
38:    end for
39:    first_trimester_date ← tweetDate − daysPregnant
40:    second_trimester_date = first_trimester_date + 91
41:    third_trimester_date = first_trimester_date + 182
42:    trimester_end_date = tweetDate + 280
43: end procedure
```

and then uses this index in returning the results for the text-search performed. Since we are searching a huge collection of tweets, I use Apache Lucene for indexing and searching.

Chapter 6

DRUG OUTCOME ASSOCIATION

User to Outcome Mapping

To extract outcomes of pregnancy, I use a lexicon-based approach to categorize all valid cases of pregnancy into two categories: *good* and *bad*. A pregnancy is categorized as having a *bad* outcome if there is evidence of a miscarriage, stillbirth or other birth complications. For example, phrases like *"preeclampsia"*, *"neonatal death"*, *"had a miscarriage"* etc. are searched for using the lexicon. A pregnancy is classified as having a good outcome if the person's timeline shows clear hints of a healthy baby being born. For example, phrases like *"its a boy"* and *"beautiful miracle"* convey the message that at the time of birth, the newborn is healthy. With this, we try to obtain an outcome for each individual user. However, there are cases where there is no evidence of either, and we ignore such users for my study. From the initial list of 15,530 twitter users, 7172 were classified as having had a *good* pregnancy outcome and 1065 users (7.1%) were classified as having experienced a *bad* outcome.

User to Drug Mapping

I then perform a search for each drug from the categorical list to obtain the drug mentions by users. Here, an assumption is made: that all drug mentions are admissions of drug intake by the user, because of our previous classification. I query our Lucene index, and, for each drug, compute the number of users who have consumed it.

Drug to Outcome Mapping

To predict the safety quotient of drug among pregnant women with a considerable accuracy, we need to rely on various factors such as user reviews of the drug, popularity of the drug and user reports of an adverse effect as a result of a drug. In this work, we use a quantitative method for determining the approximate category of the drug based on its frequency of discussion in social media. We posit that this method in combination with other classification techniques, would help to obtain higher accuracy rates.

After determining the pregnancy outcomes of the users as *good* or *bad*, we search for the drugs consumed by the users.

For each drug across the 5 categories, we record the count of unique users who have mentioned the drug across both *good* and *bad* outcomes. Having obtained the counts for each drug, we propose a simple measure to make a data-centric estimate of the risk associated with the drug R_d, which is as follows:

$$R_d = \frac{U_g}{U_b + U_g} \tag{6.1}$$

where U_b is the fraction of users with *bad* outcome who have consumed the drug, and U_g is the fraction of users with *good* outcome who have consumed the drug. The risk R_d, which lies between 0 and 1, are calculated for all drugs mentioned by users to arrive at a linear scale which helps in predicting a drug's safety and estimate its potential category. As mentioned earlier, the official categorization of drugs vary significantly between the bodies that perform the categorizations (*e.g.*, FDA), and, in addition, for some drugs, the safety profile is simply not known. Because of this, we use the data that we have described to make our own social media-based safety estimations, and compare them to the official categorizations. To measure the *close-*

ness of our predictions to the official categories of drugs, we map the risk estimates on to a 5-point scale, from 1 for A to 5 for X. Sorting the list of drugs based on the computed risk for the drug and assigning severity to each drug from 1 to 5 based on the count of drugs in each category, we arrive at the range of values for a particular category by observing the highest and lowest value in the range. Finally, to measure the *accuracy* of our predictions relative to the official categorizations, we observe the number of correct predictions of all drugs and calculate the absolute error. We state the ranges as a statistic that can be used as reference for predicting the safety of a drug.

Chapter 7

RESULTS

7.1 Evaluation of Pregnancy announcement classification

Using the annotated data of 1200 tweets, I performed 10-fold cross validation experiments to assess the accuracy of my approach in detecting real announcements. Table 7.1 summarizes the performance results of the classifiers.

