Classification of Online Patient Reviews Based on Effectiveness Using Machine Learning Algorithms

MSc Research Project
Data Analytics

Srinivas Prakash Anvekar
Student ID: x18130704

School of Computing
National College of Ireland

Supervisor: Dr. Pierpaolo Dondio

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Srinivas Prakash Anvekar
Student ID:	x18130704
Programme:	Data Analytics
Year:	2019/20
Module:	MSc Research Project
Supervisor:	Dr. Pierpaolo Dondio
Submission Due Date:	12/12/2019
Project Title:	Classification of Online Patient Reviews Based on Effectiveness Using Machine Learning Algorithms
Word Count:	8526
Page Count:	26

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 required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	28th January 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

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):	

Classification of Online Patient Reviews Based on Effectiveness Using Machine Learning Algorithms

Srinivas Prakash Anvekar
x18130704

Abstract

In the past decade, use of diverse expedited review approaches has expanded and due to various improvements in the field of medicine the time to market for the drugs has reduced considerably. This has resulted in lesser testing of drugs in clinical trials leading to Adverse Drug Reactions upon usage of those drugs in a real-world scenario. One of the leading causes of mortality rate over the years in past decade has been Adverse Drug Reactions and hence it has become a concern to oversee this situation. The boom in online forums and social media in recent years has brought about immense wealth of information submitted by patients related to drugs and their reactions. Valuable knowledge can be leveraged from this information using machine learning techniques. This research is concerned with processing the textual data about drugs from online patient reviews and classify them based on their effectiveness using SVM, kNN, Random Forest, XGBoost Classifier and Logistic Regression algorithms and evaluate which model best suits in obtaining the objective of classifying the reviews. The analysis is performed based on in-domain data and also cross-domain data. The transfer learning approach can be used to find the similarity across domains and is promising technique in field of review analysis. This study also tries to explain the correlation between sentiment of a review with that of effectiveness of the review. Among the different models, the XGBoost Classifier had the best performance over different approaches proving the viability of this research and further improvements on the model in future.

Keywords: Drug reviews, Sentiment Analysis, machine learning algorithms, feature extraction, Tf-Idf, Correlation.

1 Introduction

The steep rise in usage of social media on World Wide Web has dramatically changed people's way of conveying their opinions and how it affects their decision making in life. There is plenty of information related to medical field online and also with the rapid growth of social media platforms, people often tend to internet to learn about a drug before taking any further actions. There are numerous reasons for which a patient is recommended a drug for, it could to battling a disease, personal disorders or pain relief etc. Not every prescribed drug works in the same way for everyone. Some patients on taking these drugs find quick relief or for some it can cause side effects. The efficacy of the drugs also depends on the patient, a medicine can work really well, or it can be less effective.

Study of sciences concerned with collection, assessing, preventing, monitoring and detection of adverse effects of pharmaceutical commodities is known as pharmacovigilance. This study of extracting information and opinions about drugs from online reviews is a formidable task. This research focus is mainly on assessing and classifying the reviews for pain relief drugs, obtained from online public forums, based on their effectiveness.

1.1 Importance of pharmacovigilance in medical field

In 1961, there was a huge outbreak when thousands of infants were born with congeniality disorder as a result of exposure in the utero towards medicine that was promoted for use to mothers who were pregnant. This resulted in creation of WHO Pilot Research Project for International Drug Monitoring in 1968. The main reason for developing this system was to monitor and study the medicines. This led to the emergence of practice and science known as pharmacovigilance.

Pharmacovigilance and issues related to drug safety are applicable for everyone whose life is concerned in any way with medical interference. Powerful arrangement for drug regulations provide confidence for public in medicine. These pharmacovigilance programmes require firm links with the regulators to make sure that the safety issues are well educated to authorities in everyday practice. In order to achieve their objectives, these programmes must also be sufficiently supported in their activities. Any drug before approval must pass three regulations namely, good quality, effectiveness and safe to use Kumar et al. (2011)

Hospitalization of nearly 6.7% with a fatality rate of 0.32% was seen in United States due to adverse drug reactions (ADRs) as per study Lazarou et al. (1998). In another study conducted in United Kingdom (UK), cases directly associated with ADR was near to 80% with annual cost projected to be around $867m to NHS. It was also seen that 0.15% of the cases turned out to be fatal in nature. As a result, it has become a very important course of action to study drugs and their effects on people Pirmohamed et al. (2004).

With the advent of world wide web, limitless information related to multiple topic in medical fields has become available. With the rise in public forums and social media platforms, people are willing open to put across their views online about any product or service they use. At the same time, many people are also tending to these online forums to obtain information about the drugs they have been prescribed. Using this information available online, machine learning techniques like text processing, also known as natural language processing, can be applied on simple text to obtain critical information. This project mainly focuses on critically assessing the reviews of the prescribed drug based on the effectiveness and classify them using machine learning techniques.

1.2 Research Question

RQ: "How accurately can the drug reviews be classified based on their effectiveness as compared to the state of art model?"

