Anti-Vaccine Sentiment Classification of Tweets

By

Azin Ramezani

A thesis submitted to Macquarie University

for the degree of Master of Research

Department of Computing

October 2016

MACQUARIE
University
SYDNEY·AUSTRALIA

Typeset in LaTeX 2_ε.

Except where acknowledged in the customary manner, the material presented in this thesis is, to the best of my knowledge, original and has not been submitted in whole or part for a degree in any university.

Azin Ramezani

Acknowledgements

I would like to thank my supervisor, A/Prof. Mark Dras, for his guidance and encouragement during the preparation of this thesis.

Abstract

An important public health issue is the spread of diseases that could be prevented by changing individual beliefs and opinions about vaccination. Monitoring the spread of people's opinions through online social communities may be helpful for those public health purposes. This thesis builds on work on identifying negative sentiment about human papillomavirus (HPV) vaccines on Twitter using word n-grams and direct social connection information as features in a Support Vector Machine (SVM) classifier. This thesis examines four extensions to this. First, biological models have suggested that negative opinion is transmitted contagiously; we incorporate this by adding indirect social connection information via label propagation. Second, topic models are used to infer topics associated with tweets to reduce the feature space dimensionality in classification. Third, the content of web pages that are referenced in tweets are used as new features for classification. Finally, label propagation is extended by adding more features beyond social connection information, such as n-grams, topics, and linked web pages contents. All these extensions improve classification results to some extent, with label propagation particularly effective for tweets sent in the same time period, and topic models across longer time periods.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Twitter is a widely used online worldwide social networking service. It has more than 300 million users posting around 500 million tweets every day. Tweets are text messages with a maximum of 140 characters, and so they tend to be produced fairly rapidly. Twitter has thus been described as the largest and the most dynamic and real-time user generated dataset [1]. These characteristics make Twitter a very good source for studying public opinions and analysing sentiment.

Sentiment analysis has become important in recent years because it can be used for various purposes [2]. For example, manufacturers can examine users' opinions about their products or services, and specialists and government decision makers are able to monitor public opinion and issues and take appropriate actions in a short time [1].

There are many tasks associated with sentiment analysis: polarity classification, which is the process of classifying an opinionated piece of text as positive or negative or somewhere in between; subjectivity detection, which identifies if a piece of text is objective or subjective; determining the degree of polarity; and determining the reason for the writer's opinion polarity [2].

One of the areas in sentiment classification that has attracted many researchers in recent years is sentiment analysis on Twitter data for public health purposes, to keep track of public opinion and concerns about spread of diseases, epidemics, and different types of medications and vaccinations, and take appropriate actions in regard to this information. Traditional public health surveillance approaches, like sentinel surveillance programs or questionnaires, are difficult to implement and time consuming, cover limited number of people, and they are not real-time. As a result, an automatic real-time surveillance method is in high demand.[3]

Pro-vaccine	*As a New Year's gift to your kids vaccinate them!*
Anti-vaccine	*The risk of this vaccination are under reported and 44 confirmed deaths of young girls have been confirmed.*

Figure 1.1: Examples of Anti- and Pro- vaccine tweets for sentiment classification.

One of the important public health issues is the spread of diseases that could be prevented by changing individual beliefs and opinions about vaccination. It is essential to monitor public health opinions towards vaccination, over time and space, and evaluate the effectiveness of people's opinions on vaccination rates.[4]

In this thesis, the focus is on anti-vaccine sentiment classification in Twitter. Figure 1.1 shows two examples of opinionated tweets about vaccination. It differs from standard sentiment classification in that it deals with a different labeling scheme; and support for or opposition to vaccines might be expressed via emotive language, or quoting scientific articles, etc. We examine social connections, topics, and referenced web pages in tweets along with tweets texts for sentiment classification.

1.1 Objectives

In this thesis, we are interested in identifying which tweets express anti-vaccine sentiment. So text classification is the appropriate framework, as implemented in the work by Zhou et al. [5]. They used machine learning algorithms to identify anti-vaccine tweets about human papillomavirus vaccines (HPV), using text features and direct social connections.

[5] examined if social connections information along with tweets about HPV vaccine could be used to train classifiers. They selected 2098 HPV vaccine related tweets between October 2013 and March 2014, and labeled them as anti- or pro-vaccine. The first three months of data was used to train machine learning methods, including tweet contents (8,261 n-grams) and social connections (10,758 relationships). The second three months of data was used to test the performance of classifiers. Connection-based classifiers obtained similar results to content-based classifiers on the first three months of training data, whereas connection-based

classifiers performed more consistently than content-based classifiers in the subsequent three months. The best performing classifier was on the test data using only social connection features with accuracy 88.6%. This thesis re-implements [5] as a baseline and extends their work with the following goals:

- to study the impact of using broader social connections network as features on the sentiment classification results. The work by Zhou et al. [5] utilizes the simplest way of encoding relationships, while we will examine more complex ways of encoding relationships, such as using indirect social connections via label propagation.
- to examine the results of using topic modelling in polarity classification. Topic modeling is a kind of clustering to identify themes within a set of tweets and reduce the size of the feature space.
- to incorporate external information like contents of URL pages with sentiment classification.
- to examine the impact of extending label propagation with more features than social connections, like tweet n-grams, hashtags, emoticons, topics, URLs, and web pages content n-grams.

1.2 Findings of the Thesis

Below, we summarize the major contributions of this thesis in anti-vaccine sentiment classification on HPV vaccine tweets.

Exploitation of Social Connections in Label Propagation. In order to explore the possibility of using Twitter social connections to improve sentiment classification, label propagation with Twitter follower and source graph is implemented. We construct graphs with users as their nodes. Users are connected based on their relationships (follower or source) and they are seeded with the polarity of the tweets users posted. This model had significantly higher results in the first period using 10-fold cross validation.

Topic Modelling. Because we have a small dataset with a high dimensional feature set, it may be useful to decrease the size of the feature space in sentiment classification. Topic modelling

is a clustering approach that finds the themes that run through tweets. Using the DMM model, topic modelling will assign one topic to each tweet. These topics are used as new features in sentiment classification. Combining topics with tweet n-grams into an SVM classifier with RBF kernel slightly improves the results.

URL content classification. Tweets include associated information that is not otherwise used and could be useful for sentiment classification. Some users add URLs to their tweets to reference a web page that contains related information to their post. The idea is to extract the textual context of those web pages and exploit their information in sentiment classification. The results of using content of referenced URLs in sentiment classification are comparable to results of conventional sentiment classification.

Extending Label Propagation. We extended the label propagation with more features of tweets and features obtained in previous approaches. Graphs are created with users, tweets, hashtags, tweet n-grams, emoticons, topics, positive and negative words, URLs, and n-grams of contents of URLs. Users are connected together based on their social connections and they are connected to their tweets. Tweets are connected to n-grams, hashtags, emoticons, topics, positive and negative words, and URLs. URLs are connected to the n-grams of their contents. These graphs are seeded with tweet labels, polarity of positive and negative words, and polarity of emoticons. Results obtained from extended label propagation are comparable to the results obtained by basic sentiment classification.

1.3 Thesis Outline

The rest of this thesis is organized as follows: Chapter 2 gives an overview of the literature on sentiment classification; in particular, sentiment classification on Twitter and about public health surveillance. It also describes the baseline paper in details.

Chapter 3 explains how the baseline paper is re-implemented to use the results to compare with the results of later chapters.

Chapter 4 discusses the label propagation model using Twitter social connections. It explains how graphs are created and labels are propagated among nodes.

Chapter 5 describes topic modelling and how it is used to extract topics from tweets, as well as using topics in sentiment classification.

Chapter 6 discusses the extension of sentiment classification using distant supervision with content of web pages referenced in tweets.

Chapter 7 presents the integration of multiple features and creating an extended label propagation model. This chapter presents the experimental results of extended label propagation and compares them with results of previous chapters.

Chapter 8 summarizes the findings of the thesis and examines future extensions of the work.

Chapter 2

Literature Review

Sentiment analysis has become important in recent years because it can be used in various applications, such as recommendation systems, detection of emotions in emails or other types of communication, detection of appropriate places for ads on websites, information extraction, question answering, summarization, citation analysis, and intelligence applications [2]. For example, sentiment analysis can be used to infer interests or ratings values for users from textual reviews of items such as books, films, or products, and recommend them their preferred items.

This chapter will provide an overview of sentiment analysis and its key concepts. The focus of this thesis is on sentiment classification, one of the concepts of sentiment analysis. This chapter covers sentiment classification approaches, specifically sentiment classification on Twitter. There is a lot of work in the field, so just some selected work is described. In addition, an overview on domain specific sentiment classification, anti-vaccine sentiment on Twitter is included.

2.1 Important Concepts of Sentiment Analysis

Pang and Lee [2] give a survey on sentiment analysis; this section draws on that for its definitions of key concepts. One of the key concepts is sentiment polarity classification, which is the process of classifying an opinionated piece of text as positive or negative or locating in between. For instance in sentence, “This laptop is great”, the writer’s point of view is positive. A large portion of work in sentiment analysis has been performed in this category on contexts like online reviews and opinions.

Polarity classification is a challenging task, because a piece of text is not always opinionated

and could be objective or containing factual information. Moreover, objective documents could contain opinions. Determining whether a document is subjective or objective, known as subjectivity detection, and whether an objective piece of text contains positive or negative opinion is difficult and is one of the important concepts of sentiment analysis. As an example consider sentence, "This laptop has a long battery life". This is an objective sentence with an overall positive polarity.

Subjectivity detection is useful in extracting opinions from news, Internet forums, product reviews, and information retrieval systems. Some studies have done subjectivity detection at sentence level by considering the lexical features of adjectives in sentences like [6]. Others have carried out this task at clause or document level. In addition, genre classification has been used in some studies in order to perform subjectivity detection at document level.

Extracting information about why a reviewer's point of view is positive or negative is another concept of sentiment analysis and could be utilized in summarizing reviews. Another concept in this category is determining the degree of polarity in a document; e.g. one to five "stars" for a review, while one corresponds to the most negative and five corresponds to the most positive opinions. This can be treated as a multi-class classification problem. The last but not least concept of sentiment analysis relevant to the discussion below is considering the interactions between topics and opinions in a document, because documents could contain various opinions about different topics rather than a single opinion about one topic.

2.2 Sentiment Classification Techniques

In this section, first, I will describe traditional machine learning approaches to sentiment classification, and then I will note more recent deep learning work.

Features. One of the important parts of sentiment classification is extracting, selecting, and preparing prominent features for classification. Feature selection is a challenging task in text classification in general, because of the high dimensionality of the feature space. It is often beneficial to find an automatic approach that shrinks the feature space dimensionality, by removing non-informative terms and combining some features, without sacrificing the accuracy [7]. One of the standard approaches of extracting features is converting a piece of text into a numerical vector wherein each entry shows the presence of a word (unigram) or adjacent words (bigrams)

or their frequency in a document. Some work ([2], [8]) has found that the binary valued feature vector typically outperforms vectors with words occurrence frequency in polarity classification.

In addition, part-of-speech (POS) information is often used for selecting features for sentiment analysis. These features could be used for both polarity classification and subjectivity detection [2]. [8] combined POS tags of words with unigrams as features for sentiment classification, because words with various parts of speech can have different sentiment levels. For example, the word "love" in the sentence "I love this movie" has been used as a verb and has positive polarity, while in the sentence "This is a love story" it is a noun and from a sentiment point of view is neutral.

There are interactions between topics and sentiment which could make them important features for opinion mining tasks [2]. A topic is considered as a subject that a document is focused on and is different from the topics in topic modelling, discussed bellow. Using topics as features for sentiment classification can reduce dimensionality of the feature space. Sentiments depend on topics or domains. Riloff et al. [9] combined subjectivity classification with topic classification. First, they implemented a topic-based classifier to eliminate irrelevant texts, and then ran a subjectivity classifier on top of that. [10] proposed an unsupervised approach to identify sentiment and topic at one stage out of an unlabeled set of documents. [11] studied sentiment classification on financial news using news topics.