	Precision	Recall	F-measure
Naive Bayes	0.749	0.748	0.749
SVM	0.803	0.809	0.805

Table 7.1: Classification Performance

We can clearly see that SVM performed better with an F-score of 0.805 which is a significant improvement from Nave Bayes classifier (F-score: 0.749). I employed this optimized SVM classifier with a large set of rich features extracted from the feature extraction step on the unannotated data. It resulted in the discovery of 15,523 legitimate pregnant women from a total of 35,355 users.

7.2 Outcome Detection

I employed the outcome classification method on the timelines of these users, and over 30 million user posts, and categorized the different types of bad outcomes. Figure 7.1 depicts the major categories of bad outcomes on twitter. In twitter, out of the 15,523 user handles, 11982 user timelines were classified as *good* pregnancy outcome and 1048 users timeline talked about *bad* outcomes.

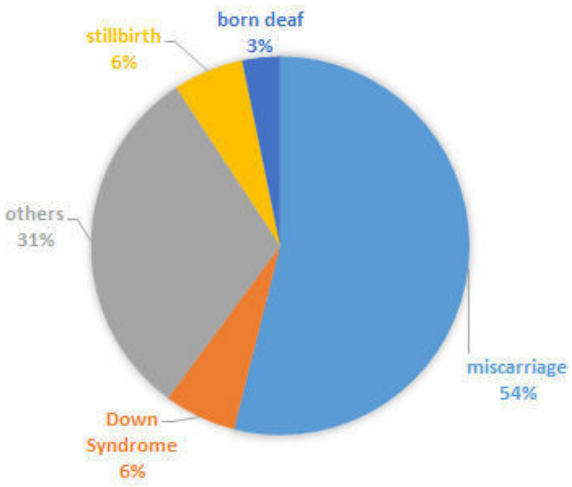


Figure 7.1: Percentages of Bad Birth Outcomes from Twitter

Tweet	Outcome
A letter to the baby I lost #miscarriage #prolife	bad
I Would have Aborted my Down Syndrome Baby	bad
So my 14 month old daughter has Spina Bifida L5 So has been holding food in her mouth for long periods of	bad
The best that happened to my life was being blessed be with child have a healthy baby boy and raise teach and watch him grow	good
My beautiful miracle	good
14 weeks and one day I cant believe its been 14 weeks since he was born This week not much has wow were gonna be weird adults	good

Table 7.2: Examples of User Tweets Presenting Good and Bad Pregnancy Outcomes

Of the bad outcomes, 699 were cases of miscarriage, 77 were cases of down syndrome, 76 were cases of stillbirth, and 439 represented other cases. The pie chart shows that by far the major reason for such a classification has been the mentions of miscarriage. Examples of tweets from the users are shown in Table 7.2.

7.3 Drug Intake Extraction

After identifying the users with the good and the bad outcome, I then performed a lexicon matching to extract the drugs taken by the users. This was done in two different setups. First, the entire timeline of the users was considered. In the second setup, tweets during the pregnancy period in the timeline was considered.

Setup 1 - Entire timeline

Summaries of the results from drug search is listed in Table 7.3. From the drug search performed for the 7396 unique drugs, 1163 drugs were mentioned by 7920 unique users across 204,775 tweets.

Drug Category	No. of Drugs (Good)	Total mentions	No. of Drugs (Bad)	Total mentions	Total Drugs
A	84	5829	110	1780	1127
B	291	11916	451	5408	3729
C	97	2127	170	3864	1050
D	43	1047	93	615	576
X	114	10076	190	3881	905
Total Users	11962		1048		

Table 7.3: Number of Drugs Across Each Category During Entire Timeline

Setup 2 - Pregnancy time period

Table 7.4 summarizes the drug search performed on the tweets during the pregnancy time period.