Prior to using drugs as remedy for diseases and personal disorder it is important to observe them for any side-effects, adverse reactions. The focus of this study is to implement machine learning techniques on drug reviews provided by patients on online forums like Drugs.com[1], WebMD.com[2] etc. The aim is to extract information from these online text reviews using natural language processing technique and classification algorithms to classify them based on effectiveness. Developing and using such a well-developed model can help in finding the effectiveness of a new drug through reviews provided by patients during clinical trials or in real world scenario thereby providing better knowledge about the drug and also contribute towards data mining.

1.3 Objectives

- To extract useful and critical information from online drug reviews by applying natural language technique.
- To critically assess and classify drug reviews submitted by patients in online public forums based on their effectiveness.
- To find the correlation between the sentiment of a review and the drug effectiveness of a review.

The rest of the paper is structured in the following manner, Literature review has been carried out in Section 2. The methodological process followed to carry out the research is written in Section 3. Section 4 covers design specification. Implementation and Evaluation has been detailed in Section 5 and 6 respectively. Section 7 covers conclusion and future work that can be carried out.

2 Related Work

Drugs do not restore structures or functions in the body that are beyond repair but can only fasten or slow down the biochemical reactions within the body. The functioning of the drug can be measured in terms of potency, efficacy and effectiveness. This research focus is mainly on finding the effectiveness of the drug through analysis of reviews provided by patients on drugs in online forums.

2.1 Text Classification

Many studies in the past have been conducted that focused primarily on sentiment analysis using machine learning techniques. A cross domain sentiment model approach was proposed for sentiment analysis of reviews for objects like camera, laptops, radio, television etc. A model that has been trained in one of the domains was used for classification of reviews from other domains. SVM model was used as a base classifier and k-Fold cross validation was applied to evaluate the process. Through this work, they demonstrated that model trained in one domain can be applied across domains to classify the reviews based on their sentiment Whitehead and Yaeger (2009). In another study, ensemble models was used in performing sentiment classification. SVM model was used along with bagging, boosting, random subspace and bagging random subspace methods. Evaluation of these models was performed using the cross-validation method and results showed that

[1] https://www.drugs.com/

[2] https://www.webmd.com/

ensemble methods performed better as opposed to single base classifier in accuracy terms Whitehead and Yaeger (2010)

Opinion mining can also be performed by using lexicons. In 2012, creation of lexicon was done using two methods namely general approach lexicon and medical domain lexicon created using medical terms. A vote-flip algorithm was used for classifying the reviews using the two obtained lexicons. When the general lexicon was integrated with medical domain lexicon better results were seen with the precision score for positive results found to be 76% Goeuriot et al. (2012). Movie reviews were classified based on their sentiment using an effective pre-processing technique. This repetitive technique of pre-processing was modelled to procure a wordlist of tokens, and these tokens were used for review classification based on the sentiment. SVM Linear model had the best overall performance with an Accuracy of 97.25% Manek et al. (2013).

Mining of drug reviews was performed using a subjectivity-based lexicon method. Terms were extracted using a reiterative process on medical seed cache and this lexicon was expanded for antonyms and synonyms using SentiWordNet (SWN). It was seen that this method had a considerable coverage of medical terms and its performance was better than SWN and hybrid techniques in review classification Asghar et al. (2013). Similarly, in another research a condensed lexicon was formed using a hybrid approach that consisted of bootstrapping and corpus-based technique. This method showed that it was highly effective as compared to other traditional approaches of sentiment classification Asghar et al. (2016).

Analysis of sentiment was performed on reviews from forums based on hearing loss using machine learning techniques like SVM, Naïve-Bayes, Logistic Regression using lemmatization method. The lemmatized sentences were further vectorized using a bag-of-words approach. A bag-of-words approach is a method where features are extracted from unstructured text and the reviews are represented in the form of vector containing set of integers. The features thus obtained are assigned a score based on their polarity for a particular review. It was seen that performance of SVM model was better than other classifiers with a kappa score of 0.64 Ali et al. (2013).

An information extraction task known as Named Entity Recognition (NER) is a method where the unstructured text is classified into predetermined groups such as organization, medical terms, expressions, people name etc. It is seen that near human like performance can be obtained using this state-of-art technique for research studies. In recent years, NER has been used in medical field for different purposes. In one of the studies, NER was used to identify name of the drug, side effects and literals that provide relation between drug and side-effects. Classification and relationship between drug and side-effect was performed using Hidden Markov Model based on the output of NER model. It was seen that information obtained from this method matched the information that is published on Drug Package Inserts Sampathkumar et al. (2014).

Similar to NER approach, another advancement in this field is feature engineering for sentiment analysis. Feature engineering is a technique where-in domain knowledge is used to extract key features using a feature selection algorithm. ADR mentions from tweets were extracted using this method. Significant results were obtained providing a way for improvements in medical field Dai et al. (2016).

There have been many studies conducted on aspect-based sentiment analysis in recent past. In one of the researches, algorithm was developed through a linguistic methodology at a clause-level for sentiment analysis. This algorithm assigned sentiment scores to the reviews and classified each one into either positive, negative and neutral reviews. This approach was compared with a rule-based model using SVM algorithm and it was seen that the linguistic approach performed better analysis Na and Kyaing (2015).