Sentiment Classification. One major way of tackling sentiment classification tasks is via conventional text classification algorithms. However, there are some contrasts between sentiment classification and other kinds of text classification. In the text categorization task there can be many categories while in sentiment classification there are only a few classes generalized among many domains and users, like positive or negative. Furthermore, in sentiment classification classes are related and opinions have a regression-like nature. In addition, features of opinion-oriented texts are different from fact-based text. [2]

Supervised Sentiment Classification. Much work has been conducted on supervised sentiment classification with a labeled set of data as training set. There are some early works which established approaches that are now standard. [8] employed three machine learning algorithms, Naive Bayes, Maximum Entropy classifier, and Support Vector Machines (SVM) in order to classify opinions on movie reviews domain, and got the best accuracy by SVM around 82%, against baselines ranging from 50% to 69%. [12] tried SVM for product reviews polarity classification

using feature extraction and scoring techniques, with highest accuracy of 85%, higher than the 84.6% accuracy of bigrams (baseline). In [13], sentiment classification was implemented on online travel destinations reviews, using supervised machine learning algorithms; Naive Bayes, SVM, and character based N-gram model, and all three approaches reached accuracies of at least 80%. [14] proposed a supervised sentiment classification with linear SVM using senses of words, from WordNet, as set of features, and had the best performance with an accuracy of 90.2%, against a baseline of 84.90%.

There are some recent works on supervised sentiment classification. Socher et al. [15] introduced the "Recursive Neural Tensor Network", which is trained on the "Stanford Sentiment Treebank" and does not use standard features as described above, but instead "word embeddings", low-dimensional vector representations. This model is able to capture sentiment of long phrases and sentences using labelled parse trees. The highest accuracy of this model is 87.6%, that outperforms the baseline traditional supervised sentiment classification with improvement of 9.7%. [16] proposed a supervised term weighting model for sentiment analysis, based on the importance of a term in a document (ITD) and importance of a term sentiment (ITS). This approach showed that term weighting schemes based on supervised learning performs better than schemes originated from information retrieval without considering the correlation between terms and sentiment polarity.

Semi-Supervised Sentiment Classification. Semi-supervised approaches are useful when the number of labelled data is much less than unlabeled data. Work by Li et al. [17] was on sentiment classification of imbalanced data, such that negative samples are the minority class and positives are the majority with an imbalanced ratio in the training dataset. Firstly, in order to produce balanced training data they applied under-sampling on their imbalanced training set. Then they proposed a novel iterative semi-supervised classification approach by creating random sub-spaces with less feature space dimensions than the original dataset. Subsequently, a deep learning semi-supervised method, called ADN, was proposed by [18]. ADN utilized embedding information of the unlabeled data and selected a number of training data to be labelled manually. Then the ADN architecture was trained by the labelled and unlabeled data. [19] proposed a two-step sentiment classification semi-supervised algorithm called "Fuzzy Deep Belief Network (FDBN)". In the first step, the general deep belief network was trained on the labelled data using semi-supervised learning algorithm and created a fuzzy method based on that. Using the fuzzy

knowledge of the first step, performance of semi-supervised sentiment classifier improved.

One principled approach to implementing a semi-supervised approach, label propagation, has been adapted from graph theory to sentiment analysis. Label propagation algorithms propagate labels from a small set of labeled data to a large group of unlabeled data, until a global stable stage is achieved. In the label propagation process, data is represented as a graph in which nodes are data features and weights of edges are the pairwise distances (closeness, equality, or similarity) between the features. Labels are propagated on the graph, from labeled nodes to unlabeled ones [20]. [21] utilized a semi-supervised label propagation method for polarity detection. Work done by [22] was on document-level sentiment classification on resource-scarce languages, using graph-based label propagation with phrases as nodes and their similarities as edge weights.

Unsupervised Sentiment Classification. In the case that no labeled data is available, unsupervised learning approaches play an important role. [23] proposed a deep learning unsupervised feature extraction approach to improve the performance of sentiment classification, which was applicable for large scale data. [24] presented an unsupervised framework for sentiment classification of reviews based on modelling two emotional signals: emotion indication and emotion correlation. This approach was beneficial because traditional lexicon-based unsupervised algorithms did not have good performance on unstructured and informal pieces of texts like reviews. The work by [25] was an unsupervised sentiment analysis on online social websites data and was applicable in subjectivity detection and polarity classification contexts.

A general approach to finding useful hidden structure in unlabeled data is topic modeling, and this has been applied to sentiment analysis in several pieces of work. Topic modelling algorithms are typically unsupervised statistical methods that determine the thematic topics that run through documents. A topic is defined as a distribution over a fixed vocabulary [26]. Figure 2.1 shows a part of an article about using data analysis to determine the number of genes. Some of the topic words are highlighted in the article, such as “computer” and “prediction” for *data analysis*, and “sequenced” and “genes” for *genetics*. [27] utilized a latent topic model of review texts for document level sentiment analysis. [28] utilized topic modeling in conjunction with sentiment unification model for sentiment analysis on reviews of electronic devices and restaurants. [29] proposed a dynamic joint sentiment-topic model to detect and track opinions from online social media. In all these cases, including topic models improved results.

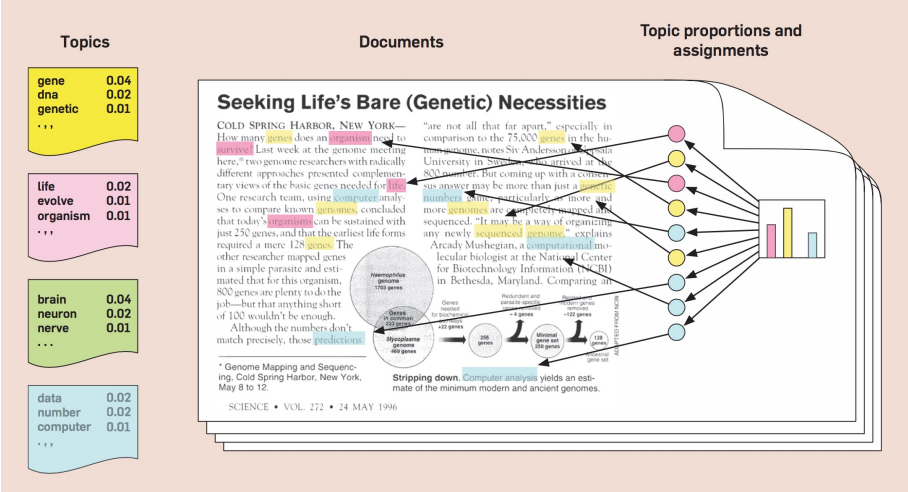


Figure 2.1: Topics and topic assignments in a document (adapted from Fig.1 in [26])

2.3 Sentiment Classification on Tweets

The genre of documents influences the results of sentiment analysis. The genre of this project is Twitter and there is a wide range of studies about sentiment classification on Twitter. Many companies and manufacturers want to study people's reviews on online websites. Twitter is one of the most popular online micro-blogs that enables users to send and read short 140-character messages called "tweets". Users can post tweets, read other users' posts and reply to them, or they can re-post or forward a tweet posted by another user, known as a "re-tweet". Tweets often contain a large amount of opinionated information. Tweets are also different from other reviews and texts, as they are informal and short (140 characters). Sentiment analysis on Twitter data has attracted many researchers in recent years.

[30] built a supervised sentiment classification model on Twitter data to classify tweets as positive, negative, or neutral. This was done with three models; unigram model as baseline, feature based model, and tree kernel model. In the feature-based model, 100 features like POS tags and prior polarity of words were used, while in the tree kernel approach, a tree representation of tweets was utilized. The tree kernel base model outperformed the other two models, with 73% accuracy, while accuracy of the unigram baseline and the feature-based model were 71%. In addition, [30] combined those models together and implemented sentiment classification with combination of unigrams and features selected in feature based model, and combination of features with tree kernels.The best performing model was the combination of unigrams and sentiment features with 75% accuracy. The work by Chikersal et al. [31] was a tweet level

sentiment classification model by exploiting supervised learning and applying linguistic rules and sentic computing resources. Their best result is with f1-score of 88.25%.

[32] proposed a target-dependent and context-aware sentiment classification approach on Twitter. A target is a user's query to look for tweets that contain positive or negative sentiment about that target and classify their sentiment based on the target. For example, the tweet "*Windows 7 is much better than Vista!*", contains positive sentiment about *Windows 7* target, while contains negative sentiment about *Vista*. The proposed approach used syntactic features that were dependent on the target, in order to prevent returning tweets that were irrelevant to the target. In addition, context-aware approach refers to considering related tweets rather than considering isolated tweets in sentiment classification. Related tweets are tweets published by a user, tweets that reply another tweet, and re-tweets. Considering related tweets in this approach enhanced the classification performance.

Because there are usually not sufficient labelled tweets, [33] presented a sentiment classification approach using distant supervision. In this model, instead of training classifiers (SVM, Naive Bayes, and Maximum Entropy) with manually labelled tweets, they used tweets with emoticons as noisy labels for training set and unigrams, bigrams, and POS tags as features. [1] proposed a semi-supervised sentiment classification model using graph-based label propagation on Twitter data with diverse set of features. [34] performed target-dependent sentiment classification on Twitter creating an "Adaptive Recursive Neural Network (AdaRNN)"model. [35] proposed an unsupervised sentiment classification method on Twitter data by expanding concepts expressed in tweets.

There are some studies about sentiment analysis on Twitter using topic modelling. Si et al. [36] used topic modelling in order to perform sentiment analysis on Twitter to predict the stock market, by calculating topics every day. Then, they regressed the stock index and sentiment time series from daily topic sets and their derived words distribution to predict the market. [37] proposed an unsupervised approach for sentiment analysis on Chinese online social reviews using an LDA model. Xiang and Zhou [38] built a method for sentiment analysis on Twitter using topic modelling. First, they built a sentiment classifier SVM with linear kernel. They trained and tested the classifier with different set of features, such as N-grams, lexicons annotated as positive or negative, emoticons, last sentiment word, etc. Then, they applied Latent Dirichlet Allocation (LDA) to tweets to identify their topics. Finally, data was clustered based on the topics distribution and used to train the sentiment classifier and integrated to the previous model.

Pro-vaccine	*As a New Year's gift to your kids vaccinate them!*
Anti-vaccine	*The risk of this vaccination are under reported and 44 confirmed deaths of young girls have been confirmed.*

Figure 2.2: Examples of Anti- and Pro- vaccine tweets for sentiment classification.

[39] proposed two sentiment topic models in order to discover latent topics and associate them with social emotion classification.

2.4 Sentiment Classification for Public Health

In recent years, sentiment classification on Twitter data for public health purposes has attracted many researchers. The goal here is to keep track of public opinions and concerns about the spread of diseases and epidemics, or about different types of medications and vaccinations, and take appropriate actions in regard to this information. Traditionally, public health surveillance approaches included sentinel surveillance programs, surveys of people all around the world, questionnaires and clinical tests, and laboratory-based examinations. However, traditional surveillance systems are difficult to implement and time consuming, cover limited numbers of people, and they are not real-time. As a result, an automatic real-time surveillance method is required to monitor public opinions and enable specialists and government decision makers to take immediate actions. Twitter is a good source of up-to-date opinionated data with millions of tweets posted everyday, for public surveillance purposes. [3]

[3] proposed a sentiment classification approach to measure the "Degree Of Concern (DOC)" of Twitter users about the impact of diseases. It was an automatic two-step sentiment classification which could identify negative and non-negative tweets about diseases. The first step was identifying personal tweets from news tweets, followed by classifying personal tweets as negative and neutral. In addition, a system called "Epidemic Sentiment Monitoring System (ESMS)" was presented by [3] to visualize individual's concerns about diseases and provide the opportunity to monitor progression and peaks of public concerns.

One important public health issue is the spread of diseases that could be prevented by

changing individual beliefs and opinions about vaccination. It is essential to monitor public health opinions towards vaccination, over time and space, and evaluate the effectiveness of people's opinions on vaccination rates. Twitter, as a kind of social media, is a good source of real-time data for measuring public health trends about vaccination. Figure 2.2 (repeated from chapter 1) shows two opinionated examples of vaccine tweets.[4]

[4] used tweets from August 2009 to January 2010 in United States, when pandemic influenza A (H1N1) was epidemic and its vaccine became available later that year. Not only tweets with related keywords were selected for this study, but also information on users and their followers were collected to identify the network of information flow. Some of the tweets were manually labeled as positive, negative, and neutral to be utilized for training a machine learning classifier. Labels of the remaining tweets were predicted automatically by the classifier and a network of information flow was created to study sentiment distribution and its effects on individual vaccination decisions. [4] found clusters of users with positive or negative opinions in Twitter. These clusters can strongly affect disease outbreak risks.

[40] studied user communities in Twitter to analyze users' anti-vaccine opinions and behaviors in a community, zone, or country and in particular how these opinions propagate. Their work focused on re-tweet graphs using Community Detection Algorithms and studying user interactions about vaccination and their social influence. The results obtained by [40] reveals that opinions spread through user communities in a similar manner to how viruses propagate and can affect people's decisions about vaccination.