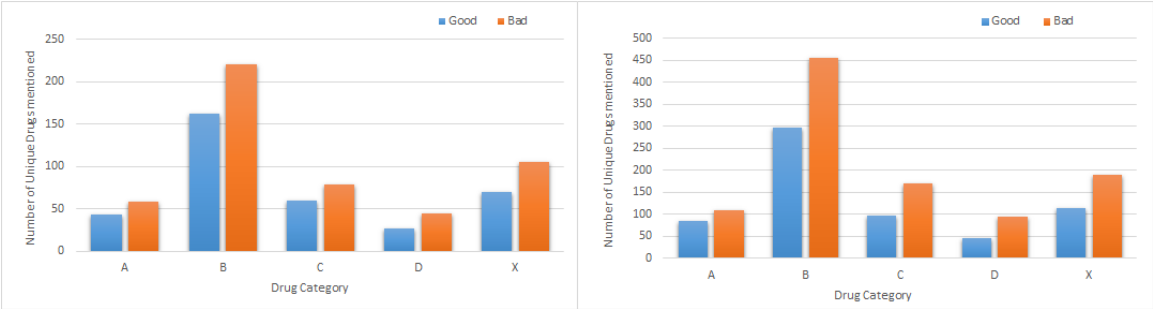
Drug Category	No. of Drugs (Good)	Total mentions	No. of Drugs (Bad)	Total mentions	Total Drugs
A	43	1216	59	466	1127
B	158	3274	216	2451	3729
C	60	490	79	598	1050
D	25	278	43	153	576
X	69	2298	105	917	905
Total Users	11962		1048		

Table 7.4: Number of Drugs Across Each Category During Pregnancy Period

There is clear evidence from both the setup that users who were classified as having had bad pregnancy outcomes are more likely to discuss a given drug than users with good outcomes. On twitter, in spite of having 11 times the population of users who were classified for having had a *bad* outcome (11962 *vs.* 1048), users with *good* outcomes discuss only about 50% of the drugs included in this study when compared to users with *bad* outcomes.

Figure 7.2 further shows that the users who had bad outcomes mention more drug intake in individual categories as well. The first graph in figure 7.2 shows the distribution of drug intake among pregnant women during the pregnancy period and the second graph in figure 7.2 gives the drug intake distribution during the entire timeline.

Figure 7.3 shows how the users mention drug intake in individual trimester. We can clearly see that Category B drug are the ones that are most mentioned which is followed by Category X. We can also see that the number of drugs intake by users is more in the third trimester and the most intakes are by the users who are classified as having had bad outcomes.



(a) Pregnancy period (b) Entire timeline

Figure 7.2: Drug Usage Across Categories Grouped by Outcomes

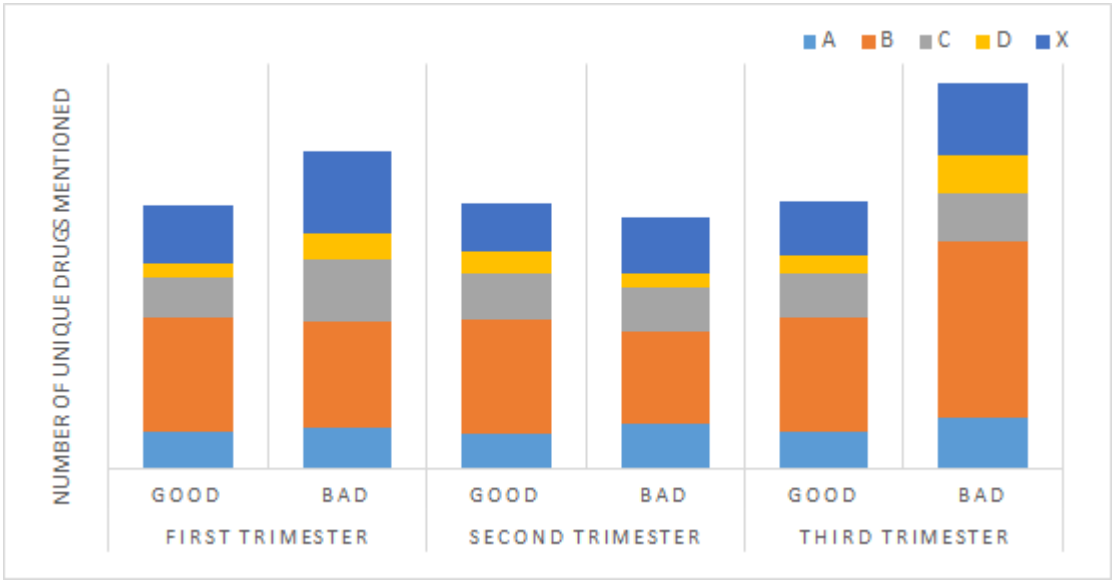


Figure 7.3: Drug Usage Across Categories in Individual Trimesters

While most of the drug search results were admissions of drug intake, there were a few ambiguous sentences which require closer look. The first mention shows an example of actual consumption, while the other two appear ambiguous. *"I took a Zyrtec this morning and I guess youre not suppose to consume more than 1 in 24hrs the struggle"*, *"Claire if it were me, I would not stop taking the progesterone. I have heard that it can be dangerous to stop it abruptly and actually can cause a*

miscarriage...", "Daily aspirin could increase chance of #pregnancy by 17" . This suggests that a classification-based approach for this task, where posts containing drug mentions are classified into personal and non-personal categories, similar to the classification for legitimate pregnancy outcomes in Step 1, would help achieve better precision. 118 drugs out of 7396(1.6%) were mentioned only during the pregnancy period.

