National College of Ireland

Sentiment Analysis of Medicine Reviews using Ensemble models

MSc Research Project
Data Analytics

Prashanth Avverahalli Ramesha
x16137591

School of Computing
National College of Ireland

Supervisor: Dr. Keith Maycock

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Prashanth Avverahalli Ramesha
Student ID:	x16137591
Programme:	Data Analytics
Year:	2016
Module:	MSc Research Project
Lecturer:	Dr. Keith Maycock
Submission Due Date:	11/12/2017
Project Title:	Sentiment Analysis of Medicine Reviews using Ensemble models
Word Count:	5877

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	10th December 2017

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Sentiment Analysis of Medicine Reviews using Ensemble models

Prashanth Avverahalli Ramesha
x16137591
MSc Research Project in Data Analytics

10th December 2017

Abstract

The need to analyze user generated data over the web has recently gained importance due to the abundance of knowledge which can be acquired by careful analysis of such data. Majority of such data is available via online networking websites like Facebook, Twitter, LinkedIn, etc. The data available in such platforms are in the form of opinions and reviews of products, movies, medications, hotels, etc. Mining and analyzing of such data has become an important aspect for the companies to understand the people's opinion on a particular subject. There has been enough research done in the application of sentiment analysis across domains like product reviews, movies, hotels, etc. However, utilization of such methodologies in the field of medicine has to be given more importance as there are several studies conducted by United States Food and Drug Administration on the effects of adverse drug reactions on patients. Studying the effects of commonly used drugs on patients is important for the pharmaceutical companies to understand the positive and negative effect of drugs on the patients. The motive of this research project is to apply machine learning models for the sentiment analysis of reviews posted by patients to determine the polarity of opinion expressed in the reviews which can be positive or negative.

1 Introduction

Sentiment analysis is the process of categorizing the opinion expressed in the form of text to determine the individual's attitude on a particular subject. It has been applied across various fields. In the field of finance, Chan and Chong (2017) proposed a linguistic approach to perform phrase level analysis on financial texts. In the field of e-commerce, Liu et al. (2017) proposed a model which utilizes fuzzy set theory with sentiment analysis for the ranking of products based on their reviews. In the field of infrastructure, Estévez-Ortiz et al. (2016) employed sentiment analysis to analyze the opinion of people expressed in social networking sites, micro-blogging, crowd sourcing platforms and multimedia platforms to study their attitude towards the local government of smart cities. Ali et al. (2017) proposed fuzzy-ontology based sentiment analysis techniques and semantic web rule language decision making to address traffic congestion problems.

1.1 Importance of sentiment analysis in medical field

There are several studies focusing on adverse drug reactions on patients. There have been multiple identified cases where individuals were being hospitalized because of the unexpected responses from the medications prescribed to them. According to the study conducted by Jolivot et al. (2016) on a subset of patients admitted to ICU of a medical firm in France between February 2013 and 2014, 23.3% of 743 admissions to ICU were due to adverse drug reactions. Out of which 13.7% cases were preventable adverse drug reactions which resulted in total of 528 days of hospitalization in ICU incurring the costs of about €747,000. Although pharmaceutical organizations test the impact of medications before releasing them to the public, some drugs show undesirable effects only just when they are consumed for longer periods. Such effects are not captured by the drug manufacturers.

In a meta-analysis conducted by Miguel et al. (2012) on several health-related databases consisting of data on 18,818 patients, about 16% of patients incur adverse drug reactions during hospitalization. In a survey conducted by Gavaza et al. (2011) on the pharmacists of Texas reported that about 67.9% of the pharmacists never reported the adverse drug reactions to the FDA and about 65.7% pharmacists lacked the knowledge of reporting procedures on such events.

A study performed by Kumar (2017) stressed upon the importance of pharmacovigilance which is defined by World Health Organization as "The science and activities related to the detection, assessment, understanding and prevention of adverse effects and other drug-related problems" (Organization et al.; 2002). The author argued that it is necessary to study the adverse effects caused by medications and how it impacts the patients. According to the 8 year study conducted by Shepherd et al. (2012), the number of deaths due to adverse drug reactions steadily increased between 1999 to 2006 with an average death rates ranging from 0.08 to 0.12 per 100,000 people and trended upward for 8 year period.

Additionally, there are numerous health-related forums and blogs such as askapatient[1], WebMD[2], drugs.com[3], druglib.com[4] which act as a platform for the patients to post their experiences on medications. Such information carry valuable insights which can be analyzed to determine the side effects of drugs on different patients with different health conditions and improve the drug accordingly.

1.2 Background

Ensemble models have been used to perform sentiment analysis of reviews on products of e-commerce websites.Many studies have reported getting better classification accuracy by using such models. However, there are only a few studies which utilize ensemble models to analyze the reviews on medications.

Ensemble modeling is a technique of weighing individual opinions and combining them to arrive at a final decision (Polikar; 2006). These techniques have been successful in improving the accuracy of machine learning models by training several individual classifiers and combining them to improve the overall predictive power of the model. They utilize several weak classifiers by combining them in some manner to obtain the result either by performing weighted average or majority voting of the individual classifiers to improve the overall accuracy of the model when compared to the accuracy obtained by using a single classifier.

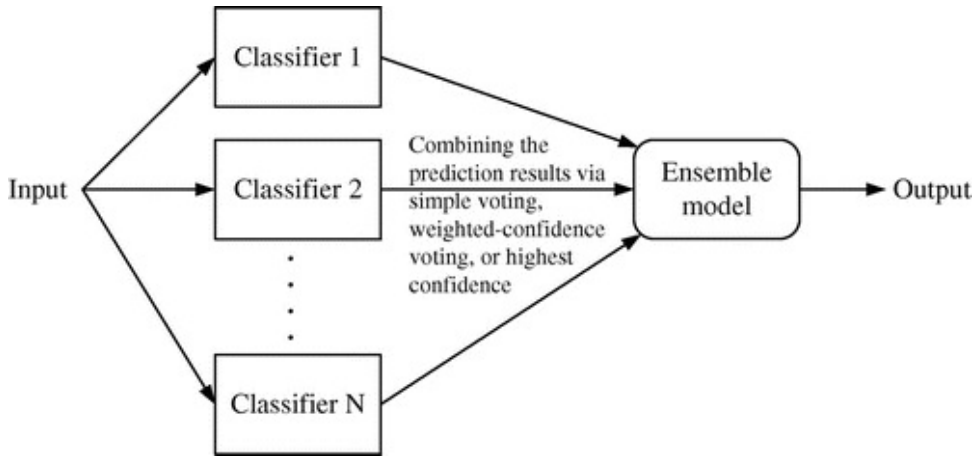

Figure 1: Illustration of Ensemble model by Chou and Lin (2012)

This research project aims at employing ensemble models in the sentiment analysis of reviews on medications and determine which combination of machine learning models give the best performance in terms of classification accuracy.