2.4.1 Foundation Work: Sentiment Classification on Vaccination Tweets

The work in this thesis follows up specifically on the work done by Zhou et al. [5], on examining the impact of using social connection information as features to improve sentiment classification on identifying anti-vaccine rhetoric about human papillomavirus (HPV) vaccines. So, this work is described in some detail. Using social connections (followers and sources) as features has been found useful in general tweet sentiment classification, as mentioned earlier in section 2.3 about related tweets. [5] noted this could be particularly true in public health contexts due to the following three characteristics:

1. "Homophily", which means people with similar opinions make connections together.

2. “Contagion of Opinions”, which means social connections form a channel to flow information and opinions.

3. “Temporal Dynamics”, which means social connections change slower than tweet contents.

In order to provide data, [5] collected 42,533 English-language tweets from October 1, 2013 to March 31, 2014 containing HPV vaccine related keywords, such as HPV, vaccine, Gardasil, Cervarix, vaccination, cervical, and cancer. In addition, for each of the 21,166 tweet users, the set of users they follow (termed SOURCES) and set of users who follow them (termed FOLLOWERS) were collected. [5] randomly selected 1050 tweets from the first three months as training data and 1100 tweets from next three months as a test set. Selected tweets were labelled manually as anti-vaccine or otherwise. After deleting identical tweets or tweets with suspended users, the final dataset contained 884 tweets for training with 247 (28%) anti-vaccine labels and 907 tweets for test set with 201 (22%) anti-vaccine labels.

After identifying training and test set, data was pre-processed by removing punctuation, stop words, and non-word elements (like URLs) from tweets. Then data was converted to a set of unigrams and bigrams as features for classification. Finally, direct follower or source relationships were added as features for classification.

[5] chose a support vector machine (SVM) with a radial basis function (RBF) kernel. In order to improve the classification performance and removing features that are not useful, a hybrid feature selection method of forward selection and backward elimination was utilized.

[5] applied a feature selection approach, which is not a standard feature selection approach in NLP. By applying Fisher’s exact tests and Bonferroni corrections, [5] identified that only 1 feature of 24 important features in the first 3 months was still significant in the second period, while 73 features out of 73 connection-based follower features and 80 features out of 82 connection-based source features were still significant in the second period. They argued that these results show that social relations, in contrast to content features, are still significant in different time periods.

The classifier was trained with different sets of features on the first period and was tested on first period (using 10-fold cross validation) and second period data. The best classification results on training set were from a combination of content and connection features with 94.4% accuracy, using 23 social connections and 28 content features. Accuracy of classification with connection-based source features was slightly higher than accuracy of classification with follower connection information.

Classifier testing in second period showed different results than classifier testing in first period. The classification accuracy with content features decreased in second period more than 30%, while classifiers with connection-based set of features had similar results in both periods. On the basis of these results [5] argued that social connections have more consistent distribution than text-based contents.

2.4.2 Critique and Possible Extensions

There are some possible extensions to the work done by [5]. They used only a limited set of features, such as unigrams and bigrams and direct sources and followers, while tweets contain more hidden information that could improve classification accuracy. In the current project, [5]'s approach is extended by trying different classifiers and adding some more features.

As a baseline for polarity classification, positive and negative words frequencies are calculated and used as features, like the work done by [41]. [41] counted positive and negative words of tweets based on the subjectivity lexicon from OpinionFinder. Also, I combine positive and negative words frequencies with unigrams and bigrams in classification.

[5] utilized direct sources and followers while opinions could flow among the network of users that are not directly connected to each other, as noted in the work by [1]. In addition to trying connection-based social information of sources and followers directly, I try indirect relationships between users in a graph-based label propagation approach motivated by the work done by [1].

Speriosu et al. [1] exploited the Twitter followers graph to boost sentiment classification results, because tweets are not isolated and they are connected to each other based on their writers and their writers' followers, in a graph. In label propagation algorithms, labels are distributed from a small set of labeled nodes to unlabeled ones in the graph, as noted in section 2.2. Speriosu et al. extended the graph by adding more features to it. They created a graph with users, tweets, unigrams and bigrams, hashtags, and emoticons as nodes. Users were connected to their tweets from one side and to their followers from another side. Tweets were connected to the unigrams and bigrams, hashtags, and emoticons they contain. Initial seeds of the graph came from the polarity values in OpinionFinder lexicon, emoticon polarities, and labels assigned to tweets from a maximum entropy classifier. This can be adapted to the problem that we are working on.

Also noted in section 2.2, was the potential usefulness of topic models in reducing the feature space dimensionality. This is particular suitable here, as [42] suggests that there might be useful

themes that distinguish anti-vaccine from pro-vaccine tweets. Surian et al. [42] found that the alignment between topic modelling and community structure detection can be a useful way to characterize Twitter communities for public health surveillance.

Chapter 3

Re-Implementation

This chapter explains how the the work done by Zhou et al. [5], outlined in section 2.4.1 of Literature Review, is broadly re-implemented as a basis for extensions in later chapters. In addition, it is explained how some of their non-standard approaches are replaced by standard ones, expecting to find comparable results.

3.1 Study Data

The dataset[1] is a collection of English-language tweets containing human papillomavirus (HPV) vaccine keywords, collected by [5] between October 1, 2013 and March 31, 2014. Selected tweets were divided into two contiguous three-month periods and were randomly sampled for classification. Tweets were labeled by five labels; "A" as "Anti-vaccine", "P" as "Positive", "AP" as "Anti-Positive", "AA" as "Anti-Anti-Vaccine", and "N" as "Neutral". In the current project, all A and AP tweets are labeled as anti-vaccine and all others are considered as pro-vaccine, as in [5]. The final data set contains 1018 records as training set from the first three month period, with 273 (27%) tweets labelled as anti-vaccine, and 1080 records as test set from the second three month period, with 227 (21%) anti-vaccine tweets. Figure 3.1 shows two tweets, one labeled as pro-vaccine and one labeled as anti-vaccine.

In addition to the tweets dataset, for each of the users responsible for tweets there are two associated files; one contains all the users they follow (sources) and one contains all the users follow them (followers). This information is used to extract social connection information. Figure 3.1 shows social network of two users, with their followers and sources. In the figure, for

[1]We thank the authors for access to this dataset.

example, both User A and User B follow User 3, so User 3 is a source for A and B; User 7 is a follower of User B and User 5 is the follower of both Users A and B.

3.2 Data Preparation

In order to prepare data for sentiment classification, tweets need to be pre-processed, following broadly the same process as [5]. First of all, URLs are removed from tweets. Secondly, all non-word elements like punctuation and numbers are removed and all tweets are converted to lowercase. Then, words that essentially have no sentiment or denotational meaning (called stop words) are excluded from tweets. To do this, tweet words are compared with a list of default English stopwords[2] to find stop words in the dataset. Finally, all the words in tweets are lemmatized to remove inflectional endings and to return the base or dictionary form of a word, using NLTK WordNetLemmatizer[3]. Figure 3.1 shows how tweets change after applying the data preparation process.

Tweet texts are then converted into unigrams and bigrams. The unigram model creates a sparse vector for each sentence with the length equal to the size of the vocabulary, representing the presence or count of each word in a tweet. Bigram is same as unigram, except for considering the presence or count of pairs of adjacent content words in a tweet.

[5] did not give a baseline to compare their results with, but a baseline is important for context, especially given the existing skewed dataset. This chapter uses two baselines. The first is a standard majority class baseline. The second is a basic measure of positive versus negative sentiment: a reasonable first guess is that anti-vaccine tweets might have more negative words, as in the anti-vaccine example in figure 3.1 the negative word “Dead” is used. In this second baseline, the total number of positive and negative words in a tweet is calculated. The tweet is labeled as pro-vaccine if it has more positive words than negative ones, otherwise it is labeled as anti-vaccine. In this thesis, the OpinionFinder subjectivity lexicon[4], which contains 2,304 positive and 4,153 negative words, is used to specify positive and negative words in tweets. In the OpinionFinder subjectivity lexicon, words are in stem form. As a result, tweet words are stemmed and then compared with OpinionFinder lexicon to calculate the total number of positive and negative words in a tweet. If the number of positive and negative words are equal

[2]http://www.ranks.nl/stopwords

[3]http://www.nltk.org/api/nltk.stem.html

[4]lexicon http://mpqa.cs.pitt.edu/opinionfinder

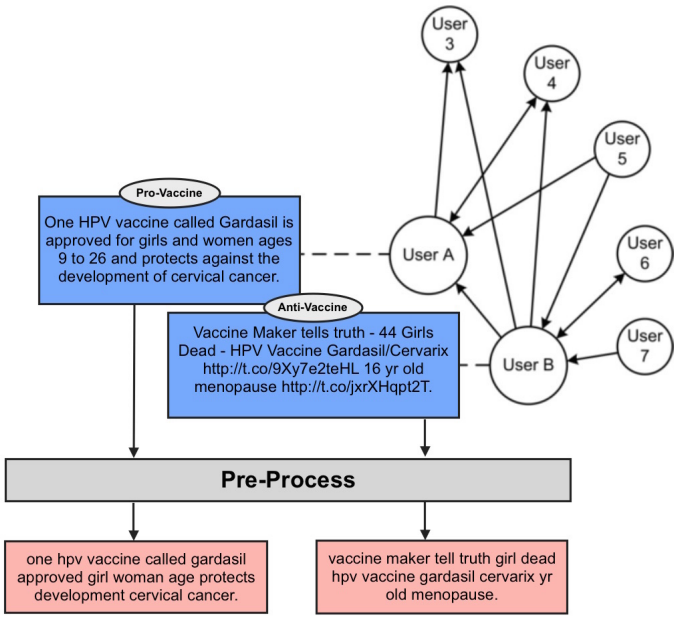


Figure 3.1: The original tweet texts are pre-processed (left). The social network for the two users posting the tweets is decomposed into source and follower features (right) (Adapted from figure 1 in [5])

in a tweet, the label would be selected uniformly randomly.

Finally, direct social connections among users are used as features for sentiment classification, as in [5]. In order to prepare social connection features, first direct source and follower relationships for all of the 1725 distinct users in the dataset are determined. Then, for each user social relationships are translated to two separate binary vectors with size of 1725 (total number of users), one for sources and one for followers. Let the follower binary vector for user i be $\{f_{i1}, \ldots, f_{iN}\}$, where $N = 1725$ is the total number of users. If user j follows user i, in the follower vector f_{ij} is 1, otherwise it is 0. In the dataset metadata, each tweet is associated with a user, who is the author of that post. Those binary vectors are added to the datasets as new features, using user identifiers associated with tweets. The first and second period datasets will thus have 3450 extra features in addition to the content unigrams and bigrams, 1725 for followers and 1725 for sources.

3.3 Classification Methods

Zhou et al. [5] constructed their classifier using Support Vector Machines (SVM) with radial basis function (RBF) kernels. In this chapter, in addition to SVM with RBF kernels, SVM with linear kernel and logistic regression are also implemented.

In supervised learning, training data consists of n tuples $(x_{i,1}, ..., x_{i,m}, y_i)$, where $x_{i,1}, ..., x_{i,m}$ are predictor variables (known as features) and y_i is the experimental or predicted variable (known as label). The purpose is to find a mapping between features and labels that fit the data (training and test) as correctly as possible.[43]

Logistic Regression Logistic regression predicts the probability that label Y belongs to a particular category C (either anti-vaccine or pro-vaccine), given a set of predictors X:

$$P(Y = C|X) = \frac{exp(\beta_1 X + \beta_0)}{1 + exp(\beta_1 X + \beta_0)}, \tag{3.1}$$

where β_0 and β_1 are coefficients estimated based on the available training data, using *maximum likelihood method*. The *likelihood function* is formulated as follows:

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y'_i=0} (1 - p(x'_i)), \tag{3.2}$$

where $p(x_i)$ is the predicted probability of each individual in training set. For logistic regression, the values $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to maximize the likelihood function.[43]

Support Vector Machines The idea of SVMs is to find a hyperplane vector in a way that the classes are well separated and the margins are as large as possible. SVM characterizes the separating hyperplane in terms of inner products of observations. The inner product of two observations $x_i, x_{i'}$ is:

$$< x_i, x_{i'} >= \sum_{j=1}^{p} x_{ij} x_{i'j}, \tag{3.3}$$

where p is the number of predictors.

The linear SVM classifier is in the form of:

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i < x, x_i >, \tag{3.4}$$

where there are n α_i parameters, one per training observation and α_i is non-zero iff the observation is a support vector.