7.4 Drug Categorization

As described in Step 3, I calculate the risk associated with each drug R_d according to Equation 6.1 and sort them in the increasing order of risk. I divide this list proportionally, according to the number of drugs in original categories. I then calculate the difference in between the actual and predicted encoding to arrive at the absolute error. The summary of absolute error for all categories are as shown in Figure 7.4. An absolute error of 0 indicates a correct prediction and it can be observed that Category B has the highest percentage of accuracy and accuracy was considerably low for Category X. I observed that of the 1165 drugs, the prediction was found to be accurate for 406 drugs (with an absolute error of 0) and 398 drugs were predicted close to their original category (with an absolute error of 1).

With the majority of the predictions (60%) within the absolute error of 1 across five categories, we posit that classification of drugs into just two categories, say safe and unsafe, would have more accuracy. We also note that by tracing an alternative sequence in Step 3 by tracking number and type of drugs taken per user, we could predict the outcome of pregnancies based on the category of drugs mentioned in their timeline. This shows that classification methods in conjunction with quantitative prediction techniques such as this can be very useful in predicting the perceived safety of drugs by monitoring social media. We observe that the risk factor R_d range

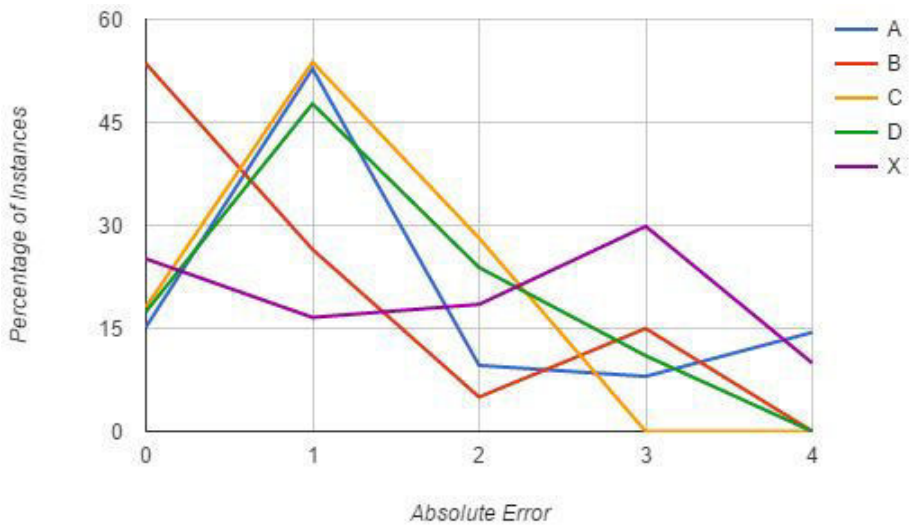


Figure 7.4: Absolute Error for Drug Category Predictions

in between 0 and .75 would be a range for most Category A and Category B drugs. For Category C, D and X we observe that most values of R_d lie in between 0.75 and 1. These ranges can be used for determining risk associated with it by calculating the proportion mentioned in Equation 6.1 and making a guess based on the ranges mentioned above.