Medical analysis at sentiment level on review was performed by constructing a model that captures common issues faced by patients who were under medication. Patient reviews were obtained by crawling through websites and sentiment analysis was performed using medical corpus like MedDRA Wood (1994) and SIDER Kuhn et al. (2010) to obtain a list of medical terms. SVM algorithm produced an accuracy of 59%. The study had certain drawbacks since there was an imbalance in review polarity and size of data that was taken for analysis resulting in low accuracy Mishra et al. (2015).

Opinion mining was used to mine drug satisfaction levels from patient reviews. Their research focused on using SVM classifier along with two neural network methods Probabilistic Neural Network (PNN) and RBFN (Radial Basis Function Neural Networks). Universal approximations used in RBFN method had better performance compared to other techniques, it overcomes the drawbacks of PNN as it utilizes radial basis function to perform approximations Gopalkrishnan and Chandrasekaran (2017). Information extraction from opinion is a demanding task, in one of the researches SVM and rule-based technique was used to extract drug side effects from reviews. This technique was found to be effective and provided grounds for drug investigations prior to releasing it providing vital information for pharmaceutical companies and physicians Ebrahimi et al. (2016).

Classification of reviews in Russian language was performed using SVM model into different categories like indications, beneficial effect, ADR and others. This research used different set of features like bag-of-words (bow), parts-of-speech (pos), word-embedding (emb) among others and significant results were seen in their study. The highest F-score for classification was obtained for a model that had a feature set which was a combination of pos, word2vec, disease lexicon and Pointwise-mutual-information Alimova et al. (2017). Most of the researches conducted before were mainly done on sentiment analysis and classification, very little work has been performed in the field of drug review classification based on its effectiveness. In a recent work in medical field, cross-domain and cross-data learning methods was applied for aspect-based analysis of drugs based on the effectiveness, side effects. It was seen that transferring learning between domains could be made use of in obtaining similarities and is a promising method in medical analysis Gräßer et al. (2018).

In one another research, the text classification was performed using improved tf-idf algorithm. A weighting factor E(t) was introduced in their algorithm that reflected the inter-class and intra-class dispersion and the degree of association between the categories and feature words. The paper also made use of the popular Word2Vec technique for representation of text in form of vectors. The words are represented using weighted technique and the text is classified using these vectors. It was seen that the improved Tf-Idf classification performance was better as compared to other models Fan and Qin (2018).

SMS spam detection is one of the approaches in that comes under text classification. As the name suggest, the message received is classified as spam or not. In a recent study, in order to select the best relevant features two wrapper algorithms were used for extraction. Nine optimal feature sets were obtained, and it was seen that the accuracy obtained using the XGBoost Algorithm was relatively high as compared to other boosting algorithms Mussa and M. Jameel (2019).

2.2 Handling Data Imbalance

Learning from data sets that have imbalance classes is a tedious task in data mining. Data imbalance is one of the issues that has an effect on the raw data. In this research, the positive cases outnumber the negative number cases. Modern algorithms struggle in dealing with this class imbalance and often result in higher error rates for the minority class.

Various studies have been conducted to study the effect of imbalance in data and how to handle the class imbalance. The handling of imbalance data in classification problems has been discussed in one of the studies. Different techniques have been employed and evaluated against six metrics namely, Accuracy, Precision, Recall, F1-score, Geometric mean, Mathew Correlation Coefficient and AUC. It was seen that the SMOTE method outperformed other techniques of handling imbalance data Vimalraj and Porkodi (2018).

Classification of questions is an important task in question-answering systems. One of the drawbacks of Question Classification (QC) is the imbalance of classes in data. In order to overcome the imbalance in QC, in a recent study a hierarchical SMOTE technique was employed. It was seen that the Naïve-Bayes Classifier had a better classification accuracy when classifying casual and choice questions as without applying the SMOTE technique. The results also showed that other categories of questions also had a better performance using SMOTE approach Mohasseb et al. (2018).

In one of the recent studies, various techniques were studied for handling of class imbalance data through highlighting the degree to which there is class imbalance and its corresponding effects on the performance. Through their research it was seen that SMOTE technique produce higher accuracy for binary classification problems Thabtah et al. (2019).

2.3 Conclusion

The main focus of this research study is to classifying the drugs based on the efficacy and it uses the logistic regression model used in the study Gräßer et al. (2018) as the base model and improve on the results provided in the earlier studies. Through literature review, it is seen that SVM model played an important role in giving results with high accuracy compared to other models. As a part of this study, SVM model has be used along with other models like XGBoost, KNN, Random Forest in classifying the reviews based on drug effectiveness. Through literature review it was seen that SMOTE technique is a very effective method to handle class imbalance data sets. As a part of this research study, to handle the imbalance of data, an oversampling technique known SMOTE is employed to reduce noise and error during classification and handle imbalance in class.