Kernel $K(x_i, x_{i'})$ is a generalization of the inner product. K is the kernel function that quantifies the similarity of two observations. Radial kernel is in the form of:

$$K(x_i, x_{i'}) = exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2), \tag{3.5}$$

where γ is a positive constant. [43]

Because the current dataset is small and has a relatively large feature space, feature selection could prove beneficial. [5] used non-standard feature selection approaches. Applying Fisher's exact tests and Bonferroni corrections they found which individual features are significant. However, other features which may not be individually significant may still be useful in classification, in combination with others. Instead in this chapter, a more standard feature selection approach, the chi-squared statistic (χ^2), is used to reduce the size of the feature space and the impact of feature selection on sentiment classification accuracy is examined. In addition, in selecting unigrams and bigrams only words with frequency more than 15 are included.

Measures In order to evaluate sentiment classification algorithms, f1-score is calculated. F-score is weighted average of precision and recall:

$$f1 - score = \frac{2.Precision.Recall}{Precision + Recall}, \tag{3.6}$$

where precision is the number of true positives (number of tweets correctly labeled as pro-vaccine) divided by the total number of tweets predicted as pro-vaccine, and recall is the number of true positives divided by the total number of actual pro-vaccine tweets. In this thesis, accuracy (the total number of correct predictions divided by total number of predictions) is not used to measure the performance of classifiers, because the existing dataset is imbalanced and accuracy is a bit misleading on imbalanced datasets. F1-score suggests a better model than accuracy that conveys a balance between precision and recall. F1-score is a number between 0 and 1, and it reaches the best value at 1 and worst at 0. [44]

3.4 Results

Feature Significance. [5] did an analysis of which features continue to be significant across

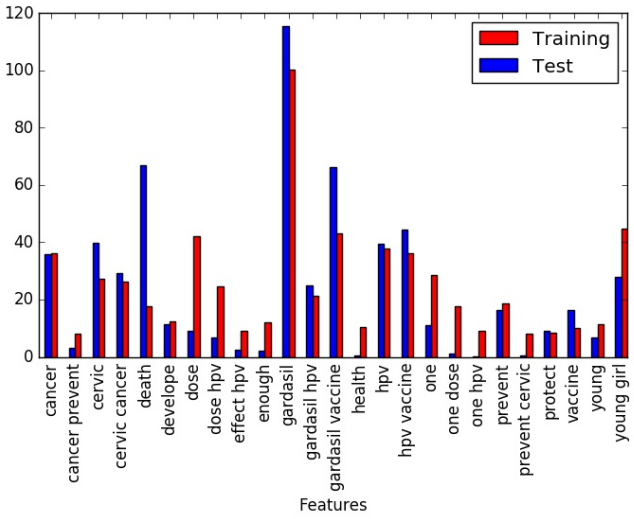


Figure 3.2: Scores of features with p-value less than 0.05 with χ^2 features selection, from the first three months (red) and second three month (blue)

time periods; we do the same here. 50 features are selected by χ^2, training in first period, for all features p-values are less than 0.05. Among them only 36 features still have p-value less than 0.05 in second period. The results show that some text-based features that are significant in first three months period are not any more significant in the second one. Figure 3.2 compares scores of some features in training and test set. All these features have p-values less than 0.05 in both periods. For example, the word "dose" is more significant in first period than the second one, while the word "gardasil" has higher score in second period. This supports the contention in [5] that important content features do change across time periods, but we note that it can still be useful to use non-significant ones as features.

Classifiers training and testing in first period As in [5], classifiers are constructed and tested in the training period using ten-fold cross validation and results are shown in table 3.1. In this period, the best result obtained by Zhou et al. [5] with combined bigrams, followers, and sources set of features, is used as the main comparison. All classifiers perform essentially the same, and the best result here is broadly the same as the best result in [5]. The classifiers constructed with connection-based source features perform slightly better than the classifiers constructed with connection-based follower features. As the results show, feature selection via χ^2 does not help.

Features	**SVM(Linear)**	**SVM(RBF)**	**Log. Regression**
N-grams + Followers + Sources ([5])	N/A	0.89	N/A
Baseline - Majority Class	0.845	0.845	0.845
Baseline - Opinion Finder (OF)	0.841	0.841	0.841
N-grams	0.857	0.847	0.868
N-grams + χ^2	0.852	0.845	0.864
N-grams + OF	0.857	0.84	0.867
Followers	0.891	0.842	0.889
N-grams + Followers	0.898	0.902	**0.903**
Sources	0.899	0.891	0.894
N-grams + Sources	0.900	0.902	0.901

Table 3.1: The performance of classifiers (f1-score) trained and tested on first three months period using 10-fold cross validation

Classifiers testing in second period Classifiers achieved slightly better f1-scores in second period and it could be because the dataset is more skewed. Table 3.2 contains the complete results of sentiment classification tested on second three months period data. All the classifiers with the total number of positive and negative words (Opinion Finder) had lowest f1-scores (0.883) and equal to majority class baseline results. Again, performance by different classifiers is broadly the same.

3.5 Summary

In this chapter, it is explained how the work done by Zhou et al. [5] was re-implemented as a baseline for this thesis. They created a classifier using support vector machines (SVM) with RBF kernel, in order to identify anti-vaccine tweets about human papillomavirus vaccines (HPV). They trained and tested their classifiers using social connection information from tweets. We used three machine learning algorithms; SVMs with linear and RBF kernels and logistic regression, with tweet n-grams and direct social connections, as the set of features. All of the classifiers obtained their best result with combination of n-grams and sources social connections with f1-score more than 0.93. Because all three classifiers obtained quite same results, we only use SVM (RBF) in the next chapters as baseline. Although in this thesis, some of the

Features	**SVM(Linear)**	**SVM(RBF)**	**Log. Regression**
Baseline - Majority class	0.883	0.883	0.883
Baseline - Opinion Finder (OF)	0.883	0.883	0.883
N-grams	0.900	0.916	0.922
N-grams + χ^2	0.917	0.912	0.917
N-grams + OF	0.899	0.905	0.919
Followers	0.924	0.922	0.922
N-grams + Followers	0.927	0.926	0.935
Sources	0.925	0.926	0.932
N-grams + Source	0.931	0.931	**0.939**

Table 3.2: The performance of classifiers (f1-score) trained on first three months period and tested on second three months period

non-standard approaches in [5], like feature selection, were replaced by standard ones, we got similar results after re-implementation. In the next chapters, the work in [5] will be extended.

Chapter 4

Label Propagation

"Tweets are not created in isolation" and they are related to each other based on the relationships between their authors [1]. Related tweets are tweets published by a user, tweets that reply another tweet, and re-tweets. Opinions can propagate like a virus among the users network in Twitter, because authors are influenced by the tweets of other authors they follow or are followed by. Not only directly connected authors can have influence on each other, but also users can be affected by others who are indirectly connected to.

Speriosu et al. [1] used label propagation to exploit indirect social connection information to improve polarity classification. Adapting the technique from [1], we implement label propagation using Twitter follower and source graphs and evaluate the results. Then, the predictions of label propagation are used within a classifier. We will be expanding this label propagation approach with more features after defining some new features in the next two chapters.

4.1 Technical Background

Label propagation has been used in a number of NLP tasks [21], [22]. Speriosu et al. [1] used label propagation for Twitter sentiment classification. The work in this thesis follows that approach, so we give some detail on it below.

Label propagation algorithms are graph-based semi-supervised machine learning methods and propagate labels from a small set of initially labeled nodes throughout the graph [1]. In the label propagation process, data is represented as a graph in which vertices or nodes are data features and weights of edges are the pairwise distances (closeness, equality, or similarity) between the features. Labels are propagated through the graph, from labeled nodes to unlabeled

ones [20]. For instance in Twitter, the relationship between users and their followers and sources could be converted into a graph, that shows how users with similar interests and ideas are connected together. Figure 4.3 depicts a graph of users and their followers in the network of a subset of the existing data.

A graph $G = \{V, E, W\}$ consists of a set of vertices or nodes V, with $|V| = n$, that are connected with set of edges E. W is an $n \times n$ matrix of weights, in which w_{ij} is the weight of the edge between nodes i and j. There are different types of graphs based on their type of relationships between their nodes. In binary graphs, the weight matrix contains 1s or 0s according to whether two nodes are connected or not, whereas in weighted graphs edges have different weights which are calculated based on the similarity of the nodes. If the edges have a direction the graph is called directed, otherwise it is undirected and the matrix is symmetric.[1]

There are different graph-based algorithms for class inference over graphs. In this work, following [1], Modified Adsorption (MAD) algorithm is used. Adsorption is an iterative algorithm, where at each iteration node labels are estimated from labels at the previous iteration [45]. The Adsorption algorithm is a random walk over the graph and controls label propagation by limiting the amount of information passed by three actions:

- injecting labels from labeled nodes to their adjacent unlabeled nodes,
- continuing walks from a node to a neighboring one,
- abandoning the walk while a final state is converged.

MAD has the desirable properties of Adsorption in addition to three input hyper-parameters μ_1, μ_2, and μ_3 to control the importance of those actions. Both Adsorption and MAD allow initial seeds to change, which is useful when noisy initial labels exist.[45]

The primary reason for [1] to use label propagation for sentiment classification in Twitter is to explore using unlabeled data, as discussed in Sec 2.2, and investigate its impact on classification performance. In this thesis label propagation is used for capturing indirect connections.

The label propagation work in this thesis follows up the work done by Speriosu et al.[1], on using label propagation over lexical links and the follower graph to improve sentiment classification for several datasets. So, this work is described in some detail. They used three different datasets of tweets annotated for polarity, as training and evaluation resources: the "Stanford Twitter Sentiment" corpus, a collection of 218 tweets with 108 positive and 75 negative ones; "Obama-McCain Debate" tweets, a dataset with 1,898 tweets with 705 positives

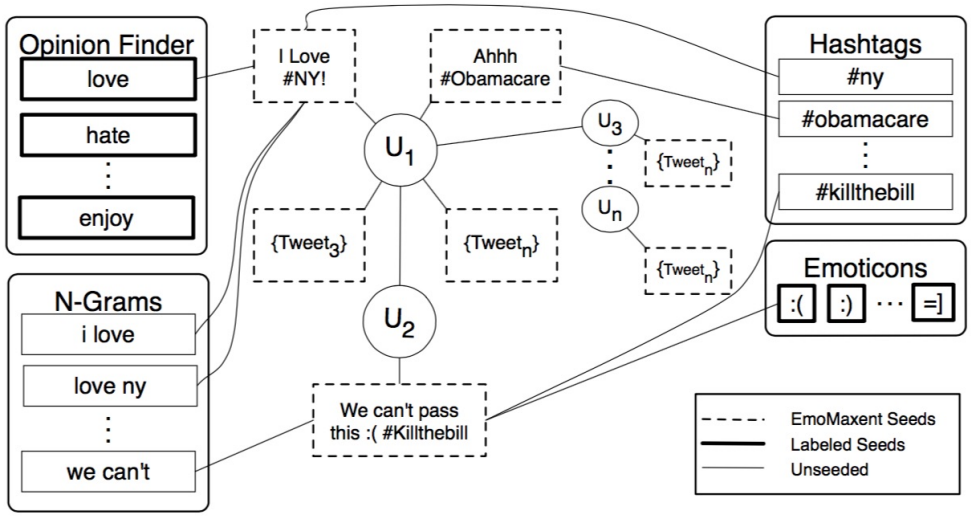


Figure 4.1: An illustration of label propagation graph (adapted from Fig.1 in [1])

and 1192 negatives; "Health Care Reform (HCR)" dataset based on the tweets about health care reform in USA with 1418 tweets. In addition, 1,839,752 tweets containing at least one emoticon were selected from a Twitter feed, annotated based on their emoticons as noisy indicators of polarity.

[1] implemented three approaches to compare the results and evaluate the performance of label propagation. First, a lexicon based label propagation was implemented by counting the number of positive and negative words in a tweet, using OpinionFinder lexicon, and select the polarity with more terms. This approach was used as a baseline. Then, Maximum Entropy classifier was used for polarity classification. Finally, label propagation was implemented to exploit the relationships between tweet users and other useful information of tweets.

In order to implement label propagation, [1] used Modified Adsorption (MAD) algorithm. A graph was created with nodes representing tweets, users, and tweet features (figure 4.1). Graph nodes were tweets, word n-grams, hashtags, emoticons, OpinionFinder words, and users. Users were connected to each other based on their follower and source relationships, and they were connected to the tweets they created. Tweets were connected to their n-grams, hashtags, emoticons, and OpinionFinder words. Edges were weighted as follows: an edge between user A who follows a user B was weighted as 1, and an edge between tweet t and feature f was weighted as w_{tf} the relative frequency ratio of the feature and was calculated by:

$$w_{tf} = \begin{cases} log\frac{P_d(f)}{P_r f} & \text{if } P_d(f) > P_r f \\ 0 & \text{o.w.} \end{cases} \tag{4.1}$$

where d is the dataset that tweet t and feature f belongs to, and r is the Emoticon dataset used as reference corpus.