Consider the drug *fluocinolone* which is considered safe during pregnancy and belongs to category A. The risk score of this drug was calculated to be 0.0125 and the system predicted it to belong to category A accurately. Similarly, the drug *simvastatin* is dangerous during pregnancy and is a category X drug. I obtained a risk factor of 0.964 and was correctly classified into category X. However, the commonly used drug "aspirin" belonged in the drug category X but was predicted to be in category B due its risk score of 0.586. There were users from both outcomes, good and bad, mentioning this drug in almost equal proportions and the majority of the drugs in our list were from category B. Another reason for such prediction is because people who had good outcomes were talking about *aspirin* in the form of *baby aspirin* which

can be seen in some of the tweets but the drug itself was missing from our drug list (*e.g.*, " *had the opposite prob High bp Had to take baby aspirin the while time.*")

Chapter 8

CONCLUSION AND FUTURE WORK

In this thesis, I present a novel approach to classify drugs based on in-depth analysis of user timelines on twitter. Most approaches until now involve processing of texts obtained from drug searches which ignore the users lifestyle habits and other information on their timeline which could be crucial. Thus, a timeline based approach to extract drug usage patterns aid in performing exhaustive variants of analysis, a few of which have been presented in this paper.

Since twitter has a limited set of 140 characters per post, majority of the users tend to use short forms and are bound to make spelling mistakes. Hence, I intend to expand the drug list by including misspellings, spelling variations, phonetic variations and abbreviations of each drug.

Similar to drug usage pattern extraction, disease and disorder extraction method could be used to classify mentions of diseases which would explain the reason why certain individuals consume a particular drug. I also plan to perform a Topic Modeling extension to uncover hidden properties in the text. Topic modeling may result in providing common topics that may interest the target demography and this could help in extracting additional features for classification.

The current method assumes that all drug mentions are admissions of consumption by the user. I find that although this may be true in most cases, drug mentions also contain recommendations, and cases where the user expresses reluctance to consuming a particular drug. Hence, a classification step is needed to separate first person admissions from mentions. This is currently a work in progress for which I plan to use a supervised classifier like I do in Step 1 to process admissions of consumption

only.

For outcome extraction, I currently use a lexicon based approach which has limitations due to non-adaptive rule based searches which make the classification strict. Hence I intend to add sentiment analysis features for performing a supervised classification for outcome extraction as well. I plan to include multiple features such as lifestyle habits, mentions of diseases or disorders to help such a classifier achieve higher accuracy.

We have already applied the same technique to DailyStrength data, which is an online support forum and have obtained similar results. I intend to expand our data sources to other online health and support forums in addition to other social media outlets similar to twitter to provide a common platform that can be used by pharmacists to determine the social reception of a particular drug after clinical trials.

In this thesis, I have shown a method to categorize drugs into safety classes based on the risk factor computed for each drug form social media. Although this risk factor in itself would be insufficient in assessing the risk of a drug, it can be combined with other features to assess its overall reception among consumers in pharmacovigilence. By interchanging the steps in the pipleline where we know the risk associated with the drug, we can predict the outcome given the timeline.

Although this study focuses on drug usage in pregnancy, we can use a similar approach with minimal changes by monitoring the timeline to address drug consumption by other special populations such as old-age and people suffering from particular disorders(like depression, ADHD, PTSD). For example, to find drug usage patterns in senior citizens and monitor adverse reactions, we only need to modify two components in the system. We replace the search terms to "i am 70-100 years old" in step 1 and change the outcome list to health complications.

Monitoring the timeline of users does spark a debate in online privacy versus

perceived benefits of pharmacovigilance and it needs to be addressed through public discourse. I believe that the proposed method has a great significance in pharmacovigilance in addressing drug consumption in special populations. With the world moving towards Personalized Medicine, the results of the proposed method can be combined with the clinical data to determine the right treatment for each individual.

REFERENCES

"Cdc deaths in 2013", http://www.cdc.gov/nchs/data/nvsr/nvsr64/nvsr64_02.pdf (2013).

Aphinyanaphongs, Y., B. Ray, A. Statnikov and P. Krebs, "Text classification for automatic detection of alcohol use-related tweets", in "International Workshop on Issues and Challenges in Social Computing", (2014).