[1] http://www.askapatient.com/
[2] http://www.webmd.com/
[3] https://www.drugs.com/
[4] http://www.druglib.com/

2 Related Work

There are few studies which focus on applying sentiment analysis techniques using machine learning models. Whitehead and Yaeger (2009) proposed cross domain sentiment mining model where a model trained in a particular domain can be used as a classifier for the different domain. Data in the form of reviews for different subjects like camera, laptops, summer camps, lawyers, drugs, radio, restaurant and television were chosen for the study. Support Vector Machine (SVM) was used as a base classifier. To test the performance of classifier trained in one domain and its classifying power in another domain, a new SVM classification model was trained for each model and was used to test the model on the rest of the dataset with a K-fold of 25. The classification accuracy was calculated as the average of K-fold tests. By building two different ensembles where one used simple majority vote of its component models for each new classification and another used weighted majority votes, the authors demonstrated that it is possible to improve the accuracy by selecting the cross-domain models with lexicons similar to target domain lexicon. This work demonstrated that it is possible to deploy a model trained in one domain to a different domain and achieve an acceptable accuracy in classifying the sentiment.

In a similar study involving the utilization of ensemble models, Whitehead and Yaeger (2010) performed sentiment analysis on the dataset incorporating the reviews on cameras, restaurants, laptops, etc. employing a single SVM and bagging, boosting, random subspace and bagging random subspace methods which used SVM as their base model for bench-marking. K-fold cross validation with 10 folds was done to acquire the result. The results indicated ensemble methods as the better performers compared to single base classifier in terms of accuracy.

Ali et al. (2013) conducted sentiment analysis on reviews posted on hearing loss forums using naive bayes, SVM , logistic regression with lemmatization. Further analysis was carried out using traditional bag of words approach. A bag of words approach is the process of extracting features from the text, it involves extraction of words and their frequency of occurrence in text. The words are assigned a score based on their polarity. The two methods were compared; in each of the test cases, machine learning algorithms outperformed the bag of words approach. SVM gave the best performance with the overall agreement of kappa 0.64 which is a good agreement with the model according to the study by Landis and Koch (1977). This study showcased that machine learning methods can perform better when compared to existing traditional sentiment analysis methods.

A study conducted by Na and Kyaing (2015) proposed a linguistic approach in clause-level sentiment analysis of drug reviews on the WebMD forum. This method divided each sentence into dependent and independent clauses. Stanford parts of speech parser (De Marneffe et al.; 2006) was employed to construct a tree like structure from the clauses obtained. The extracted data from the website was manually labelled as positive, negative or neutral. For bench-marking purpose, a SVM-1 with bag of words approach and SVM-2 with additional linguistic bi-gram feature were developed. Although the accuracy of the linguistic approach was better than that of single SVM model, they encountered issues in the form of misclassified clauses, lexicon error, metamap error, user text error, etc. Additionally, separating the misclassified clauses and tagging of classifiers increased

the manual workload.

Mishra et al. (2015) performed analysis on medications by building a model which captures and incorporates the common issues faced by patients while on medication by crawling through the website. It performed sentiment analysis on such reviews at an aspect level using corpus like MedDRA (Wood; 1994) and SIDER (Kuhn et al.; 2010) to gather a list of medical terminologies. Clustering was employed to group the reviews based on the frequency of their occurrence. The sentiment analysis performed with SVM as the base classifier could only achieve the accuracy of 59% which was low due to the imbalance in the polarity of reviews as there were more number of negative than positive reviews. Also, the size of the data used for analysis was also small which resulted in model achieving low accuracy.

Siddiqua et al. (2016) proposed a rule based classifier by training the model using Bag of Words along with probabilistic Naive Bayes using 1.2 million tweets acquired from the Stanford twitter dataset (Go et al.; 2009) they concluded that employing feature selection with rule based classifier can have significant improvement in the accuracy of classifying the sentiment.

In a study conducted by (Jianqiang; 2016) where in prior polarity scores and n gram features were combined as a feature set of the tweets. This feature set was utilized for an ensemble of classifiers including SVM, Logistic regression and random forest. This model was benchmarked with n gram as the baseline. Experimental results showed ensemble models performed better with accuracy of 86% with logistic regression thus indicating that ensemble models perform better compared to single baseline model.

Salas-Zárate et al. (2017) proposed an model which incorporated aspect based sentiment analysis based on ontological methods on diabetic disorders, the proposed method involved the accumulation of tweets on diabetes which were manually tagged as either positive, negative or neutral by a group of experts which were benchmarked with the aspect level sentiment analysis conducted using N-gram before, N-gram around and N-gram after methods where: N-gram before involved extraction of n grams before the aspect, N-gram after involved extraction of n grams after the aspect and N-gram around involved the extraction of n grams before and after the aspect. The results obtained rated N-gram around as the most efficient model compared to the rest. The downside of this approach involved manual tagging of large number of tweets.

In a survey conducted by Hadi et al. (2017), the authors propose that pharmacists have a noteworthy part in detailing Adverse Drug Reactions (ADR) and investigated the present situations of ADR across several nations and led a survey of various situations experienced by the drug specialists. They conclude that information gap is the real explanation behind a few ADR not being accounted for by the pharmacists.

Perikos and Hatzilygeroudis (2017) employed naive bayes, maximum entropy and SVM along with an ensemble of classifiers comprising of the above three algorithms to perform aspect based sentiment analysis of tweets. They proposed a combination of features such as bag of words, bag of words with parts of speech tagging, dependency tree, parts of speech tagging and dependency tree in classifying the polarity of tweet. Two

thirds of data were used to train the classifier and the remaining data was supplied as test data to measure the performance. Majority voting was used to extract the result where the prediction made by majority of algorithms is considered as the final classification. Ensemble classifier was a better performer with 5.8% greater accuracy than the single base classifier which was SVM. This proposal represented how the combination of bag of words, parts of speech tagging and dependency tree features can improve the classification of the model by improving the accuracy.

3 Methodology

The proposed approach for the analysis is shown in Figure 2. Implementation of the analysis was done using RapidMiner [5] software.

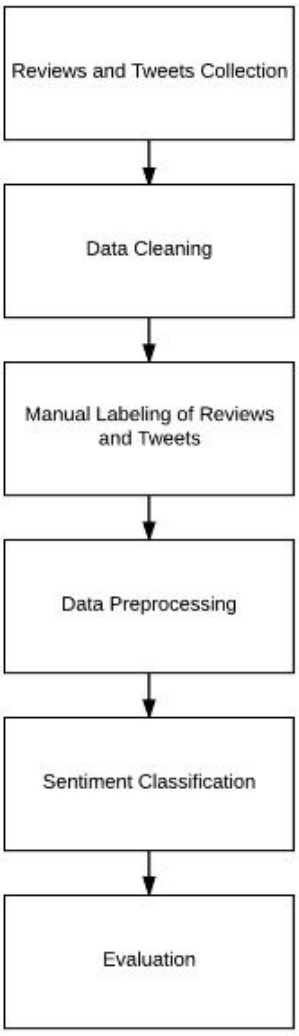

Figure 2: Proposed approach

[5] https://rapidminer.com/

The data in the form of reviews were scraped from the WebMD forum using Octoparse tool [6]. The scraped data consisted of reviews of 70 commonly used medications. Initial data consisted of about 10000 reviews posted by patients on medications. For the second analysis, about 80,000 tweets on the drug xanax were collected.

3.1 Data Cleaning

About 3500 reviews were chosen from the medications. To simplify the process of manually labeling the reviews, lengthy reviews were removed. Also, reviews containing irrelevant information such as questions and suggestions about the drug, blank reviews and repeated reviews were removed.

Twitter data was cleaned by removing duplicate tweets, tweets with links containing the sale of drugs online and empty tweets. The cleansed data contained about 20,000 tweets.