3 Methodology

Different data mining methodologies are available to develop an application based on machine learning technology. They include KDD (Knowledge Discovery in Data), CRISP-DM (Cross Industry Standard for Data Mining) and SEMMA (Sample, Explore, Modify, Model, and Assess). This research is based on natural language processing techniques and through research it was seen that KDD methodology would be most apt in developing a solution for classification of medical reviews based on their effectiveness. KDD

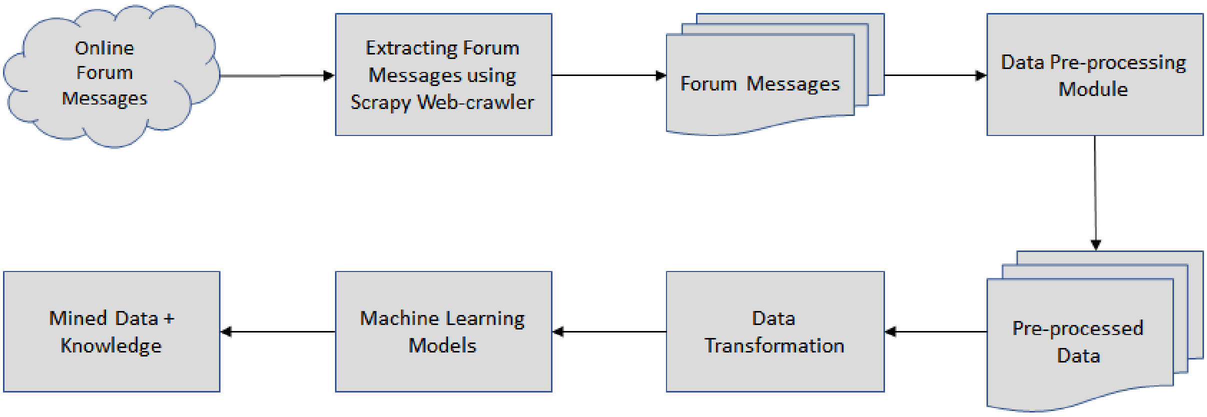


Figure 1: KDD Methodology

abbreviation stands for knowledge discovery in database, it is a method of obtaining useful information from underlying data Fayyad et al. (1996). The different steps involved in KDD methodology as a part of this study is shown in Figure 1.

The rest of this section shall explain briefly the details involved in each of processes of the KDD life cycle.

3.1 Data Collection

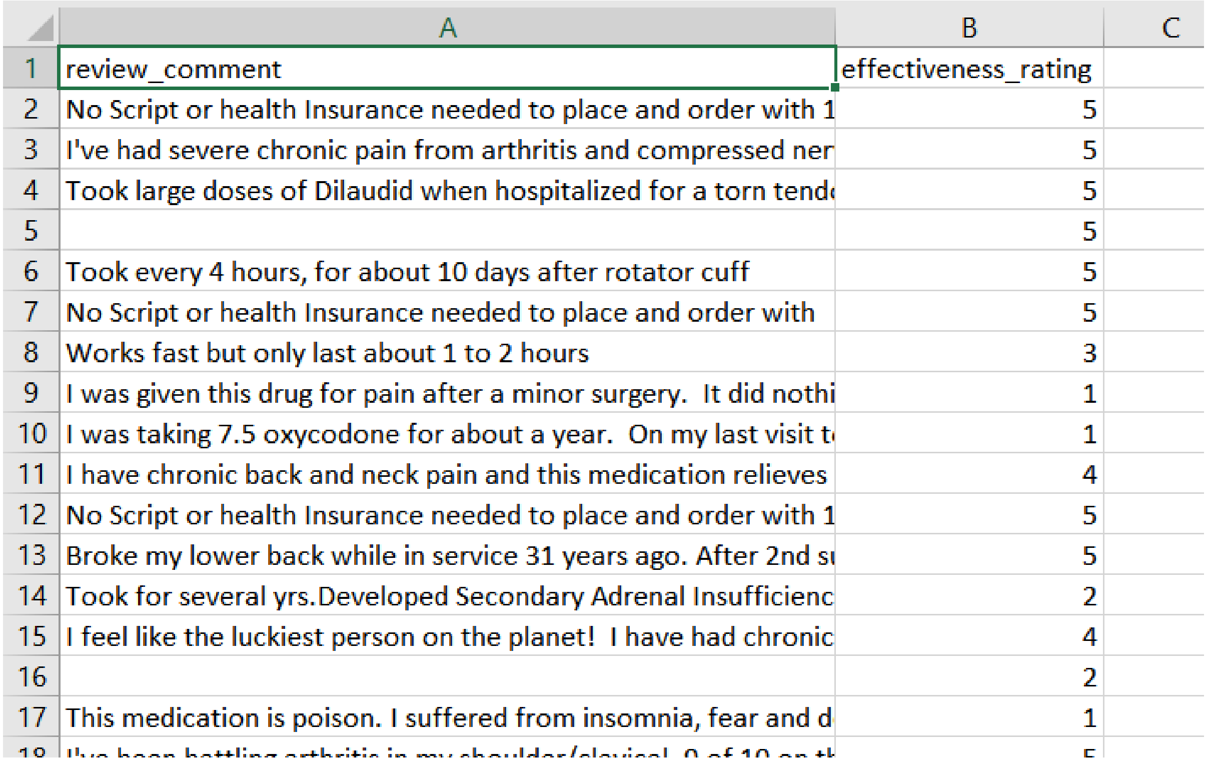
Most of the researches that have been performed on drug reviews from online forums or tweets extracted from twitter are based on sentiment analysis. The amount of research on classifying the drugs based on its efficacy is very less and focus of this research is on this topic. The data for classification is extracted from online public forums like WebMD.com , Drugs.com that contain reviews pertaining to the drugs used for pain relief by patients. The data extracted contains review comments, review rating provided by patients and this information is stored in CSV format. Figure 2 gives a sample screenshot of the drug stored in csv format. This data was extracted using the Scrapy tool developed on Python.