Aramaki, E., S. Maskawa and M. Morita, "Twitter catches the flu: detecting influenza epidemics using twitter", in "Proceedings of the conference on empirical methods in natural language processing", pp. 1568–1576 (Association for Computational Linguistics, 2011).

Banjari, I., D. Kenjerić, K. Šolić and M. L. Mandić, "Cluster analysis as a prediction tool for pregnancy outcomes", Collegium Antropologicum **39**, 1 (2015).

Benton, A., 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).

Bian, J., U. Topaloglu and F. Yu, "Towards large-scale twitter mining for drug-related adverse events", in "Proceedings of the 2012 international workshop on Smart health and wellbeing", pp. 25–32 (ACM, 2012).

BusinessWire, "Twenty six percent of online adults discuss health information online", http://www.businesswire.com/news/home/20121120005872/en/Twenty-percent-online-adults-discuss-health-information (2012).

Culotta, A., "Towards detecting influenza epidemics by analyzing twitter messages", in "Proceedings of the first workshop on social media analytics", pp. 115–122 (ACM, 2010).

De la Cruz-Mesía, R. and F. A. Quintana, "A model-based approach to bayesian classification with applications to predicting pregnancy outcomes from longitudinal β-hcg profiles", Biostatistics **8**, 2, 228–238 (2007).

Derczynski, L., A. Ritter, S. Clark and K. Bontcheva, "Twitter part-of-speech tagging for all: Overcoming sparse and noisy data.", in "RANLP", pp. 198–206 (2013).

Diaz-Aviles, E. and A. Stewart, "Tracking twitter for epidemic intelligence: case study: Ehec/hus outbreak in germany, 2011", in "Proceedings of the 4th Annual ACM Web Science Conference", pp. 82–85 (ACM, 2012).

Dos Reis, V. L. and A. Culotta, "Using matched samples to estimate the effects of exercise on mental health from twitter", in "Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence", (2015).

Fang, X. and J. Zhan, "Sentiment analysis using product review data", Journal of Big Data **2**, 1, 1–14 (2015).

Finer, L. B. and S. K. Henshaw, "Disparities in rates of unintended pregnancy in the united states, 1994 and 2001", Perspectives on sexual and reproductive health pp. 90–96 (2006).

Fleiss, J. L., B. Levin and M. C. Paik, *Statistical methods for rates and proportions* (John Wiley & Sons, 2013).

Food, U., D. Administration *et al.*, "Content and format of labeling for human prescription drug and biological products; requirements for pregnancy and lactation labeling, 73 fed. reg. 30831-68", (2008).

Fried, D., M. Surdeanu, S. Kobourov, M. Hingle and D. Bell, "Analyzing the language of food on social media", in "Big Data (Big Data), 2014 IEEE International Conference on", pp. 778–783 (IEEE, 2014).

Genes, N., "Twitter discussions of nonmedical prescription drug use correlate with federal survey data", in "Medicine 2.0 Conference", (JMIR Publications Inc., Toronto, Canada, 2014).

Gomide, J., A. Veloso, W. Meira Jr, V. Almeida, F. Benevenuto, F. Ferraz and M. Teixeira, "Dengue surveillance based on a computational model of spatio-temporal locality of twitter", in "Proceedings of the 3rd international web science conference", p. 3 (ACM, 2011).

Hanson, C. L., B. Cannon, S. Burton and C. Giraud-Carrier, "An exploration of social circles and prescription drug abuse through twitter", Journal of medical Internet research **15**, 9, e189 (2013).

Harpaz, R., W. DuMouchel, N. H. Shah, D. Madigan, P. Ryan and C. Friedman, "Novel data-mining methodologies for adverse drug event discovery and analysis", Clinical Pharmacology & Therapeutics **91**, 6, 1010–1021 (2012).

Jakarta, A., "Apache lucene-a high-performance, full-featured text search engine library", (2004).

Jiang, K. and Y. Zheng, "Mining twitter data for potential drug effects", in "Advanced Data Mining and Applications", pp. 434–443 (Springer, 2013).