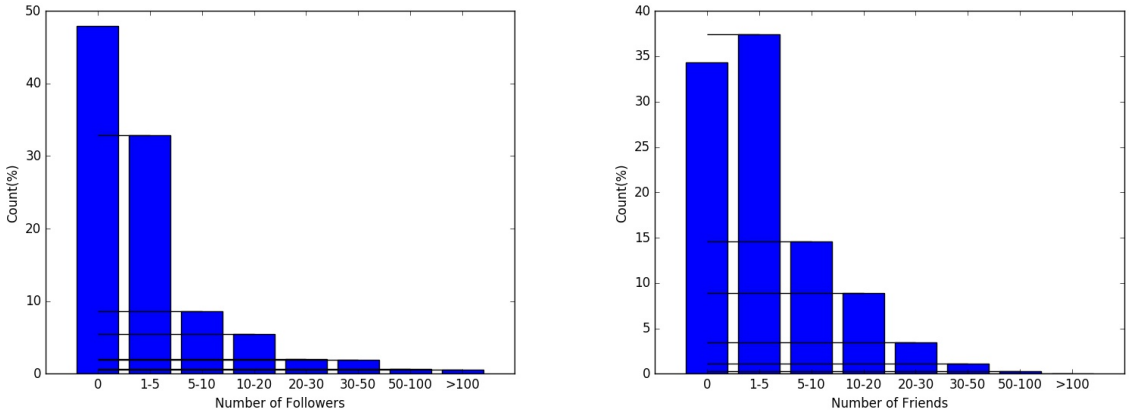
Some of the nodes of the graph were seeded with initial label information. Four groups of variants were considered as seeds:

- Labels predicted by Maximum Entropy classifier were used to seed tweet nodes. In figure 4.1 these EmoMaxent seeds are shown by dashed lines.
- All OpinionFinder words were seeded based on their polarity. Strongly positive/negative words were labeled as 90% positive/negative, and weakly positive/negative words were labeled as 80% positive/negative.
- All emoticons were labeled as 90% positive/negative based on their polarity.
- The labels in HCR database were used as seeds for the tweet nodes.

The results obtained by [1] for all datasets showed that label propagation outperformed or had same results with other approaches. The best performing label propagation was for Stanford Twitter Sentiment dataset with 84.7% accuracy.

4.2 Data preparation

In this chapter, label propagation is implemented using graphs with only users and their followers and sources as nodes. Chapter 7 incorporates more types of nodes, after these have been defined in intervening chapters. In order to prepare data for label propagation, two separate graphs with Twitter sources and followers networks are created. In the current dataset there are 1725 distinct authors for the total of 2098 tweets. From the full list of user sources and followers, only users who are in the list of tweet authors are selected. In this way, users who did not have any tweets in the dataset are removed. Figure 4.2 shows the distribution of sources and followers numbers in the current dataset, for both first and second period. More than 45% of users have no followers, while 30% have between 1 to 5 followers (figure 4.2a). As can be seen in figure 4.2b, less than 35% of users do not follow anybody in the existing dataset and more than 35% follow between



(a) Followers (b) Sources

Figure 4.2: Distribution of Followers and Sources

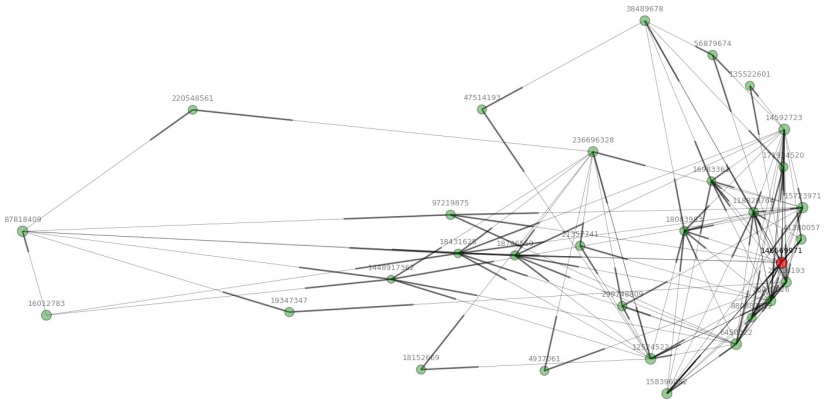


Figure 4.3: Input graph of Twitter users with more than 50 followers in the current dataset from both first and second period.

1 to 5 users. Because a large number of users do not have any followers or sources in the existing dataset, graphs have a noticeable number of isolated nodes. No label can be predicted for isolated nodes and this can affect classification performance negatively; we deal with this below by assigning the major class label to isolated vertices.

In the present work, weighted and undirected input graphs are created, as in [1], with only users as vertices. An edge between user i and j with weight > 0 shows that user i follows j or is followed by j, while an edge with weight equals to 0 means users i and j are not connected. Figure 4.3 shows a part of followers input graph, containing users with more than 50 followers.

In order to prepare initial seeds for label propagation, labels are assigned to users who are

only in the first three months period of data. For each user the total number of anti- and pro-vaccine tweets are calculated. If a user has more anti-vaccine tweets than pro-vaccine ones, the node representing that user in the graph is labeled as anti-vaccine otherwise pro-vaccine. In the case of having equal number of anti- and pro- vaccine tweets labels are selected randomly.

4.3 Experimental Setup

To implement label propagation, Junto label propagation Toolkit[1] with Modified Adsorption (MAD) algorithm [45] is used.

Preparing input files. Junto requires 3 input files: input graph, seeds, and golden labels. In addition, three MAD parameters and the number of iterations to propagate labels are specified. The input graph file contains all the nodes information to create the graph. The seeds file contains initial seeds to start walk through the graph, while the golden file contains all labels from training and test set to be compared with the predicted labels and is used to evaluate the algorithm.

For the vaccination Twitter sentiment task, firstly, two separate input graph files are created. One file contains the list of all users and their followers and the other one contains the list of users and their sources. The input graph file contains triples in the form of: <*source_ node*> <*target_ node*> <*edge_ weight*>. In the followers graph <*target_ node*> follows <*source_ node*>, while in the sources graph the order is reversed. <*Edge_ weight*> is calculated from the log of total number of source node's followers or sources.

Secondly, a set of initial seeds is required to create the seeds file. The seeds file is created in the format of: <*node_ name*> <*seed_ label*> <*seed_ score*>, where <*seed_ score*> is a number between 0 and 1, depending on the level of confidence on the <*seed_ label*>. This score is used to identify noisy labels. Finally, the Gold label file is created with the same format with the seed file as evaluation set.

The output file created by label propagation contains a line for each node. In each line, there is an estimated label and score in the format of: *[AssignedLabel Score]+*, in which a score (a number between 0 to 1) is calculated for each predicted label and shows the probability of being assigned to its corresponding node. One sample of estimated label and score is: L_1

[1] https://github.com/parthatalukdar/junto

0.9621008357244348 L_2 0.003069755513487841 _ _ DUMMY _ _ 0.0028080883894480387, where in this case a score 0.962 (rounded) is assigned to label L_1, 0.003 to label L_2, and score 0.003 to the special label _ _ DUMMY _ _. DUMMY is a special label which has the semantics of "none of the other labels"label and it is assigned high score by the algorithm if there is not enough evidence to assign any of the other labels.

In the current work, for label propagation the number of iterations is set to 100 and the injection parameter μ_1 is changed to 0.005, and for other parameters default values are used, following Speriosu et al. [1].

Models. The first model will just be the straight application of label propagation, as in [1]. The second model will take the output of label propagation as an indication of the indirect influence of users. As such, it can be added to the classifier in the same way as direct followers and sources were in chapter 3. Two new features are added to the first and second period data. One feature contains the predicted probabilities for anti-vaccine label and the other one contains the predicted probabilities for pro-vaccine labels. SVM (RBF) classifier is trained and tested with these two new features and combination on these features with n-grams.

To assign a label to isolated nodes and calculate f1-score, all of the tweets with no predicted labels were labelled as pro-vaccine. Pro-vaccine is the majority class in both training and test set.

4.4 Results

In this section, label propagation models are evaluated and the results are compared with the results of SVM (RBF) classifier, from chapter 3.

First, label propagation is implemented by information only in first three months period training set and is tested using 10-fold cross validation. The results are listed in table 4.1: the first block contains the baseline, the second block contains the best results from chapter 3, the third block gives the results of the first model; and the forth block gives the results of the second model.

The label propagation models trained and tested on first period of data using 10-Fold cross validation did not have better results than SVM classifier with RBF kernel trained and tested on

[2]The best results from chapter 3

Methods	**F1-score**
Baseline - Majority Class	0.845
SVM(RBF)[2]- Direct Followers	0.842
SVM(RBF) - N-grams + Direct Followers	0.902
SVM(RBF) - Direct Sources	0.891
SVM(RBF) - N-grams + Direct Sources	0.902
Label Propagation - Followers	0.853
Label Propagation - Sources	0.861
SVM(RBF) - Label Prop. (Followers)	0.951
SVM(RBF) - N-grams + Label Prop. (Followers)	**0.952**
SVM(RBF) - Label Prop. (Sources)	0.938
SVM(RBF) - N-grams + Label Prop. (Sources)	0.936

Table 4.1: The performance of label propagation testing on first period data compared to the performance of classifiers testing on first period data using 10-fold cross validation.

same set of data with 10-fold cross validation (table 4.1). However, label propagation results are comparable to the results of SVM(RBF) with direct followers and sources. The best result is for SVM(RBF) classifier with both n-grams and label propagation results for followers as features, with f1-score of 0.952. Table 4.1 shows that label propagation models with connection-based nodes and their combination with SVM classification perform better than classifiers using direct connection-based features, trained and tested on training set.

Then, label propagation is trained by initial seeds from first three months period and tested in second three months period. After implementing label propagation using social connections network, about 40% of 1080 tweets from 922 unique users in second three months period were not labeled. This is because of the large number of isolated nodes in both followers and source graphs. The complete results are given in table 4.2: the first block contains the baseline, the second block contains the best results from chapter 3, the third block gives the results of the first model; and the forth block gives the results of the second model.

The performance of label propagation was slightly better in the second period than the first period, and the f1-score values increased for both followers (0.875) and sources (0.872) networks (Table 4.2); however, it is below the higher baseline for the period. SVM classifiers

[3]The best results from chapter 3

Methods	**F1-score**
Baseline - Majority class	0.883
SVM(RBF)[3]- Direct Followers	0.922
SVM(RBF) - N-grams + Direct Followers	0.926
SVM(RBF) - Direct Sources	0.926
SVM(RBF) - N-grams + Direct Sources	**0.931**
Label Propagation - Followers	0.875
Label Propagation - Sources	0.872
SVM(RBF) - Label Prop. (Followers)	0.883
SVM(RBF) - N-grams + Label Prop. (Followers)	0.902
SVM(RBF) - Label Prop. (Sources)	0.883
SVM(RBF) - N-grams + Label Prop. (Sources)	0.919

Table 4.2: The performance of label propagation testing on second period data, compared to the performance of classifiers testing on second period data.

with direct followers and sources as features outperformed label propagation. Using label propagation results as features for SVM(RBF) classifier, slightly improved the classification results. However, the results show that label propagation using only social connection features is not helpful to improve polarity classification across data separated by large periods of time.

4.5 Summary

Label propagation was implemented as a semi-supervised tweet polarity classification method, using only social connections information. For label propagation, followers and sources graphs were created separately and seeded by labels assigned to users from their posted tweets. While results were substantially better for label propagation in the first time period, indicating that indirect social connection was useful, this improvement was not sustained in the second. There is a room for improvement in the way the graph is extended by adding more features, as in [1], and some external features, like the textual context in the referenced pages from a tweet. Label propagation is extended in chapter 7.

Chapter 5

Topic Modelling

Section 3.4 noted that there is a large feature space relative to the size of the dataset. We found (as did Zhou et al. [5]) some features that were significant in the first period dataset were for the most part not significant in the second period. This suggests that some kind of dimensionality reduction could be useful.

Surian et al. [42] have investigated the use of topic modelling along with community structure detection for the present domain. They found that the alignment between topic modelling and community structure detection can be a useful way to characterize Twitter communities for public health surveillance.

In this chapter we look at including topics from topic models as features in our classification task, as in [38],[39].