3.2 Data Pre-processing and Transformation

The data collected from the websites is in an unstructured format. This data is pre-processed to a format that could be understood by machine learning algorithms. The pre-processed data is transformed in to a vectoral format using python packages namely CountVectorizer and TfidfVectorizer. This vector will be used by the models for classification of data. A detailed explanation of the pre-processing and transformation phase has been described in Section 5 along with the implementation of the model.

3.3 Data Mining Models

One of the objectives of this research is classifying the reviews based on their effectiveness using machine learning algorithms. Since the reviews in the datasets are labelled, a supervised learning approach is followed in classifying the data. In this research, the logistic regression algorithm used in the study Gräßer et al. (2018) is used as the baseline



	A	B	C
1	review_comment	effectiveness_rating	
2	No Script or health Insurance needed to place and order with 1	5	
3	I've had severe chronic pain from arthritis and compressed ner	5	
4	Took large doses of Dilaudid when hospitalized for a torn tend	5	
5		5	
6	Took every 4 hours, for about 10 days after rotator cuff	5	
7	No Script or health Insurance needed to place and order with	5	
8	Works fast but only last about 1 to 2 hours	3	
9	I was given this drug for pain after a minor surgery. It did nothi	1	
10	I was taking 7.5 oxycodone for about a year. On my last visit t	1	
11	I have chronic back and neck pain and this medication relieves	4	
12	No Script or health Insurance needed to place and order with 1	5	
13	Broke my lower back while in service 31 years ago. After 2nd su	5	
14	Took for several yrs.Developed Secondary Adrenal Insufficienc	2	
15	I feel like the luckiest person on the planet! I have had chronic	4	
16		2	
17	This medication is poison. I suffered from insomnia, fear and d	1	
18	I've been battling arthritis in my shoulder/clavical. 9 of 10 on th	5	

Figure 2: Data stored in CSV format

model. This model will be contrasted with other machine learning algorithms like Random Forest, SVM, kNN (k-nearest neighbours) and XGBoost(XGB) Classifier and the best model obtained can be used for further studies.

3.4 Evaluation Metrics

The dataset containing the reviews is split in the ratio 70:30 into training dataset and test dataset respectively. Initially the models are trained with the training dataset and results are obtained using the test dataset based on trained model. The obtained results are be evaluated against four evaluation metrics namely Accuracy, Precision, Recall, F1-score. The values obtained are in terms of percentages showing how well did our models perform in classify the reviews. Below are the equations for calculating the metrics.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)}$$

$$Precision = \frac{TP}{(TP+FP)}$$

$$Recall = \frac{TP}{(TP+FN)}$$

$$F1 = \frac{2*(Recall*Precision)}{(Recall+Precision)}$$

In order to check if there is statistical significant difference between two models, a t-test can be performed using the below formula[3]

[3] https://www3.nd.edu/~rjohns15/cse40647.sp14/www/content/lectures/28%20-%20Classifier%20Comparisons.pdf