3.2 Data Labeling

This phase involved carefully going through the reviews and labeling them as either positive or negative. Each review contained the effectiveness rating, ease of use rating, satisfaction rating and the patient's comment on the drug, where: effectiveness is whether the drug has worked for the patient, ease of use is how easy it is to use the drug and satisfaction is whether the patient has had a good experience with the drug.

Firstly, the entire review was gone through to get an understanding of whether if it was positive or negative. Then, the review was labelled with respect to the satisfaction rating as effectiveness and ease of use rating did not contribute in categorizing the reviews.

From an analysis of the reviews it was found out that in some cases, patients who had a negative experience with the drug had given a positive rating. Also, patients who had a positive experience with the drug rated it poorly. Na and Kyaing (2015) and Yalamanchi (2011) also reported these issues in their study as this is a common phenomena while mining review data. Such instances are called opinion spam, this concept was first introduced by Jindal and Liu (2008). There are few studies which have focused on identifying opinion spam. Chen and Chen (2015) proposed binomial regression model in identifying the product reviews which have anomalous proportions deviating from the majority opinion. Lin et al. (2014) proposed supervised technique coupled with threshold based solution in the identification of opinion spam on product reviews. In this analysis, such reviews were labelled solely on the comment made by the patient.

To provide the training data for machine learning models, 2400 reviews were manually labeled with equal number of positive and negative reviews. As there were no readily available labelled dataset on the medicine domain, a subset of tweets collected were manually labelled for the purpose of training the machine learning models. The training

[6] `https://www.octoparse.com/`

data for twitter analysis of xanax consisted of 1585 tweets with 629 negative, 790 neutral and 166 positive tweets.

3.3 Data Preprocessing

For the purpose of training the machine learning model, The data had to undergo several preprocessing steps. They were:

Nominal to Text: In order to perform text processing, the data was converted from nominal to text values.

Transforming cases: The entire reviews were converted to lower case to avoid ambiguity.

Tokenization: The text in the reviews were broken down into constituent words with each word being termed as a 'token'. Tokenization removes unwanted symbols and punctuation marks which have no meaning. Also, tokenizing is done to construct a Term frequency - Inverse document frequency (TF-IDF) dictionary for constructing n grams. Term frequency is the number of a times a word appears in a document and inverse document frequency is how much information that a word provides and is generally calculated as logarithmic quotient of total number of documents by the number of documents containing the word.

Filtering Stopwords: Words such as a, an, the, if, etc. were removed as they do not carry any value in the analysis.

Filtering Tokens: Tokens with length lesser than three and greater than twenty were filtered out to remove unwanted characters.

Stemming: The remaining words were converted to their root word or stemmed. For example, running will be converted into run, eating will be converted into eat. The purpose of stemming is to group similar words together by converting them into their root word. Stemming process reduces the size of vocabulary and hence decreases the redundancy of word occurrence. For this analysis, porter stem algorithm (Porter; 1980) was used as it is widely employed stemming algorithm for sentiment analysis.

3.4 Choosing the Model

The below machine learning algorithms were used to build the model as Whitehead and Yaeger (2009), Sharma and Dey (2013), Perikos and Hatzilygeroudis (2017), and Siddiqua et al. (2016) successfully employed these algorithms in their work which involved analysis of similar data.

1. **Support Vector Machine**

Support Vector Machine (Cortes and Vapnik; 1995) is a supervised learning algorithm used for classification and regression. It is a non probabilistic binary classifier which builds a model and assigns examples to one of the two categories. It works by building two parallel hyperplane separating the two classes of data such that the distance between them is large as possible. The region bounded by the two planes is called as margin.

2. **Decision Tree**

Decision Tree (Quinlan; 1986) is also a supervised learning algorithm used for classification. It works by creating a model to predict the value of target variable based on the input variable. There are two types of decision trees: classification and regression tree. Classification tree works on categorical variables and regression trees predicts the outcome which can be real number. It works by splitting the classification set into subsets in a recursive manner called recursive partitioning. The process is stopped when splitting no longer contributes to predictions and the subset at a node has all the values of a target variable.

3. **Random Forest**

A Random Forest (Liaw et al.; 2002) is an ensemble learning algorithm which builds multiple decision trees by selecting random subsets of training data. It overcomes the problem of over-fitting by decision trees. It works by averaging different decision trees built on different subsets of training data, hence reducing the variance. However, there can be an increase in bias and interpretability as a trade off for increased accuracy.

4. **Naive Bayes**

Naive Bayes (McCallum et al.; 1998) is a family of classification algorithms which work on the same principle that every feature to be classified is independent of the value of rest of the features. Each feature contributes independently to the probability of a particular class regardless of the correlation between the features. It is primarily used in document classification applications.

5. **Generalized Linear Model**

Generalized Liner Model (McCullagh; 1984) is an extension of liner model where dependent variable is linearly related to the factors as well as the covariates via link function. Unlike linear regression, it allows the dependent variables to follow non normal distribution. The algorithm has three components: random component, systematic component and link function. Random component refers to the probability distribution of response variable, systematic component specifies the linear combination in explanatory variables and link function specifies the link between random and systematic components.

6. **Adaptive Boosting (AdaBoost)**

AdaBoost(Freund and Schapire; 1995) is a machine learning algorithm which is used in conjunction with other algorithms to increase the classification accuracy. Initially, a base algorithm is chosen and the training examples are assigned equal weights. After each iteration, the weights of incorrectly classified examples are increased and the process is repeated. The resulted obtained is the weighted sum of the learners. This algorithm is very useful in training weak classifiers.

7. **Bootstrap Aggregation(Bagging)**

Bagging (Breiman; 1996) decreases the variance in prediction by generating different subsets of training data with same cardinality. Some subsets can contain repetitions of examples. Different models are built on different random subsets of data. Finally, voting mechanism is employed to get the final prediction from all the models.

4 Implementation and Results

In order to test the models, Cross validation was chosen as they split the training data into chunks of equal length and after each iteration, different splits of data are chosen as training and test data. The purpose of choosing K-fold cross validation over split validation is because it overcomes high variance by splitting the training dataset into k subsets. Also, every subset is chosen as the test data at some point over different iterations thereby reducing the variance.

For the model built for the analysis, K-fold cross validation with the K values 6,8,10,12,15 and 20 were applied and models were tested. The accuracy obtained for the value K=10 was higher compared to other values of K.

For sampling the data, three different sampling types were tested: linear sampling, shuffled sampling and stratified sampling. Linear sampling divides the training data into partitions without changing the order. Shuffled sampling builds random subsets of training data and stratified sampling also builds the random subsets of training data with an additional factor that it maintains the class balance in every subset.

Out of three sampling types, shuffled sampling gave the best results and hence it was chosen.

The results obtained from the respective models for 2400 reviews with 10 fold cross validation and shuffled sampling are reported in terms of accuracy, precision and recall. They are defined below:

Accuracy: It is the ratio of correctly predicted observations to the total total observations and is calculated by the formula:

Accuracy(A) :

$$\frac{tp + tn}{tp + tn + fp + fn} \tag{1}$$

Precision: It is the ratio of correctly predicted number positive classes to the total number of positive predictions and is calculated as:

Precision(p) :

$$\frac{tp}{tp + fp} \tag{2}$$

Recall: It is the ratio of correctly predicted positive classes to all the observations in the positive class and is calculated as:

Recall(r) :

$$\frac{tp}{tp + fn} \tag{3}$$

F1-Score: It is also a measure of model's performance, it provides a balance between precision and recall by calculating their weighted harmonic mean. It is calculated as:

F1-Score(F) :

$$\frac{2(p)(r)}{p + r} \tag{4}$$

Where: **tp** is the correctly predicted positive review, **tn** is the correctly predicted negative review, **fp** is when the actual review is negative but predicted as positive and **fn** is when the actual review is positive but predicted as negative respectively.