Landis, J. R. and G. G. Koch, "The measurement of observer agreement for categorical data", biometrics pp. 159–174 (1977).

Laopaiboon, M., P. Lumbiganon, N. Intarut, R. Mori, T. Ganchimeg, J. Vogel, J. Souza and A. Gülmezoglu, "Advanced maternal age and pregnancy outcomes: a multicountry assessment", BJOG: An International Journal of Obstetrics & Gynaecology **121**, s1, 49–56 (2014).

Leaman, R., 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", pp. 117–125 (Association for Computational Linguistics, 2010).

Liu, S., W. Ma, R. Moore, V. Ganesan and S. Nelson, "Rxnorm: prescription for electronic drug information exchange", IT professional **7**, 5, 17–23 (2005).

Lobo, I. and K. Zhaurova, "Birth defects: causes and statistics", Nature Education **1**, 1, 18 (2008).

Mejova, Y., H. Haddadi, A. Noulas and I. Weber, "# foodporn: Obesity patterns in culinary interactions", in "Proceedings of the 5th International Conference on Digital Health 2015", pp. 51–58 (ACM, 2015).

Nikfarjam, A. and G. H. Gonzalez, "Pattern mining for extraction of mentions of adverse drug reactions from user comments", in "AMIA Annual Symposium Proceedings", vol. 2011, p. 1019 (American Medical Informatics Association, 2011).

Odlum, M., "How twitter can support early warning systems in ebola outbreak surveillance", in "143rd APHA Annual Meeting and Exposition (October 31-November 4, 2015)", (APHA, 2015).

OConnor, K., P. Pimpalkhute, A. Nikfarjam, R. Ginn, K. L. Smith and G. Gonzalez, "Pharmacovigilance on twitter? mining tweets for adverse drug reactions", in "AMIA Annual Symposium Proceedings", vol. 2014, p. 924 (American Medical Informatics Association, 2014).

Patki, A., A. Sarker, P. Pimpalkhute, A. Nikfarjam, R. Ginn, K. OConnor, K. Smith and G. Gonzalez, "Mining adverse drug reaction signals from social media: going beyond extraction", Proceedings of BioLinkSig **2014** (2014).

PAUL, M. J., A. SARKER, J. S. BROWNSTEIN, A. NIKFARJAM, M. SCOTCH, K. L. SMITH and G. GONZALEZ, "Social media mining for public health monitoring and surveillance", in "Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing", vol. 21, p. 468 (2016).

Phase, I., I. Phase, I. Phase, I. Phase III and P. I. P.-a. S. Reporting, "Pharmacovigilance: ensuring the safe use of medicines", World Health (2004).

Sarker, A., R. Ginn, A. Nikfarjam, K. OConnor, K. Smith, S. Jayaraman, T. Upadhaya and G. Gonzalez, "Utilizing social media data for pharmacovigilance: A review", Journal of biomedical informatics **54**, 202–212 (2015).

Sarker, A. and G. Gonzalez, "Portable automatic text classification for adverse drug reaction detection via multi-corpus training", Journal of biomedical informatics **53**, 196–207 (2015).

Sharma, S. and M. De Choudhury, "Detecting and characterizing nutritional information of food and ingestion content in instagram", Proc. WWW Companion (2015).

Toutanova, K., D. Klein, C. D. Manning and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network", in "Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1", pp. 173–180 (Association for Computational Linguistics, 2003).

Vapnik, V. and C. Cortes, "Support-vector networks", Machine learning **20**, 3, 273–297 (1995).

Wettach, C., J. Thomann, C. Lambrigger-Steiner, T. Buclin, J. Desmeules and U. von Mandach, "Pharmacovigilance in pregnancy: adverse drug reactions associated with fetal disorders", Journal of perinatal medicine **41**, 3, 301–307 (2013).