5.1 Topic Modelling Techniques

As described in [26], topic modelling methods are appropriate for identifying themes within a set of Twitter posts. Identified topics can be used as features for classifiers, in order to reduce the feature space dimensionality.

Topic modelling algorithms are statistical methods that analyze the words in documents in order to:

- identify themes that run through documents,
- identify how those themes are connected,
- identify how those themes change over time.

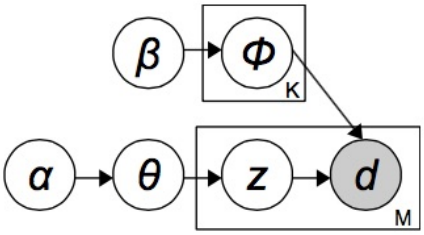


Figure 5.1: Graphical model of DMM (Adapted from figure A2 in Appendix 1 [42])

One advantage of topic modelling algorithms is that they do not need any prior annotation of documents, as topics are identified iteratively by analyzing texts. A topic is considered as a distribution over words. For example, a *data analysis* topic will have words about data analysis with high probability, such as "computer" and "prediction ".

Applying topic modelling on Twitter is a challenging task, because of the short length of tweets (140 characters). Mainly, topic modelling is used for documents with greater lengths than tweets.

In this thesis, Dirichlet Multinomial Mixture (DMM) models, designed specifically for short texts, are used to infer topics associated with tweets. A DMM model [46] is a generative topic model and assigns one topic to each tweet. Figure 5.1 shows the graphical model of DMM. Each circle in the graphical model shows a variable and each rectangle shows a repetition of a variable. In this model, there are M documents and each document $d_m \in \{d_1, \ldots, d_M\}$ is considered as a bag of words. In the generative process of DMM, a distribution ϕ_k over words is generated for each topic $k \in \{1, \ldots, K\}$. ϕ_k is chosen from the Dirichlet distribution with a hyper-parameter β ($\phi_k \sim Dir(\beta)$). Topic mixtures θ are created from Dirichlet distribution with a hyper-parameter α ($\theta \sim Dir(\alpha)$). In order to generate the document d_m, DMM first chooses a topic z_m from the Multinomial distribution on θ ($z_m \sim Multinomial(\theta)$). Then DMM generates the words w_m in document d_m from the Multinomial distribution conditioned on the topic z_m ($w_m \sim Multinomial(\phi_{k=z_m})$). The probability of the document d generated from topic k is calculated by:

$$p(d|z = k) = \prod_{w \in d} p(w|z = k). \tag{5.1}$$

5.2 Experimental Setup

Data preparation for Topic Modelling. In order to implement topic modelling, all tweets about HPV vaccine collected by Zhou et al. [5] from October 1, 2013 and March 31, 2014 are utilized. The complete dataset contains 131,286 tweets.

Several pre-processing steps are required to prepare data for topic modelling, as in section 3.2. The only difference is that all the words in the tweets are stemmed using NLTK Porter Stemmer[1]. After pre-processing, all duplicate tweets are removed. The final corpus contains 81381 unique tweets.

Topic Modelling Configuration. In order to implement topic modelling, a Java package for the DMM topic models, jLDADMM[2] is used [47].

The DMM model from jLDADMM is used with the set of unique tweets. The input corpus is a text file, each line representing one tweet. In addition to input corpus, some input parameters are configured before applying DMM model. In the current work, for topic modelling the number of topics is set as 80 and the number of Gibbs sampling iterations is set to 2000. The default values of hyper-parameters α and β are used.

After running DMM model from jLDADMM, there are five outputs from the DMM:

- document-to-topic distributions,
- topic-to-word distributions,
- top topical words,
- the parameters used in the model, and
- the list of topics assigned to tweets. In this model one topic is assigned to each tweet.

Topic Models as Features. In order to utilize DMM model results in sentiment classification, the file topicAssignments is mapped to the training and test set. Using one-hot representation, a binary vector with length 80 (the number of topics), $v_1, \ldots, v_{80}$, is added to each tweet in the classification training and test sets. The values in each binary vector are obtained from respective tweets in the list of topics assigned. If a topic m is assigned to a tweet, in its vector v_m is one and all other entities are zero. These vectors are utilized as new features for classification.

[1] http://www.nltk.org/api/nltk.stem.html

[2] http://jldadmm.sourceforge.net

5.3 Results

After adding topic modelling assigned topics to the set of features, sentiment classification is carried out with these features only and in combination with previous features. The results are compared with the results of SVM (RBF) classifier, from chapter 3.

Classifiers training and testing in first period Classifiers are trained and tested in the training period using ten-fold cross validation and results are shown in table 5.1: the first block gives the earlier baselines, the second block gives the best results from chapter 3; and the third block gives the new results. In this period, best result obtained by Zhou et al. [5] with combined n-grams, followers, and sources set of features and majority class, are used as baselines. The best-performing classifier using DMM topics, in the training period is SVM(RBF) constructed from combination of n-grams and DMM topics with 0.889 f1-score, which is comparable to the baseline results. Adding DMM topics to set of features does not perform better than n-grams and connection-based sources features in classification. However, classifiers using combination of n-grams and DMM topics set of features outperform classifiers using only n-grams, so they do appear to be contributing some useful information.

Methods	**F1-score**
Baseline - Majority Class	0.845
SVM(RBF) - N-grams + Followers + Sources ([5])	0.89
SVM(RBF) - N-grams + Direct Sources[3]	**0.902**
SVM(RBF) - N-grams	0.847
SVM(RBF) - DMM topics	0.826
SVM(RBF) - N-grams + DMM topics	0.889

Table 5.1: The performance of classifiers (f1-score) trained and tested on first three months period using 10-fold cross validation

Classifiers testing in second period Classifiers achieved slightly better f1-scores in second period. Table 5.2 contains the complete results of sentiment classification tested on second three month period data: the first block gives the earlier baseline, the second block gives the best

[3]The best results from chapter 3

results from chapter 3; and the third block gives the new results. All the classifiers using DMM topics as their set of features have better results than the baseline results considering majority class. The best-performing classifier in the test period is SVM(RBF) constructed from both content features (n-grams) and DMM topics with 0.942 f1-score. This is slightly better than the best-performing classifier in section 3.4, classifiers constructed from n-grams and sources set of features. We have previously found that content features were less useful across time periods, but the results here show that topic models have more than compensated for this.

Methods	**F1-score**
Baseline - Majority class	0.883
SVM(RBF) - N-grams + Direct Sources[4]	0.931
SVM(RBF) - N-grams	0.916
SVM(RBF) - DMM topics	0.909
SVM(RBF) - N-grams + DMM topics	**0.942**

Table 5.2: The performance of classifiers (f1-score) tested on second three months period

5.4 Summary

DMM topic model algorithm was implemented to reduce the dimensionality of feature space, by finding the themes that run through the tweets and assign one topic to each tweet. Assigned topics were used as new features for sentiment classification. SVM (RBF) classifier with the combinations of tweet n-grams and DMM topics improved over just using an n-gram representation of content alone. Results show that topics are especially useful across time periods. In the light of their usefulness, DMM topics are exploited to extend label propagation later.

[4]The best results from chapter 3

Chapter 6

Incorporating URL contents with Sentiment Classification

In Twitter some users add URLs to their tweets to reference web pages that contain related information to their tweets. A key issue in Twitter is that tweets are short, so the content of linked web pages as a source of additional context is particularly important. The goal of this chapter is to incorporate that external information with anti-vaccine sentiment classification and examine the impact on classification performance.

There might be useful information in contents of web pages that could enhance the accuracy of sentiment classification. For example, figure 6.1 shows an anti-vaccine tweet with a link to a website that contains an article about HPV vaccine. The article is against HPV vaccine and as shown in figure 6.1 it contains anti-vaccine content, which can be used as extra source of information for sentiment classification.

6.1 Data Preparation

In order to prepare data, first, all the URLs are extracted from tweets in first and second period data. Among them there are some URLs generated from generic URL shortening services, in a format like "http://bit.ly". Some of the generic URLs are replaced by the actual expanded URLs in the metadata associated with the tweets, if identified in the list of expanded URLs. Otherwise they are excluded from the dataset, as in [42].

Then, all the textual contents of html pages are extracted. Two separate files are created containing URLs from first and second period of data, along with their textual contents. The

Figure 6.1: An example of web page content extracted (bottom) from the URL in tweet (top).

label assigned to the textual contents was the label of the associated tweet, making this an instance of distant supervision, and the dataset was balanced by removing neutral tweets to have the equal number of anti- and pro-vaccine tweets. The final URL datasets contain 406 records in the first period and 601 in the second one.

Finally in order to prepare data for sentiment classification, URL contents need to be pre-processed, following broadly the same pattern as data preparation in section 3.2. First, all non-word elements like punctuation and numbers, and stop words are removed and all contents are converted to lowercase. Then, all the words are lemmatized to remove inflectional endings and to return the base or dictionary form of a word, using NLTK WordNetLemmatizer[1].

Web pages content texts are then converted into unigrams and bigrams. The unigram model creates a sparse vector for each content with the length equal to the size of the vocabulary, representing the presence or count of each word in a content. Bigram is same as unigram, except for considering the presence or count of pairs of adjacent words in a text.

[1]http://www.nltk.org/api/nltk.stem.html

6.2 Experimental Setup

Two separate models are used to incorporate content of web pages referenced in tweets with sentiment classification. In the first model, n-grams of web pages contents are added to initial datasets to be used in classification directly. This makes the feature vector much larger by adding 8755 new features (the size of the vocabulary of web pages content). Web pages n-grams are used alone or in combination with the tweet n-grams, as input features for the SVM (RBF) classifier. In this model, a significant number of tweets do not have URLs or web page contents, which can affect results negatively.

The second model implements a two level sentiment classification using distant supervision. First two new URL training and test sets are used for classification. A SVM (RBF) classifier is trained with n-grams of web pages contents from the first period URL dataset and tested on both training and test set to predict labels for corresponding URLs. For example, "vactruth.com" is assigned the label anti-vaccine, because it is linked from multiple anti-vaccine tweets. Then, those predicted labels for URLs are added as new features to initial tweets training and test set, and they are used for sentiment classification, on their own or in combination with other features.

6.3 Results

Classifiers training and testing in first period. First, classifiers are constructed and tested in the training period using ten-fold cross validation. Table 6.1 contains the complete results of sentiment classification trained and tested on first three month period data: the first block contains the baselines, the second gives the results of the first model; and the third block gives the results of the second model. Majority class and SVM (RBF) classifier trained with tweet n-grams as set of features, are used as baselines to compare the results with. In this period, sentiment classification using n-grams of web pages contents directly as features did not obtain better results than the baselines. However, combining those features with tweet n-grams slightly increased f1-score from 0.837 to 0.859. Adding predicted URL labels under supervision to sentiment classification resulted in a slightly better f1-score than using URL n-grams. The best-performing classifier in the training period is with the combination of URL predicted labels and tweet n-grams with 0.897 f1-score. The results show that the second model performed better.

[2]The result from section 3.4

Methods	**F1-score**
Baseline - Majority Class	0.845
SVM(RBF)[2]- Tweet N-grams	0.847
SVM(RBF) - URL N-grams	0.837
SVM(RBF) - Tweet N-grams + URL N-grams	0.859
SVM(RBF) - URL Predicted Labels	0.841
SVM(RBF) - Tweet N-grams + URL Pred. Labels	**0.897**

Table 6.1: The performance of classifiers (f1-score) trained and tested on first three months period using 10-fold cross validation

Methods	**F1-score**
Baseline - Majority Class	0.883
SVM(RBF)[3]- Tweet N-grams	0.917
SVM(RBF) - URL N-grams	0.893
SVM(RBF) - Tweet N-grams + URL N-grams	0.895
SVM(RBF) - URL Predicted Labels	0.891
SVM(RBF) - Tweet N-grams + URL Pred. Labels	**0.931**

Table 6.2: The performance of classifiers (f1-score) trained on first three months period and tested on second three months period

Classifiers testing in second period As usual, classifiers achieved slightly better f1-scores in second period. Table 6.2 shows the complete results of sentiment classification tested on second three months period data: the first block contains the baselines, the second gives the results of the first model; and the third block gives the results of the second model. The best-performing classifier in the second period is with the combination of URL predicted labels and tweet n-grams with 0.931 f1-score, which is better than the majority class and classification using only tweet n-grams. Unlike the result in the first period, the combination of tweet n-grams and URL content n-grams with f1-score 0.895 did not perform better than SVM(RBF) with only tweet n-grams. Under distant supervision, adding the predicted URL labels to the classifier resulted in f1-score of 0.891.