Algorithm	Accuracy(%)	Precision(%)	Recall(%)
Support Vector Machine	70.45	75.17	61.02
Random Forest	47.31	47.74	60.29
Decision Tree	52.85	67.65	39.18
Naive Bayes	62.86	62.74	61.79
Generalized Linear Model	73.37	72.72	74.85
Ensemble(SVM,GLM, Decision Tree)	70.57	75.49	63.64
GLM(AdaBoost)	69.57	68.66	71.86
Bagging(GLM)	**74.65**	**74.13**	**75.9**

Table 1: Results for Sentiment Classification of Review Data

As seen in Table 1 it is evident that Generalized Linear model with bootstrap aggregation has the best performance with the accuracy of 74.65% and F1-Score of 0.75.

The second part of this research aimed at answering the question: *Is it possible to make use of tweets on medications instead of reviews to determine the sentiment of people on different medications?*

For this purpose, tweets on the drug xanax were collected and sentiment analysis was performed. The analysis was limited to the drug xanax as it was not possible to retrieve suitable number of tweets for other drugs to perform an analysis.

The training data consisted of 1585 tweets with 629 negative tweets, 790 neutral tweets and 166 positive tweets. The number of positive tweets were relatively small due to the death of a celebrity rapper of xanax overdose which resulted in majority of negative tweets condemning the medication. Cross validation was performed by choosing different values of K (6, 8, 10, 12, 15 and 20). The accuracy obtained for 10 folds was greater than rest of the values of K. The results obtained are reported in terms of Accuracy and Kappa Statistic since this analysis involved multiclass classification.

Algorithm	Accuracy(%)	Kappa
Support Vector Machine	66.47	0.387
Random Forest	50.00	0.004
Decision Tree	50.94	0.042
Naive Bayes	54.73	0.276
Generalized Linear Model	66.16	0.368
Ensemble(SVM,GLM, Naive Bayes)	**66.72**	**0.394**
SVM(AdaBoost)	66.28	0.384
Bagging(SVM)	65.97	0.376

Table 2: Results for Sentiment Classification of Tweets

As seen in Table 2, the ensemble model consisting of SVM, GLM and Naive Bayes achieved the highest accuracy of 66.72% with Kappa statistic of 0.394 which is close to 0.4 signifying a moderate agreement with the model in accordance with Landis and Koch (1977) .

5 Evaluation

The models with highest classification accuracy for patient reviews and twitter data were chosen to classify the reviews and tweets of xanax respectively. Visualization was done using Tableau [7]. Initial analysis revealed that there were more female reviews than compared to males.

5.1 Case Study 1

In this case study, the predicted polarity of the xanax reviews were categorized by the age group and gender.

As seen from Figure 3, we can observe that there are more reviews for the females with age range 35-44 and males with the age range 45-54. Also, there are more negative reviews from females compared to males with an average of 15.5 negative reviews among males and 28.8 negative reviews for females across all the age groups. Females with age group 13-18 are significantly higher than males in terms of usage of the drug.

[7] http://www.tableau.com/

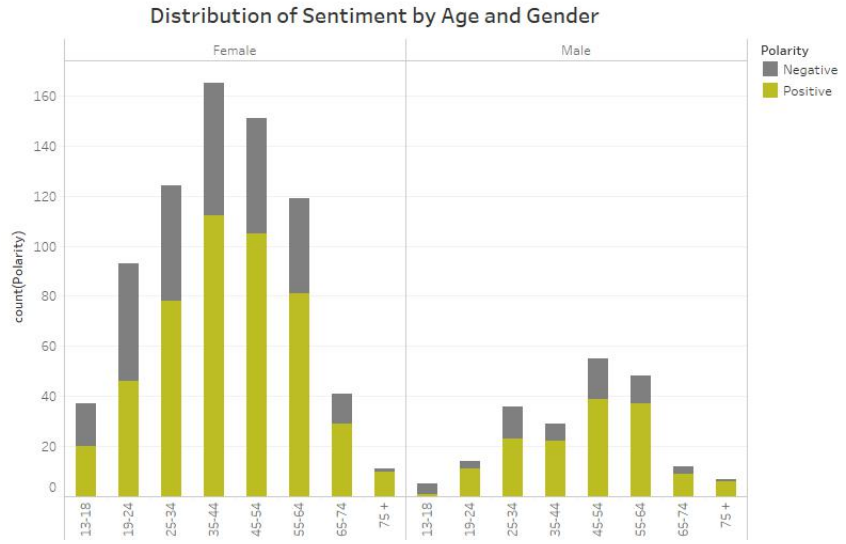

Figure 3: Distribution of polarity by Age and Gender

According to a study done by Woodlock (2005), females make up 70% of prescribed anti depressant consumption and are suffering from mental distress at twice the rate of males. The author also stated that marketing of antidepressant drugs are primarily aimed at female audience. In another study done by Green et al. (2009), the author reported that women were more likely to report the use of prescription opioid (29.8% females vs. 21.1% males) and abuse of any prescription opioid (15.4% females vs. 11.1% males). The data for this study was collected from the Addiction Severity Index Multimedia Version Connect (ASI-MV Connect) database between November 2005 to April 2008. These studies support the gender bias in the reviews.

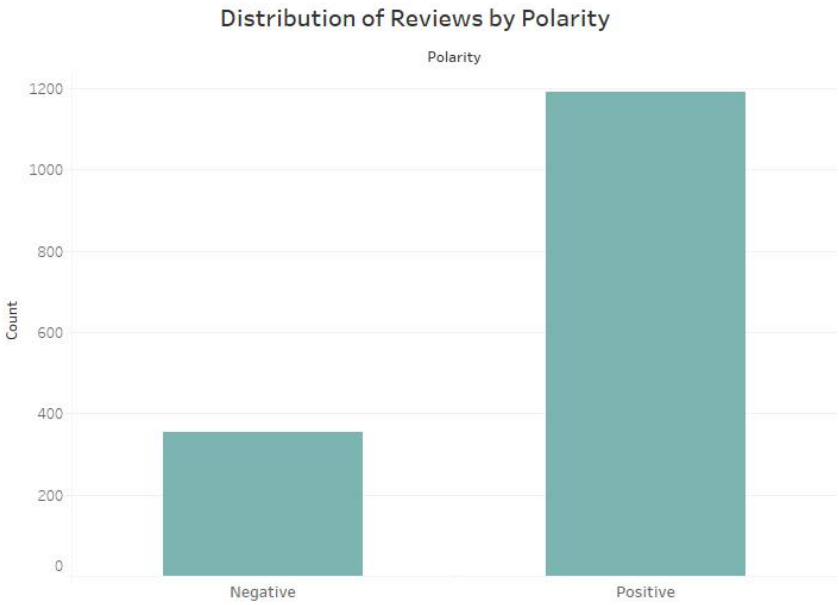

Figure 4: Distribution of Polarity

As seen from Figure 4, the predictions obtained from the model had comparatively more positive than negative reviews. The average rating of xanax in WebMD forum is 4.47 out of 5 suggesting that it is a helpful drug for the patients. For the purpose of verification, the ratings for xanax was recorded across different web forums. They were: 3.9 out of 5 in askapatient website, 8.19 out of 10 in druglib website, 8.9 out of 10 in drugs.com website indicating that patients see this as an useful medication across web forums.