APPENDIX A

TWITTER SEARCH QUERIES

i am "weeks pregnant" lang:en since:2014-01-01 until:2015-01-01
i am "months pregnant" lang:en since:2014-01-01 until:2015-01-01
im "weeks pregnant" lang:en since:2014-01-01 until:2015-01-01
im "months pregnant" lang:en since:2014-01-01 until:2015-01-01
i'm "weeks pregnant" lang:en since:2014-01-01 until:2015-01-01
i'm "months pregnant" lang:en since:2014-01-01 until:2015-01-01
i'm "weeks prego" lang:en since:2014-01-01 until:2015-01-01
i'm "months prego" lang:en since:2014-01-01 until:2015-01-01
i am "weeks prego" lang:en since:2014-01-01 until:2015-01-01
i am "months prego" lang:en since:2014-01-01 until:2015-01-01
im "weeks prego" lang:en since:2014-01-01 until:2015-01-01
im "months prego" lang:en since:2014-01-01 until:2015-01-01
i'm "weeks preggers" lang:en since:2014-01-01 until:2015-01-01
i'm "months preggers" lang:en since:2014-01-01 until:2015-01-01
i am "weeks preggers" lang:en since:2014-01-01 until:2015-01-01
i am "months preggers" lang:en since:2014-01-01 until:2015-01-01
im "weeks preggers" lang:en since:2014-01-01 until:2015-01-01
im "months preggers" lang:en since:2014-01-01 until:2015-01-01

APPENDIX B

OUTCOME LIST

Outcome	Outcome List
bad	miscarriage, miscariage, stillbirth, stilbirth, preterm birth, pre term birth, neonatal death, premature birth, low birthweight, low birth weight, Anencephaly, Anotia, Microtia, Cleft Lip, Cleft Palate, Atrial Septal Defect, Atrioventricular Septal Defect, endocardial cushion defect, atrioventricular canal defect, AV canal defect, avsd, Coarctation of Aorta, Hypoplastic Left Heart Syndrome, HLHS, aortic valve is not formed, aortic valve is very small, mitral valves is very small, mitral valves is not formed, Pulmonary Atresia, Tetralogy of Fallot, pulmonary stenosis, ventricular hypertrophy, Ventricular Septal Defect, Total Anomalous Pulmonary Venous Return, TAPVR, Supracardiac tapvr, Cardiac tapvr, infracardiac tapvr, Transposition of the Great Arteries, dtga, Conoventricular Ventricular Septal Defect, cvsd, Perimembranous Ventricular Septal Defect, pvsd, Inlet Ventricular Septal Defect, ivsd, Muscular Ventricular Septal Defect, mvsd, Tricuspid Atresia, Truncus Arteriosus, Craniosynostosis, synostosis, Sagittal synostosis, Coronal synostosis, scaphocephaly, Bicoronal synostosis, brachycephaly, anterior plagiocephaly, Lambdoid synostosis , posterior plagiocephaly, Metopic synostosis, trigonocephaly, baby Down Syndrome, boy Down Syndrome, girl Down Syndrome, baby Downs Syndrome, boy Downs Syndrome, girl Downs Syndrome, Trisomy, Hip dislocation, Hirschsprung disease, Intestinal blockage, lattened face, almond shaped nose, short neck, loose joints, exomphalos, Encephalocele, Gastroschisis, Hypospadias, Microcephaly, Omphalocele, Spina Bifida, Upper and Lower Limb Reduction Defects, Anophthalmia, microphthalmia, Common truncus, Diaphragmatic hernia, deformity, deformed limbs, Limb deficiency, Edwards syndrome, phenyhlkentonuria, pku, Rett Syndrome, muscular dystrophy, xald, born deaf, born blind, born dumb
good	babygirl, baby girl, babyboy, baby boy, babyannouncement, baby announcement, littleone arrived, little one arrived, littleprincess, little princess, itsagirl, its a girl, it is a girl, itsaboy, it is a boy, its a boy, newbaby, firsttimemom, newbornbaby, doubly blessed, Theyre Twins, they are twins, boy healthy, baby healthy, girl healthy, child healthy, son healthy, daughter healthy, son active, daughter active, boy active, baby active, girl active, child active, born healthy, born active, beautiful baby, precious baby, beautiful daughter, beautiful son, he was born, she was born, was born our baby, be a big brother, be a big sister, beautiful miracle

Table B.1: Good and Bad Outcome List