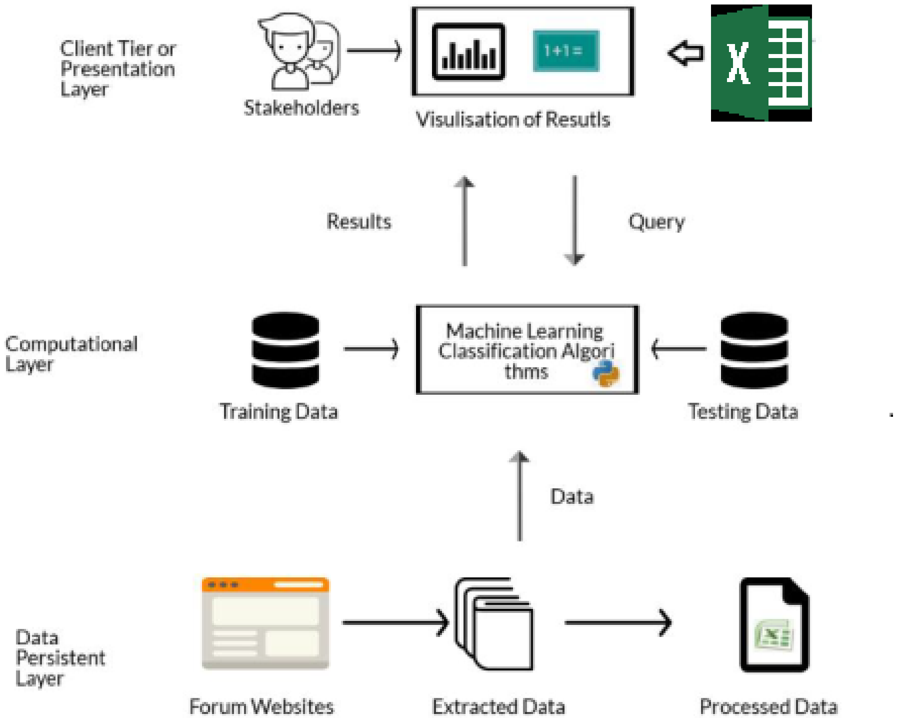


Figure 3: Three-Tier Architecture

$$t\_score = \frac{Acc1-Acc2}{\sqrt{((Acc1*(1-Acc1)/n)+(Acc2*(1-Acc2)/n)}}$$

where Acc1 is the accuracy of model 1, Acc2 is the accuracy of model 2 and n is the size of dataset.

If the t_score between the two models is:

- greater than 1.64, statistical difference at 90% confidence level
- greater than 1.96, statistical difference at 95% confidence level
- greater than 2.32, statistical difference at 99% confidence level

Section 6 gives details about various use cases and performances of different models used for classification.

4 Design Specification

This section shall outline the architecture that has been followed to implement the solution for the research study. A three-tier architecture is followed in this research. The top tier is the client layer, followed by business layer consisting of machine learning models as the middle tier and a data extraction layer as the bottom tier of the architecture. The three-tier architecture is as shown in the Figure 3.

The top tier of design is the client tier where the results of the analysis is presented to the clients/stakeholders. The results from analysis is presented in a visualisation format that is understood by the stakeholders. The middle tier of design is the computational layer consisting of machine learning algorithms where the pre-processed data is analysed. The results obtained here are passed to the top tier for presentation. The data for this layer is obtain from the bottom tier namely data persistent layer. The data for this layer is obtained from online public forums where patients have provided their views on the prescribed drugs in the form of online reviews. This data is pre-processed before sending it to the middle layer for computation.

5 Implementation

The following section shall provide details on how the implementation was performed to obtain an efficient model to classify the drug reviews based on their effectiveness. This section will also briefly describe the feature extraction and selection process.

5.1 Environmental Setup

The implementation for this research study was carried out on a 64-bit Windows Operating System with 8 GB of RAM. The models are developed using Python coding language. The Spyder IDE (Integrated Development Environment) provided by Anaconda software is used for writing the code and running them for results. Latest version of Python 3.7.3 was used. Microsoft Excel is used for obtaining visualization reports.

5.2 Data Selection

The first and the foremost step of this research study is the collection of appropriate data required for analysis of effectiveness of the drugs. For this study, the reviews provided by the patient in online public forums are taken into consideration. With the evolution of social media, more and more people have shown willingness to share their thoughts about various products and services they have used.

There are several public medical forums that contain information related to the reviews provided by patients about the drugs they have been prescribed for different purposes, for this research reviews have been taken from WebMD.com, Drugs.com. The data is scraped from these websites and it consists of patient review comments, sex of the patient, effectiveness of the drug in form of ratings. The scrapping of data from the website is done in python with the help of a web crawler toolkit known as Scrapy . The dataset consists of 15890 number of reviews and their attributes i.e. review effectiveness rating is stored in a csv format as shown in Figure 2. This data is then pre-processed to a format that can be understood by the machine learning algorithms.

5.3 Data Pre-Processing

The raw unstructured data obtained from websites cannot be understood by the machine learning algorithms. Hence it is very important to remove any noise from the data and bring it to a format that can be utilized by the algorithms for analysis. This process is known as pre-processing of data and is the most important part of KDD cycle and research

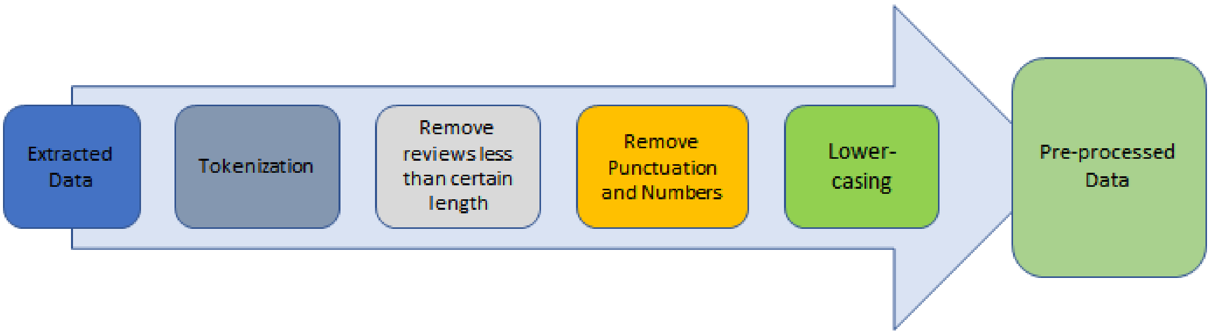

Extracted Data
Tokenization
Remove reviews less than certain length
Remove Punctuation and Numbers
Lower-casing
Pre-processed Data

Figure 4: Data Pre-Processing

studies. The data pre-processing is carried out in python using Spyder IDE. The initial step in pre-processing was to handle the missing values. The dataset contained certain entries in which the review comments were empty, such rows were dropped from the dataset since they do not help in analysis of effectiveness of the drugs. For text analysis, there are certain pre-defined steps that have to be performed on the unstructured text for the algorithms to be able to understand the data. The different steps involved in text processing is as shown in the Figure 4.