5.2 Case Study 2

For this case study, the distribution of polarity was analyzed by the different conditions the medications was used.

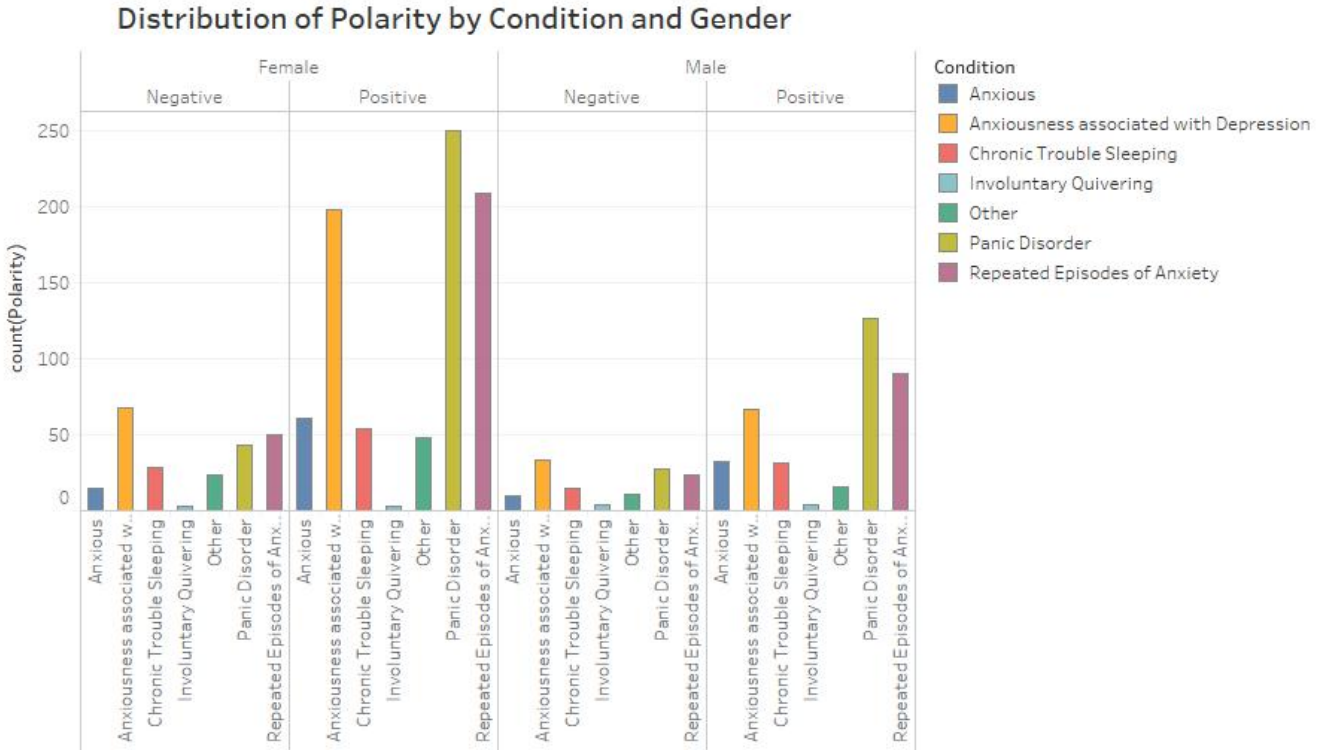

Figure 5: Distribution of Polarity by Condition and Gender

As seen in Figure 5, the distribution of sentiment is similar for both genders for different conditions with more cases of anxiousness associated with depression, Panic disorder and Repeated episodes of anxiety. From this distribution we can infer that most of the patients irrespective of gender take xanax for anxiety related issues. According to the study done by Stahl (2002), despite the limitations of benzodiazepines, they are still widely used along with serotonergic antidepressants for the treatment of anxiety disorders. Xanax was the most prescribed drug for anxiety with 31 million prescriptions and was the top prescribed medication for anxiety. Benzodiazepines are class of agents that directly affect central nervous system and are usually used in the treatment of anxiety, sleep disorder, seizures and panic disorder. Xanax is a part of the list of benzodiazepines and is the most commonly prescribed medication for anxiety.

5.3 Case Study 3

In this case study, the distribution of reviews with respect to the condition and age were analyzed.

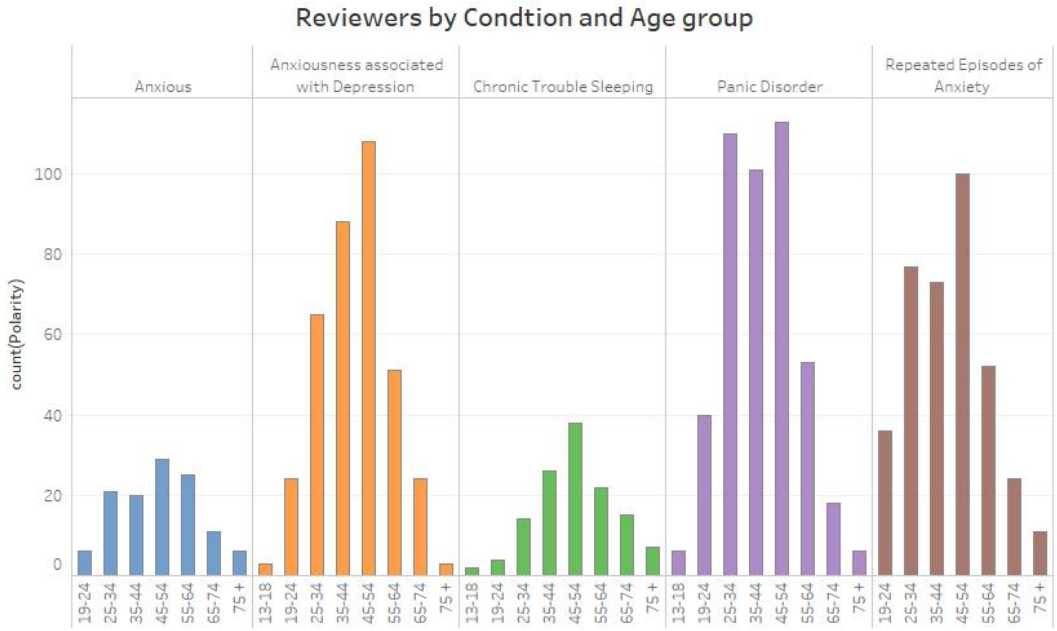

Figure 6: Reviewers by Condition and Age group

As seen from Figure 6, for all the conditions, the age group 45-54 has most number of reviews. However, there are significantly more reviewers for the age group 25-34 and 35-44 for Panic disorder and repeated episodes of anxiety.