[3]The result from section 3.4

6.4 Summary

In this chapter, it was explained how the content of web pages, referenced in tweets, can be used as features for sentiment classification. Two separate approaches were implemented for sentiment classification with URL pages contents; one with n-grams of textual contents and one using distant supervision in a two level classification.

Sentiment classification using web pages contents had comparable results with sentiment classification using only tweet n-grams as features. The important result is that adding URL predicted labels did better than either tweet n-grams alone or tweet n-grams with content n-grams added. This shows that referenced web pages may contain useful information that could improve polarity classification. This information is used in next chapter as extra nodes to extend label propagation.

Chapter 7

Extended Label Propagation

We found from previous chapters that all the previous techniques, label propagation, topic modelling, and linked webpage info can contribute to improvements. This chapter explains how label propagation from chapter 4 is extended by adding more features beyond followers and sources. The idea behind this extension is to exploit extra information in tweets with label propagation and explore the possibility of improving classification.

Label propagation has a way of incorporating multiple types of information and obtaining label distributions for different types of nodes, as noted in [1]. Motivated by this property, a graph is constructed with users, tweets, tweet words n-grams, topics, hashtags, URLs, and emoticons as nodes. Graphs constructed with only users were sparse, while adding more nodes to graphs results in more dense graphs. Speriosu et al. [1] found that using nodes of other types than social connections increases the performance on their task.

7.1 Experimental Setup

In order to prepare data for extended label propagation, two separate graphs, one containing Twitter sources connections and one containing followers relationships in Twitter are created, exactly in the same way as graphs in section 4.3. More features are added to these graphs as nodes, in order to examine the effect on the label propagation results.

Features, other than followers and sources, are extracted from tweets and added to graphs as vertices. Figure 7.1 shows an illustration of graphs created for extended label propagation. These features are a combination of the features investigated in the previous chapters, plus others that [1] used: tweetIds (tweet identifiers), hashtags, emoticons, n-grams, positive and negative

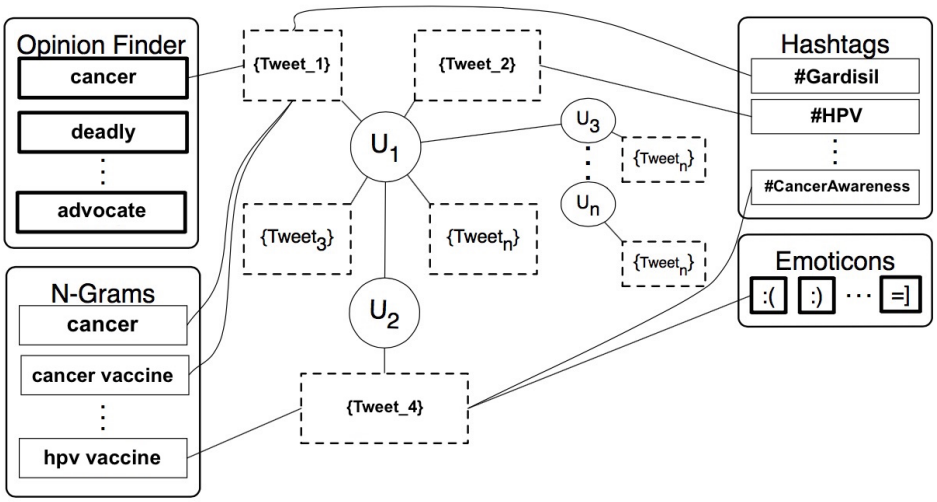


Figure 7.1: An illustration of extended label propagation graph (adapted from Fig.1 in [1])

words, URLs, URL content n-grams, and topics. Some of these features are selected directly from database and added to graphs without any changes, while others require some preparation before being added to the graphs. The graph is constructed as follows:

User vertices Users (vertices U_i in figure 7.1) are connected to each other based on the Twitter follower or source network, as described in section 4.2. Edges between followers and sources are weighted in the same way of section 4.2, with log of the number of followers our sources of users.

TweetID vertices We create a new vertex type for TweetIDs, indicated by Tweet_i in figure 7.1. An edge, with weight 1 is created from Tweet_i to user U_j if U_j posted Tweet_i.

Hashtag vertices We create a new vertex type for hashtags in tweets, with these vertices contained within the Hashtags box in figure 7.1. An edge, with weight 1 is created between hashtag h_i and Tweet_j if h_i occurs in Tweet_j.

Emoticon vertices We create a new vertex type for emoticons in tweets, with these vertices contained within the Emoticons box in figure 7.1. An edge, with weight 1 is created between emoticon e_i and Tweet_j if e_i occurs in Tweet_j. Emoticons are used as noisy indicators of sentiment and a list of positive and negative emoticons is listed in table 7.1.

Tweet N-grams vertices In order to exploit tweet texts as features for label propagation, the most frequent unigrams and bigrams are selected and added to graphs as vertices of type

N-Grams, in figure 7.1. First, tweets are pre-processed in the same way as section 3.2 and then converted to unigrams and bigrams vectors containing words with frequency more than 15. The resulting list of unigrams and bigrams with 118 records is selected and added to graphs as nodes.An edge, with weight 1 is created between n-gram n_i and Tweet_j if n_i occurs in Tweet_j.

OpinionFinder vertices We create a new vertex type for every word in the OpinionFinder lexicon[1], with these vertices contained within the Opinion Finder box in figure 7.1. An edge, with weight 1 is created between opinion finder word OF_i and Tweet_j if OF_i occurs in Tweet_j.

Topic vertices All the topics assigned to tweets in chapter 5 are added to graph as vertices of type Topic. An edge, with weight 1 is created between topic t_i and Tweet_j if t_i is assigned to Tweet_j.

URL vertices All the URLs referenced in tweets are added to graph as vertices of type URL. An edge, with weight 1 is created between URL url_i and Tweet_j if url_i occurs in Tweet_j.

URL N-grams vertices In order to exploit the textual content of referenced web pages in tweets, a new vertex of type URL content n-grams is created. First, contents are retrieved from the HTML documents of URLs. Then, html tags are removed and only text is extracted. Texts are converted to unigrams and bigrams vectors and n-grams with frequency more than 15 are selected to be added to label propagation graphs. The 8755 most frequent unigrams and bigrams are selected from URLs contents. An edge, with weight 1 is created between URL content n-gram $url - n_i$ and URL url_j if $url - n_i$ occurs in url_j.

Positive	:) :D =D =) :] =] :-) :-D :-] ;) ;D ;] ;-) ;-D ;-]
Negative	:(=(:[=[:-(:-[:'(

Table 7.1: List of positive and negative emoticons (reproduced from Table 1 in [1]

After constructing graphs, initial labels are required for seeding them. Among all nodes in graphs, only training set tweetIds, OpinionFinder words, and emoticons are initially labeled. Each tweet node from training set is seeded with its polarity label from the database. We label

[1]lexicon http://mpqa.cs.pitt.edu/opinionfinder

positive words or emoticons as pro-vaccine and negative ones as anti-vaccine. We do this on the basis of the observation made in section 3.4, although there may not in fact be a strong correlation between positive and negative sentiment and pro- and anti-vaccine sentiment. Following [1], OpinionFinder words are labeled as follows: If a word is positive and strongly subjective, it is labeled as 90% pro-vaccine, and if a word is positive and weekly subjective, it is labeled as 80% pro-vaccine. This process is done similarly for negative words with anti-vaccine labels. All emoticons are labeled as 90% pro-vaccine if they are positive and 90% anti-vaccine if they are negative.

As in section 4.3, we supply Junto label propagation Toolkit[2] with Modified Adsorption (MAD) algorithm, with the same parameter settings.

Two input graphs are constructed; one for Twitter followers network plus all other features and one for Twitter sources network with same extra features. For seed file, in order to obtain the best result the training set was under-sampled to become balanced and tweets with neutral polarity were omitted. After that, 581 labels remained from 1018 tweets in training set, with 273 anti-vaccine labels and 308 pro-vaccine ones. Noisy labels for emoticons and OpinionFinder lexicon were added and resulted in a seed file with 4466 seeds.

Models. The first model will just be the straight application of label propagation, as in [1]. The second model will take the output of label propagation to incorporate with a classifier. As such, probabilities of predicted labels from the label propagation can be added to the classifier in the same way as tweet n-grams and direct followers and sources were in chapter 3. Two new features are added to the first and second period data. One feature contains the predicted probabilities for anti-vaccine label and the other one contains the predicted probabilities for pro-vaccine labels. SVM (RBF) classifier is trained and tested with these two new features and combination on these features with n-grams.

7.2 Results

First, extended label propagation is implemented by information only in first three month period training set and is tested using 10-fold cross validation. Table 7.2 shows the results in first three month period: first block contains the majority class baseline, the second block gives the best

[2]https://github.com/parthatalukdar/junto

Methods	**F1-score**
Baseline - Majority Class	0.845
SVM(RBF)[3]- Direct Followers	0.842
SVM(RBF) - N-grams + Direct Followers	0.902
SVM(RBF) - Direct Sources	0.891
SVM(RBF) - N-grams + Direct Sources	0.902
Label Propagation - Followers	0.853
SVM(RBF) - Label Prop. (Followers)	0.951
SVM(RBF) - N-grams + Label Prop. (Followers)	0.952
Label Propagation - Sources	0.861
SVM(RBF) - Label Prop. (Sources)	0.938
SVM(RBF) - N-grams + Label Prop. (Sources)	0.936
EX-Label Prop. - Followers	0.965
EX-Label Prop. - Sources	0.966
SVM(RBF) - EX-Label Prop. (Followers)	0.841
SVM(RBF) - N-grams + EX-Label Prop. (Followers)	0.883
SVM(RBF) - EX-Label Prop. (Sources)	0.967
SVM(RBF) - N-grams + EX-Label Prop. (Sources)	**0.977**

Table 7.2: The performance of extended label propagation testing on first period data compared to the performance of classifiers testing on first period data, using 10-fold cross validation.

results from chapter 3, the third block gives the results from chapter 4, the forth block contains the new results of the first model, and the fifth block gives the new results of the second model.

Results obtained from 10-fold cross validation for extended label propagation, with either followers or sources, are better than the results from SVM classifier with RBF kernel and with direct connection based features. However, extended label propagation did not perform better than label propagation using only connection based nodes (table 7.2). SVM(RBF) classifier trained and tested on training set using 10-fold classification with n-grams and extended label propagation (with sources nodes) results has the best performance with a surprisingly high f1-score 0.977 among all classifiers.

Then, extended label propagation is trained by initial seeds from first three months period

[3]The best results from chapter 3

Methods	**F1-score**
Baseline - Majority class	0.883
SVM(RBF)[4]- Direct Followers	0.922
SVM(RBF) - N-grams + Direct Followers	0.926
SVM(RBF) - Direct Sources	0.926
SVM(RBF) - N-grams + Direct Sources	0.931
Label Propagation - Followers	0.875
SVM(RBF) - Label Prop. (Followers)	0.883
SVM(RBF) - N-grams + Label Prop. (Followers)	0.902
Label Propagation - Sources	0.872
SVM(RBF) - Label Prop. (Sources)	0.883
SVM(RBF) - N-grams + Label Prop. (Sources)	0.919
EX-Label Prop. - Followers	0.90
EX-Label Prop. - Sources	0.897
SVM(RBF) - EX-Label Prop. (Followers)	0.883
SVM(RBF) - N-grams + EX-Label Prop. (Followers)	**0.932**
SVM(RBF) - EX-Label Prop. (Sources)	0.896
SVM(RBF) - N-grams + EX-Label Prop. (Sources)	0.897

Table 7.3: The performance of label propagation testing on second period data, compared to the performance of classifiers testing on second period data.

and tested in second three months period. In contrast to label propagation using only connection based nodes (where 40% of tweets were unlabeled), after implementing extended label propagation all the tweets had predicted labels. Table 7.3 shows the results: first block contains the majority class baseline, the second block gives the best results from chapter 3, the third block gives the results from chapter 4, the forth block contains the new results of the first model, and the fifth block gives the new results of the second model.

As with plain label propagation, the performance of extended label propagation was not sustained in the testing period, although extended label propagation outperformed label propagation (table 7.3). After adding extended label propagation results to SVM(RBF) classifier, the best performing classifier is with n-grams and extended label propagation (with followers nodes) set

[4]The best results from chapter 3

of features, with f1-score 0.932 (Table 7.3), essentially the same as SVM(RBF) with n-grams and direct sources.