5.4 Feature Selection and Extraction

We want our documents to be represented in vectoral form, in doing this we need to make sure that the vector space consists of what we think is important. Feature selection is a process of selecting those words that we think are worthwhile in our documents and ignoring those words that do not have meaningful contribution and representing them in a vectoral space for analysis by machine learning algorithms. In this research, feature selection is performed using different techniques. They included n-gram approach (unigram, bigram), tf-idf approach (unigram, bigram) to obtain the corpus. The corpus thus obtained using the said approaches is used to represent the review comments provided by patient in their vectoral form.

1) **N-gram approach**: Unigram and bigram features are obtained in this method using the CountVectorizer python package. Two parameters namely max_df (used for removing terms that appear too frequently) and min_df (used for removing terms that appear too infrequently) are used to obtain the feature set for the unigram and bigram features. The max_df parameter was set 0.75 meaning the term that appear in more than 75% of the documents were removed and the min_df parameter was set to 1 meaning that the term must appear in at least one of the documents.

2) **Tf-idf approach**: Similarly, tf-idf feature weights for unigram and bigram terms was obtained from the model. This was obtained using the TfidfVectorizer python package. Tf-idf plays an important role in classification as opposed to only using the term frequency for feature extraction Li et al. (2014). Two parameters namely max_df (used for removing terms that appear too frequently) and min_df (used for removing terms that appear too infrequently) were used to obtain the feature set for the unigram and bigram features. The max_df parameter was set 0.75 meaning the term that appear in more than 75% of the documents were removed and the min_df parameter was set to 1 meaning that the term must appear in at least one of the documents.

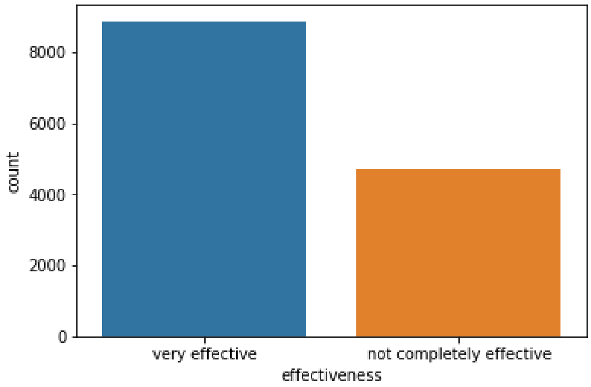


Figure 5: Class Imbalance in data extracted

5.5 Transformation

The effectiveness of drug for each review ranged from values 1 to 5 given by the patients. To help distinguish the effectiveness of the drug in a better way, the ratings were transformed into two categories by adding a new column. The ratings were represented in a new column as "effectiveness" and put into two categories namely "not completely effective" and "very effective". The ratings for review 1 to 3 were represented as "not completely effective" and ratings for review 4 and 5 were represented as "very effective".

Each review extracted from the forums, was then represented in their vectoral form using the features obtained from the feature selection and extraction stage using the respective python packages, namely CountVectorizer model was used to fit and transform the reviews based on unigram and bigram features and TfidfVectorizer model was used to fit and transform the reviews based on unigram and bigram tf-idf weights.

Upon transformation of review ratings to their respective classes, imbalance in data was seen in the respective classes as shown in the Figure 5. Using this imbalance data on algorithms would provide inappropriate results and may result in overfitting of model towards one of the classes. To overcome this drawback, a technique known as oversampling of data is performed to obtain better results. Many algorithms are proposed for handling imbalanced data, but many are complex and tend to produce some noise. For this purpose, the SMOTE proposed by Chawla et al. (2011) eliminates generation of noise and overcomes imbalance in classes effectively. The technique is performed using the SMOTE package from sklearn in python. The distribution of classes after performing oversampling technique is as shown in Figure 6.