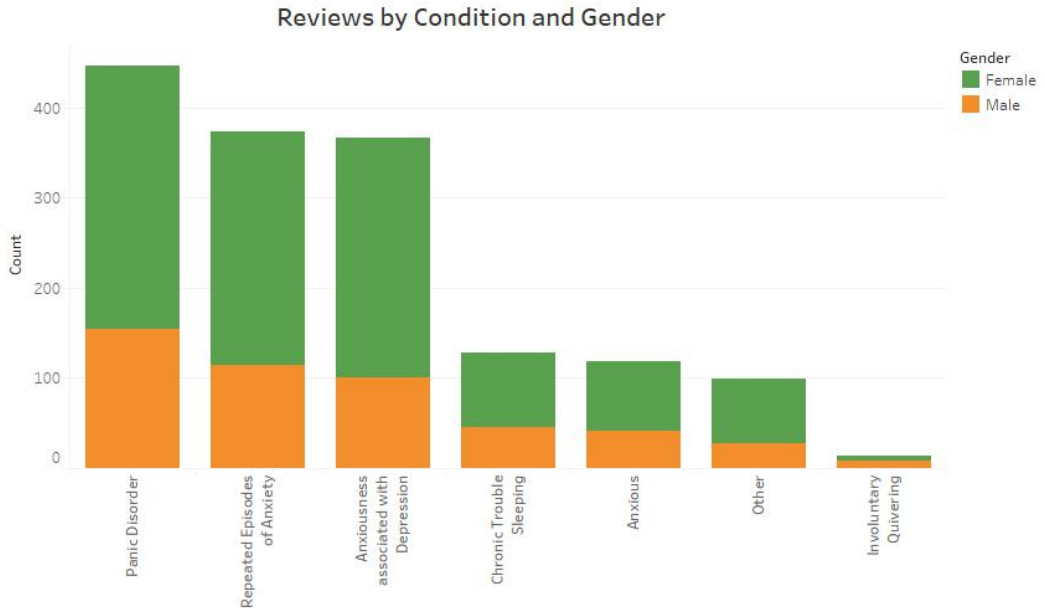

Figure 7: Reviewers by Condition and Gender

From Figure 7, we can see that there are more cases in panic disorder and anxiety related issues, females have more number of reviews compared to males in such cases. According to Anxiety and Depression Association of America (Skarl; 2015), 3.1% of the population of the United States suffer from Generalized Anxiety Disorder and it is observed as twice as common in women than men. Additionally, Enoch et al. (2003) in their study stated that women are more prone to anxiety than men. The study was conducted on two separate group of participants: 149 predominantly Caucasian individuals (92 women, 57 men), and 252 Plains American Indians (149 women, 103 men). They argued that lower activity in a particular genotype was associated with higher anxiety scores in women.

5.4 Case Study 4

This case study is the analysis done on the twitter data. The ensemble model consisting of GLM, SVM and Naive Bayes were used to classify about 20,000 tweets.

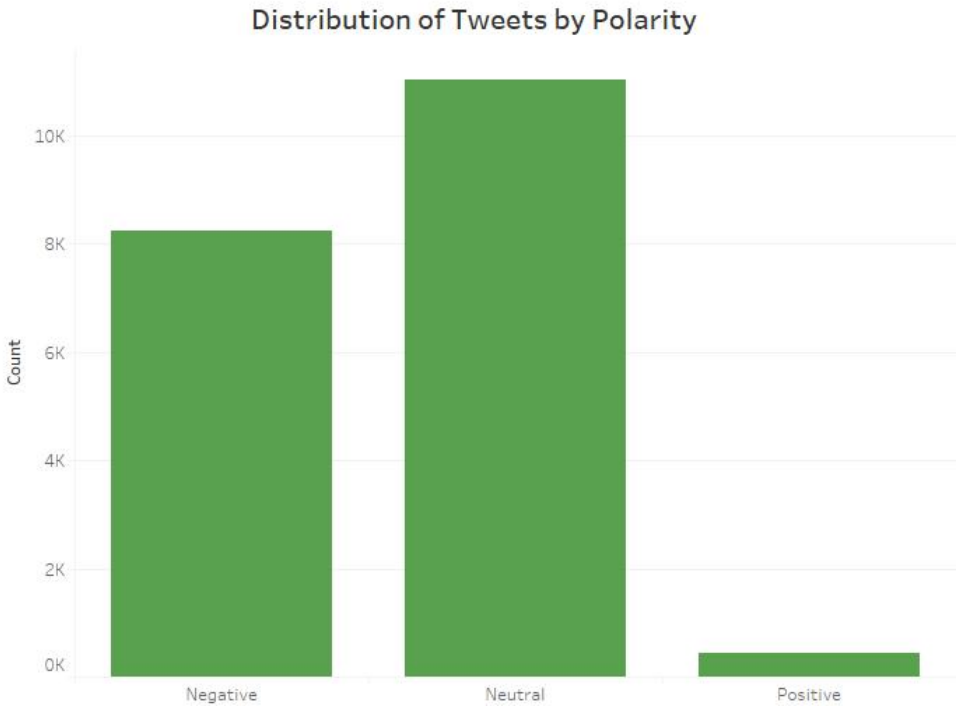

Figure 8: Distribution of Condition and Polarity

As seen in Figure 8, it is evident that there is a majority of negative tweets and very few positive tweets, this is due to the reason that the tweets were collected during the day at which a celebrity rapper died due to an overdose of xanax and hence many reviews had a negative flavor in them opposing xanax. As the most number of tweets were addressing the death of the celebrity, it was not possible to filter such tweets and retain sufficient tweets to perform analysis.