7.3 Summary

In this chapter, label propagation from chapter 4 was extended by adding some extra features to followers and sources graphs. In this approach, graphs are created with the following features as their vertices: users, tweetIds, hashtags, emoticons, n-grams, positive and negative words from OpinionFinder, URLs, URL content n-grams, and topics from topic modelling.

The results of testing label propagation on the first three month period data using 10-fold cross validation were surprisingly high. This reveals that label propagation could be a very useful approach when graphs are not changing and are consistent in a time period. Extended label propagation testing on second three month period of data had better results than label propagation with only connection-based graphs. Those results are comparable to the results of classifiers from chapter 3.

Chapter 8

Conclusion

Sentiment classification on tweets about human papillomavirus (HPV) vaccines was implemented to identify anti- and pro-vaccine tweets, using different approaches. First, the work done by Zhou et al. [5] on using direct information about social connections in addition to content of tweets, was re-implemented on an existing dataset of HPV vaccine tweets between October 1, 2013 and March 31, 2014. The dataset was divided into two contiguous three-month periods. 1018 tweets from first three months period were used as training and development set with results on that period produced under 10-fold cross validation, and 1080 tweets from second three months period were used as a held-out test set. Three machine learning algorithms (SVM with linear and RBF kernels and logistic regression) were tested and had almost similar results. The best results of SVM(RBF) classifier were with the combination on n-grams and direct connection information of sources, with f1-score 0.902 tested in first period using 10-fold cross validation, and with f1-score 0.93 tested in second period of data.

Sentiment classification task was extended by a graph-based label propagation method in order to make use of indirect social connections information. Label propagation, motivated by [1], was implemented by creating two separate graphs of users and their followers and sources as nodes. The graphs were seeded by the labels of tweets users posted. The best performing classifier with label propagation results, was SVM (RBF) with tweet n-grams and label propagation results of followers in first period with f1-score 0.952, which was quite a bit better than the best of chapter 3 with 3% improvement. In the second period, SVM(RBF) with tweet n-grams and label propagation results of sources graph had the best result with f1-score 0.919; which is comparable to the results of same classifier with tweet n-grams and direct sources.

In order to reduce the size of feature space topic modelling was used. Topic modelling algorithms find the themes that run through the tweets and assign topics to tweets. In this thesis, a topic modelling algorithm (DMM) was implemented to assign a topic to each tweet and those topics were used as new features for sentiment classification. Although using only topics as features for the SVM (RBF) classifier did not improve sentiment classification, using topics in combination with tweet n-grams outperformed previous results, with f1-score 0.942.

In addition to previously mentioned features, external features of tweets can be helpful in sentiment classification. One of the external features is the textual content of web pages that are referenced in tweets. These features were used for sentiment classification using distant supervision. First a classification was done on the text parts of web pages. Then the predicted labels were used as new features for sentiment classification. The best result was obtained by the SVM (RBF) classifier with the combination of tweet n-grams and predicted labels with f1-score 0.931 which is similar to the best performing classification with the combination of tweet n-grams and direct sources from chapter 3.

As in [1], label propagation was extended with all tweet features and external features. Graphs were created with users, tweets, tweet n-grams, hashtags, emoticons, OpinionFinder lexicon words, topics, and n-grams of referenced web pages content as nodes. Graphs were seeded by tweet labels, OpinionFinder words polarities, and emoticons polarities. The results of extended label propagation tested on first period data using 10-fold cross validation were significantly high with f1-score 0.96. The results obtained by testing on second period, with f1-score 0.90, were slightly better than the results of the label propagation using social connections information. High results of extended label propagation in first period show that this approach could be considered as an alternative for conventional methods. Although the results in second period are not better than the results of implemented supervised sentiment classification approaches, they are still comparable to the results obtained by classifiers.

The findings in this thesis suggest some future work directions of improving the graph construction in label propagation by considering asymmetric relationships rather than using undirected graphs and weighting edges precisely based on the features importance. In addition, we can link the current work up with community structure detection approach, as in [42]. All the works done on label propagation in this thesis can also be integrated as one service for semi-supervised sentiment classification.

References

[1] M. Speriosu, N. Sudan, S. Upadhyay, and J. Baldridge. *Twitter polarity classification with label propagation over lexical links and the follower graph*. In *Proceedings of the First Workshop on Unsupervised Learning in NLP*, EMNLP '11, pp. 53–63 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2011).

[2] B. Pang and L. Lee. *Opinion mining and sentiment analysis*. Foundations and trends in information retrieval **2**(1-2), 1 (2008).

[3] X. Ji, S. A. Chun, and J. Geller. *Monitoring public health concerns using twitter sentiment classifications*. In *Healthcare Informatics (ICHI), 2013 IEEE International Conference on*, pp. 335–344 (IEEE, 2013).

[4] M. Salathé and S. Khandelwal. *Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control*. PLoS Comput Biol **7**(10), e1002199 (2011).

[5] X. Zhou, E. Coiera, G. Tsafnat, D. Arachi, M.-S. Ong, and A. G. Dunn. *Using social connection information to improve opinion mining: Identifying negative sentiment about hpv vaccines on twitter* (Amsterdam: IOS Press, 2015).

[6] V. Hatzivassiloglou and J. M. Wiebe. *Effects of adjective orientation and gradability on sentence subjectivity*. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pp. 299–305 (Association for Computational Linguistics, 2000).

[7] Y. Yang and J. O. Pedersen. *A comparative study on feature selection in text categorization*. In *Proceedings of ICML-97, 14th International Conference on Machine Learning (Nashville, TN, 1997)*, vol. 97, pp. 412–420 (1997).

[8] B. Pang, L. Lee, and S. Vaithyanathan. *Thumbs up?: sentiment classification using machine learning techniques*. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pp. 79–86 (Association for Computational Linguistics, 2002).

[9] E. Riloff, J. Wiebe, and W. Phillips. *Exploiting subjectivity classification to improve information extraction*. In *Proceedings of the national conference on artificial intelligence*, vol. 20, p. 1106 (Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005).

[10] C. Lin and Y. He. *Joint sentiment/topic model for sentiment analysis*. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 375–384 (ACM, 2009).

[11] B. A. Hagedorn, M. Ciaramita, and J. Atserias. *World knowledge in broad-coverage information filtering*. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 801–802 (ACM, 2007).

[12] K. Dave, S. Lawrence, and D. M. Pennock. *Mining the peanut gallery: Opinion extraction and semantic classification of product reviews*. In *Proceedings of the 12th international conference on World Wide Web*, pp. 519–528 (ACM, 2003).

[13] Q. Ye, Z. Zhang, and R. Law. *Sentiment classification of online reviews to travel destinations by supervised machine learning approaches*. Expert Systems with Applications **36**(3), 6527 (2009).

[14] A. Balamurali, A. Joshi, and P. Bhattacharyya. *Harnessing wordnet senses for supervised sentiment classification*. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1081–1091 (Association for Computational Linguistics, 2011).

[15] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. *Recursive deep models for semantic compositionality over a sentiment treebank*. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, vol. 1631, p. 1642 (Citeseer, 2013).

[16] Z.-H. Deng, K.-H. Luo, and H.-L. Yu. *A study of supervised term weighting scheme for sentiment analysis*. Expert Systems with Applications **41**(7), 3506 (2014).

[17] S. Li, Z. Wang, G. Zhou, and S. Y. M. Lee. *Semi-supervised learning for imbalanced sentiment classification*. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, vol. 22, p. 1826 (2011).

[18] S. Zhou, Q. Chen, and X. Wang. *Active deep networks for semi-supervised sentiment classification*. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pp. 1515–1523 (Association for Computational Linguistics, 2010).

[19] S. Zhou, Q. Chen, and X. Wang. *Fuzzy deep belief networks for semi-supervised sentiment classification*. Neurocomputing **131**, 312 (2014).

[20] O. Zoidi, E. Fotiadou, N. Nikolaidis, and I. Pitas. *Graph-based label propagation in digital media: A review*. ACM Computing Surveys (CSUR) **47**(3), 48 (2015).

[21] D. Rao and D. Ravichandran. *Semi-supervised polarity lexicon induction*. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 675–682 (Association for Computational Linguistics, 2009).

[22] Y. Ren, N. Kaji, N. Yoshinaga, M. Toyoda, and M. Kitsuregawa. *Sentiment classification in resource-scarce languages by using label propagation.* In *PACLIC*, pp. 420–429 (2011).

[23] X. Glorot, A. Bordes, and Y. Bengio. *Domain adaptation for large-scale sentiment classification: A deep learning approach*. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 513–520 (2011).

[24] X. Hu, J. Tang, H. Gao, and H. Liu. *Unsupervised sentiment analysis with emotional signals*. In *Proceedings of the 22nd international conference on World Wide Web*, pp. 607–618 (ACM, 2013).

[25] G. Paltoglou and M. Thelwall. *Twitter, myspace, digg: Unsupervised sentiment analysis in social media*. ACM Transactions on Intelligent Systems and Technology (TIST) **3**(4), 66 (2012).

[26] D. M. Blei. *Probabilistic topic models*. Communications of the ACM **55**(4), 77 (2012).

[27] S. Branavan, H. Chen, J. Eisenstein, and R. Barzilay. *Learning document-level semantic properties from free-text annotations*. Journal of Artificial Intelligence Research **34**, 569 (2009).

[28] Y. Jo and A. H. Oh. *Aspect and sentiment unification model for online review analysis*. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 815–824 (ACM, 2011).

[29] Y. H. C. Lin, W. Gao, and K.-F. Wong. *Tracking sentiment and topic dynamics from social media* (2012).

[30] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. *Sentiment analysis of twitter data*. In *Proceedings of the workshop on languages in social media*, pp. 30–38 (Association for Computational Linguistics, 2011).

[31] P. Chikersal, S. Poria, E. Cambria, A. Gelbukh, and C. E. Siong. *Modelling public sentiment in twitter: using linguistic patterns to enhance supervised learning*. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 49–65 (Springer, 2015).

[32] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao. *Target-dependent twitter sentiment classification*. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 151–160 (Association for Computational Linguistics, 2011).

[33] A. Go, R. Bhayani, and L. Huang. *Twitter sentiment classification using distant supervision*. CS224N Project Report, Stanford **1**, 12 (2009).

[34] L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, and K. Xu. *Adaptive recursive neural network for target-dependent twitter sentiment classification.* In *ACL (2)*, pp. 49–54 (2014).

[35] A. Montejo-Ráez, E. Martínez-Cámara, M. T. Martín-Valdivia, and L. A. Ureña-López. *A knowledge-based approach for polarity classification in twitter*. Journal of the Association for Information Science and Technology **65**(2), 414 (2014).

[36] J. Si, A. Mukherjee, B. Liu, Q. Li, H. Li, and X. Deng. *Exploiting topic based twitter sentiment for stock prediction.* ACL (2) **2013**, 24 (2013).

[37] F. Xianghua, L. Guo, G. Yanyan, and W. Zhiqiang. *Multi-aspect sentiment analysis for chinese online social reviews based on topic modeling and hownet lexicon*. Knowledge-Based Systems **37**, 186 (2013).

[38] B. Xiang, L. Zhou, and T. Reuters. *Improving twitter sentiment analysis with topic-based mixture modeling and semi-supervised training.* In *ACL (2)*, pp. 434–439 (2014).

[39] Y. Rao, Q. Li, X. Mao, and L. Wenyin. *Sentiment topic models for social emotion mining*. Information Sciences **266**, 90 (2014).

[40] G. Bello-Orgaz, J. Hernandez-Castro, and D. Camacho. *Detecting discussion communities on vaccination in twitter*. Future Generation Computer Systems (2016).

[41] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. *From tweets to polls: Linking text sentiment to public opinion time series.* ICWSM **11**(122-129), 1 (2010).

[42] D. Surian, D. Q. Nguyen, G. Kennedy, M. Johnson, E. Coiera, and A. G. Dunn. *Characterizing twitter discussions about hpv vaccines using topic modeling and community detection*. Journal of Medical Internet Research **18**(8), e232 (2016).

[43] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, vol. 6 (Springer, 2013).

[44] Y. Tan and Y. Shi. *Data Mining and Big Data: First International Conference, DMBD 2016, Bali, Indonesia, June 25-30, 2016. Proceedings*, vol. 9714 (Springer, 2016).

[45] P. P. Talukdar and F. Pereira. *Experiments in graph-based semi-supervised learning methods for class-instance acquisition*. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 1473–1481 (Association for Computational Linguistics, 2010).

[46] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell. *Text classification from labeled and unlabeled documents using em*. Machine learning **39**(2-3), 103 (2000).

[47] D. Q. Nguyen. *jLDADMM: A Java package for the LDA and DMM topic models*. http://jldadmm.sourceforge.net/ (2015).