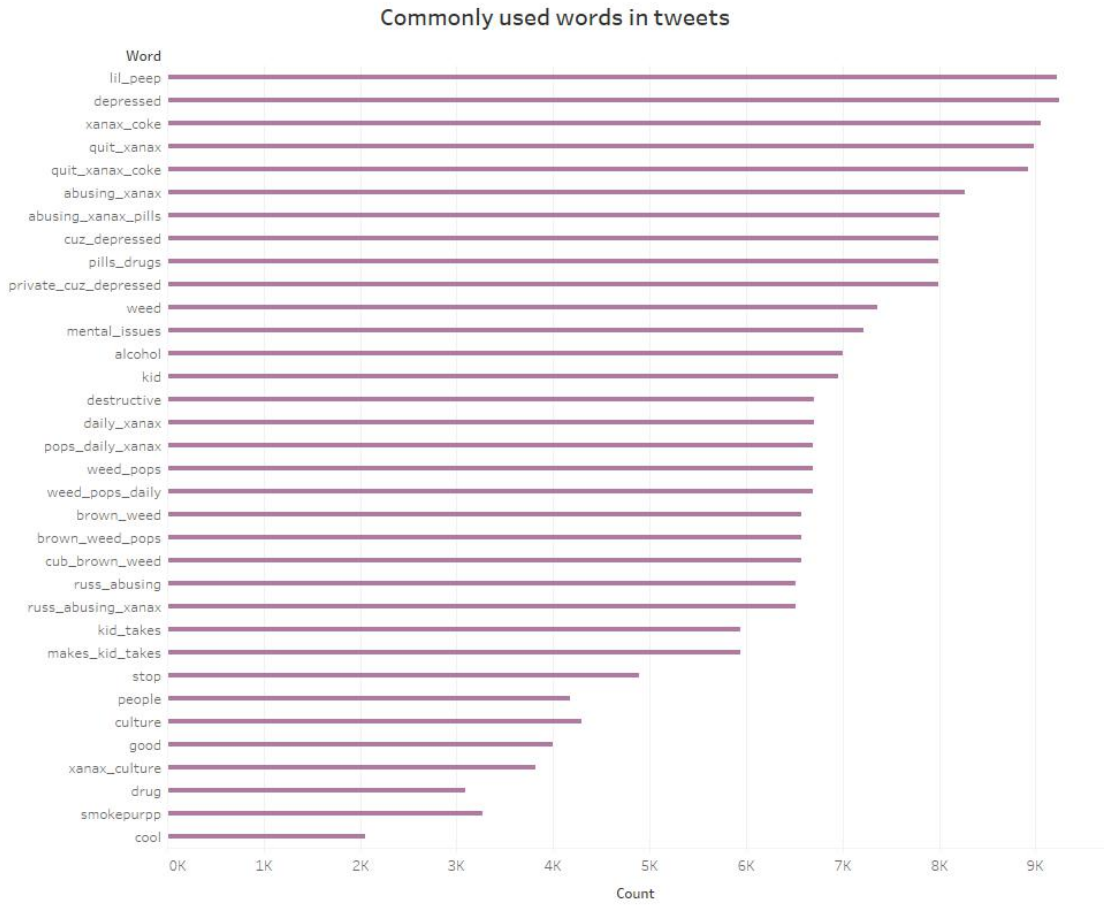

Figure 9: Commonly used words in tweets

The most commonly used words as seen in Figure 9 include lil peep which is the name of the deceased celebrity. There are other interesting words like abusing xanax, xanax_coke, pops_xanax_daily, makes_kid_take, xanax_culture which indicate that the drug is being abused along with other drugs and alcohol by the teenagers. According to the study done by Compton and Volkow (2006), there has been an increase in cases of prescription drug abuse. Also, they state that there were high rates of abuse among teenagers in United States. In another study, Rome (2001) reported that alcohol is the most abused drug with one in every 5 teenagers consuming it by the end of their high school, 52% by eighth grade. Also, marijuana is the second most widely used drug with 17% of 8th graders, 32% of 10th graders, and 38% of 12th grade students reported using it at least once.

Also, some of the tweets contained links which were discarded for the analysis. Upon further analysis it was found that those links referred to the sale of drugs online. This issue was addressed by Katsuki et al. (2015) in their study of twitter data to monitor the sale of illicit drugs online without prescription. From their analysis of 2,417,662 tweets, a document term matrix was generated to extract frequently appearing drugs in the tweets. Xanax was fourth in the list appearing in 36,486 tweets. The URLs captured from the tweets reported the sale of drugs including Ativan, Ambien, Lunesta, Valium, and Xanax. The authors argued that it necessary to analyze twitter data to identify and analyze the abuse of prescription drugs and promotion of non-medical use of prescription medications

(NUPM) online by the youth.

6 Conclusion and Future Work

This research project evaluated the application of machine learning approaches in sentiment analysis of medicine reviews and how ensemble models increase the accuracy of classifying the polarity of reviews by utilizing specialized learners. Different combination of machine learning models were tested to determine which models give the best accuracy.

This research has several findings:

1. There is a difference in opinion of people on the medication across different platform. This issue could not be properly addressed due to majority of tweets directed towards the death of the celebrity which resulted in more number of negative tweets on the drug. It was not possible to filter out such tweets as the volume of tweets were not sufficient to perform analysis.

2. Some reviewers rated the drug highly but had a bad experience on the drug while some reviewers gave a poor rating and their comment indicated a positive experience. Such reviews are called opinion spam. Na and Kyaing (2015) and Yalamanchi (2011) also reported about this issue in their study. Such instances can impact the study negatively.

3. More females experience anxiety related disorders than men. And xanax is mostly used to treat anxiety related issues. The studies done by (Skarl; 2015) , Enoch et al. (2003) and Stahl (2002) also reported these issues.

4. From the analysis of twitter data it was found that teenagers are abusing xanax along with other substances like alcohol, marijuana and coke. The studies done by Compton and Volkow (2006) and Rome (2001) also reported the abuse of prescription drugs by teenagers in United States.

There are several gaps that needs to be addressed as part of future work. They are:

1. Further investigation is required in the identification of opinion spam in the drug reviews. Existing works focus on identifying opinion spam across product review websites. More focus must be given on identifying such problems across medical domain.

2. This project aimed at finding out the relationship between the review data on web forums and the tweets on medications. However, due to the death of a celebrity of an overdose, it was not possible to correlate the relationship between reviews and tweets as there were more number of negative tweets addressing the abuse of xanax.

Further study is required to determine whether it is possible to use tweets on medications instead of drug reviews to study the people's opinion on medications. This is important because of the availability of twitter data and the people's willingness to share their thoughts across social media.

3. There is a need of surveillance of twitter data to monitor the abuse of medications and the sale of drugs online without prescription. The tweets which were eliminated from the analysis contained links to the sale of drugs online. This issue was addressed by Katsuki et al. (2015) in their study of twitter data to monitor the abuse of drugs. However, further research needs to be done to identify the sale of such drugs and to prevent youth from abusing prescription drugs for recreational purposes.

7 Acknowledgment

I would like to take this opportunity to thank my Supervisor, Dr. Keith Maycock who has provided constant support and guidance throughout the completion of the thesis. I am gratefully indebted to his valuable advice and feedback over the course of this research.

References

Ali, F., Kwak, D., Khan, P., Islam, S. R., Kim, K. H. and Kwak, K. (2017). Fuzzy ontology-based sentiment analysis of transportation and city feature reviews for safe traveling., *Transportation Research: Part C* **77**: 33 – 48.
URL: *http://ezproxy.ncirl.ie/login?url=http://search.ebscohost.com/login.aspx?direct=trueAuthTyp livescope=site*

Ali, T., Schramm, D., Sokolova, M. and Inkpen, D. (2013). Can i hear you? sentiment analysis on medical forums., *IJCNLP*, pp. 667–673.

Breiman, L. (1996). Bagging predictors, *Machine learning* **24**(2): 123–140.

Chan, S. W. and Chong, M. W. (2017). Sentiment analysis in financial texts., *Decision Support Systems* **94**: 53 – 64.
URL: *http://ezproxy.ncirl.ie/login?url=http://search.ebscohost.com/login.aspx?direct=trueAuthTyp livescope=site*

Chen, Y.-R. and Chen, H.-H. (2015). Opinion spam detection in web forum: a real case study, *Proceedings of the 24th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, pp. 173–183.

Chou, J.-S. and Lin, C. (2012). Predicting disputes in public-private partnership projects: Classification and ensemble models, *Journal of Computing in Civil Engineering* **27**(1): 51–60.

Compton, W. M. and Volkow, N. D. (2006). Abuse of prescription drugs and the risk of addiction, *Drug and Alcohol Dependence* **83**(Supplement 1): S4 – S7. Drug Formulation and Abuse Liability.
URL: *http://www.sciencedirect.com/science/article/pii/S037687160600055X*

Cortes, C. and Vapnik, V. (1995). Support-vector networks, *Machine Learning* **20**(3): 273–297.
URL: *https://doi.org/10.1007/BF00994018*

De Marneffe, M.-C., MacCartney, B., Manning, C. D. et al. (2006). Generating typed dependency parses from phrase structure parses, *Proceedings of LREC*, Vol. 6, Genoa Italy, pp. 449–454.

Enoch, M.-A., Xu, K., Ferro, E., Harris, C. R. and Goldman, D. (2003). Genetic origins of anxiety in women: a role for a functional catechol-o-methyltransferase polymorphism, *Psychiatric genetics* **13**(1): 33–41.

Estévez-Ortiz, F.-J., García-Jiménez, A. and Glösekötter, P. (2016). An application of people's sentiment from social media to smart cities., *El profesional de la información* **25**(6).

Freund, Y. and Schapire, R. E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting, *European conference on computational learning theory*, Springer, pp. 23–37.

Gavaza, P., Brown, C. M., Lawson, K. A., Rascati, K. L., Wilson, J. P. and Steinhardt, M. (2011). Texas pharmacists knowledge of reporting serious adverse drug events to the food and drug administration, *Journal of the American Pharmacists Association* **51**(3): 397–409a.

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

Green, T. C., Serrano, J. M. G., Licari, A., Budman, S. H. and Butler, S. F. (2009). Women who abuse prescription opioids: Findings from the addiction severity index-multimedia version connect prescription opioid database, *Drug and Alcohol Dependence* **103**(12): 65 – 73.
URL: *https://www.sciencedirect.com/science/article/pii/S0376871609000945*

Hadi, M. A., Neoh, C. F., Zin, R. M., Elrggal, M. E. and Cheema, E. (2017). Pharmacovigilance: pharmacists perspective on spontaneous adverse drug reaction reporting, *INTEGRATED PHARMACY RESEARCH AND PRACTICE* **6**: 91–98.

Jianqiang, Z. (2016). Combing semantic and prior polarity features for boosting twitter sentiment analysis using ensemble learning, *Data Science in Cyberspace (DSC), IEEE International Conference on*, IEEE, pp. 709–714.

Jindal, N. and Liu, B. (2008). Opinion spam and analysis, *Proceedings of the 2008 International Conference on Web Search and Data Mining*, ACM, pp. 219–230.

Jolivot, P.-A., Pichereau, C., Hindlet, P., Hejblum, G., Bigé, N., Maury, E., Guidet, B. and Fernandez, C. (2016). An observational study of adult admissions to a medical icu due to adverse drug events, *Annals of intensive care* **6**(1): 9.

Katsuki, T., Mackey, T. K. and Cuomo, R. (2015). Establishing a link between prescription drug abuse and illicit online pharmacies: analysis of twitter data, *Journal of medical Internet research* **17**(12).

Kuhn, M., Campillos, M., Letunic, I., Jensen, L. J. and Bork, P. (2010). A side effect resource to capture phenotypic effects of drugs, *Molecular systems biology* **6**(1): 343.

Kumar, A. (2017). Pharmacovigilance: Importance, concepts, and processes, *American Journal of Health-System Pharmacy* p. ajhp151031.

Landis, J. R. and Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers, *Biometrics* pp. 363–374.

Liaw, A., Wiener, M. et al. (2002). Classification and regression by randomforest, *R news* **2**(3): 18–22.

Lin, Y., Zhu, T., Wu, H., Zhang, J., Wang, X. and Zhou, A. (2014). Towards online anti-opinion spam: Spotting fake reviews from the review sequence, *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, IEEE, pp. 261–264.

Liu, Y., Bi, J.-W. and Fan, Z.-P. (2017). Ranking products through online reviews: A method based on sentiment analysis technique and intuitionistic fuzzy set theory., *Information Fusion* **36**: 149 – 161.
URL: *http://ezproxy.ncirl.ie/login?url=http://search.ebscohost.com/login.aspx?direct=trueAuthTyp livescope=site*

McCallum, A., Nigam, K. et al. (1998). A comparison of event models for naive bayes text classification, *AAAI-98 workshop on learning for text categorization*, Vol. 752, Madison, WI, pp. 41–48.

McCullagh, P. (1984). Generalized linear models, *European Journal of Operational Research* **16**(3): 285–292.

Miguel, A., Azevedo, L. F., Araújo, M. and Pereira, A. C. (2012). Frequency of adverse drug reactions in hospitalized patients: a systematic review and meta-analysis, *Pharmacoepidemiology and drug safety* **21**(11): 1139–1154.

Mishra, A., Malviya, A. and Aggarwal, S. (2015). Towards automatic pharmacovigilance: Analysing patient reviews and sentiment on oncological drugs, *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*, IEEE, pp. 1402–1409.

Na, J. and Kyaing, W. Y. M. (2015). Sentiment analysis of user-generated content on drug review websites, *Journal of Information Science Theory and Practice* **3**(1): 6–23.

Organization, W. H. et al. (2002). The importance of pharmacovigilance.

Perikos, I. and Hatzilygeroudis, I. (2017). Aspect based sentiment analysis in social media with classifier ensembles, *Computer and Information Science (ICIS), 2017 IEEE/ACIS 16th International Conference on*, IEEE, pp. 273–278.

Polikar, R. (2006). Ensemble based systems in decision making, *IEEE Circuits and systems magazine* **6**(3): 21–45.

Porter, M. F. (1980). An algorithm for suffix stripping, *Program* **14**(3): 130–137.

Quinlan, J. R. (1986). Induction of decision trees, *Machine learning* **1**(1): 81–106.

Rome, E. S. (2001). It's a rave new world: rave culture and illicit drug use in the young, *Cleveland Clinic Journal of Medicine* **68**(6): 541–550.

Salas-Zárate, M. d. P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodríguez-García, M. Á. and Valencia-García, R. (2017). Sentiment analysis on tweets about diabetes: An aspect-level approach, *Computational and mathematical methods in medicine* **2017**.

Sharma, A. and Dey, S. (2013). A boosted svm based ensemble classifier for sentiment analysis of online reviews, *ACM SIGAPP Applied Computing Review* **13**(4): 43–52.

Shepherd, G., Mohorn, P., Yacoub, K. and May, D. W. (2012). Adverse drug reaction deaths reported in united states vital statistics, 1999-2006, *Annals of Pharmacotherapy* **46**(2): 169–175.

Siddiqua, U. A., Ahsan, T. and Chy, A. N. (2016). Combining a rule-based classifier with ensemble of feature sets and machine learning techniques for sentiment analysis on microblog, *Computer and Information Technology (ICCIT), 2016 19th International Conference on*, IEEE, pp. 304–309.

Skarl, S. (2015). Anxiety and depression association of america, *Journal of Consumer Health on the Internet* **19**(2): 100–106.

Stahl, S. M. (2002). Don't ask, don't tell, but benzodiazepines are still the leading treatments for anxiety disorder, *The Journal of clinical psychiatry* **63**(9): 756–757.

Whitehead, M. and Yaeger, L. (2009). Building a general purpose cross-domain sentiment mining model, *Computer Science and Information Engineering, 2009 WRI World Congress on*, Vol. 4, IEEE, pp. 472–476.

Whitehead, M. and Yaeger, L. (2010). Sentiment mining using ensemble classification models, *Innovations and advances in computer sciences and engineering*, Springer, pp. 509–514.

Wood, K. (1994). The medical dictionary for drug regulatory affairs (meddra) project, *Pharmacoepidemiology and drug safety* **3**(1): 7–13.

Woodlock, D. (2005). Virtual pushers: Antidepressant internet marketing and women, *Women's Studies International Forum* **28**(4): 304 – 314.
URL: *http://www.sciencedirect.com/science/article/pii/S0277539505000269*

Yalamanchi, D. (2011). *Sideffective-system to mine patient reviews: sentiment analysis*, PhD thesis, Rutgers University-Graduate School-New Brunswick.