5.6 Model Implementation

5.6.1 Logistic Regression

Logistic regression is used as a classifier to assign observations to a discrete set of classes. Some of the classification related problem include Email spam detection, Online transaction fraud detection, Tumour detection etc. In this research, this model is used for classifying the data based on the effectiveness and serves as the base model as used in the study (Gräßer. et.al., 2018). Grid search is performed to obtain the best parameters for the model, they include C = 0.01, solver = newton-cg, penalty = 'l2', class_weight='balanced',

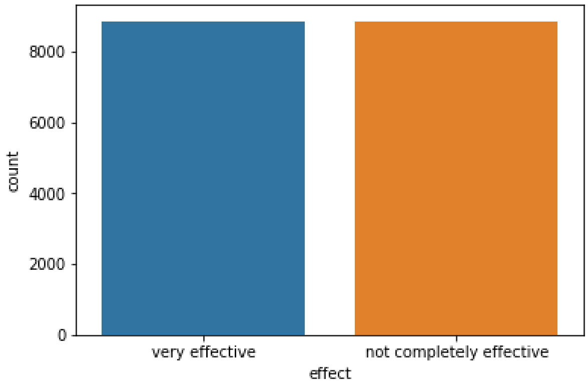


Figure 6: Balanced data after applying SMOTE

multi_class = 'multinomial'. The data is split into train and test dataset and performance of the model is measured in terms of accuracy, precision, f1_score, recall.

5.6.2 Random Forest

Random forest classifier Breiman (2001) is a meta estimator that uses a number of decision trees on different sub-samples of same size from the data set for classification. The model uses averaging for improving predictive accuracy and to control over-fitting. The size of sub-sample space is same as original size but uses replacement technique to choose samples if bootstrap is set to True. A grid search is performed using the RandomSearchCV package from sklearn to find the best parameters for the model. It was seen that the best parameters were n_estimators = 1600, min_samples_split = 10, min_samples_leaf = 1, max_features = sqrt, depth is none and bootstrap is True. The data is again split in 70:30 ratio and similar to previous model accuracy, precision, f1_score, recall is calculated to test the performance.

5.6.3 KNN (k-Nearest-Neighbours)

k-Nearest-Neighbours algorithm is a well-known non-parametric algorithm that is used for classification in pattern recognition. The output of an object is based on plurality vote of its k nearest neighbours. Since this study deals with vector representation of data, kNN is one of the chosen models. Similar to other models, a grid search is performed to obtain the best parameter for the model, they were algorithm = 'auto', leaf_size = 3, metric = 'euclidean', n_jobs = -1, n_neighbors = 3. The data for this model is split in 70:30 ratio and accuracy, precision, f1 score, recall is calculated to check the performance.

5.6.4 Support Vector Machine

Support Vector Machine (SVM) model is one of the powerful classifiers in statistical learning and according to the study Zhang et al. (2011) it has proven to be efficient for classification tasks in text categorization. The SVM model is deployed using SVC package from sklearn library. Default values are used for most of the parameters. Hyperparameter tuning was applied to obtain best parameters, the kernel used for the model is linear kernel with exponent value 1.0, gamma as 10 and C parameter set to 1.0. The dataset is split into train and test set in 70:30 ratio respectively. The accuracy, precision, f1_score, recall is calculated using the functions from metric package from sklearn library to test the performance of the model.

5.6.5 XGBoost Classifier

Gradient Boosting method is a type of ensemble learning that train and predict many models at once to produce one superior output. XGBoost (Xtreme Gradient Boosting) is a part of gradient boosting algorithms optimized for recent data problems and tools. It is very efficient and flexible, implemented using the Gradient Boosting Framework. Most of the parameters of the algorithm are kept default with other parameters set as follows, learning_rate =0.1, n_estimators=1000, max_depth=5, min_child_weight=1, gamma=0, subsample=0.8, colsample_bytree=0.8, objective= 'binary:logistic', nthread=4, scale_pos_weight=1, seed=27, random_state = 4. The dataset is split into train and test set in 70:30 ratio respectively. The accuracy, precision, f1_score, recall is calculated using the functions from metric package from sklearn library to test the performance of the model.

6 Evaluation

One of the key parts of research project is evaluating the models to prove that there performance is better than other models in the same field. A machine learning algorithm validation is not comprehensive until it has been tested using the test data for its performance. This section will detail about the different approaches taken for validating the data.

The most common evaluation metrics that are used for classification problem include Accuracy, Recall, Precision. The formulas for the same are defined in Section 3. As per the study Hossin and Sulaiman (2015), one of the drawbacks of accuracy is its limitation to produce less distinctive and less discriminating values. It also powerless in terms of informativeness and favours the minority class, hence F-measure is used in addition to others and it acts as good discriminator and provides better information in classification problems especially binary. As a result, five performance metrics are calculated for each model for testing the performance and to check which model performs better than the other.

6.1 Case Study 1 - Selection of feature set size and In-Domain Data Analysis

In order to select the size of the feature set, Random Forest model is taken into consideration. Both Bag-of-Words (BoW) model and Tf-IDF models are used to obtain the vectors for the reviews based on varying feature set size ranging from 100 features to 5000 features. Accuracy of the model in each of these cases is calculated and compared against each other. It was seen that the accuracy of the model was highest for the model when the size of the feature set was 2500 as seen in Figure 7. As a result, in each of the following analysis, the feature set size considered is 2500.

In this analysis, the overall performance of the models are tested. The reviews taken from both the websites WebMD.com and Drugs.com are combined for analysis. Feature set is obtained using the BoW and the Tf-Idf model. The maximum set of features taken

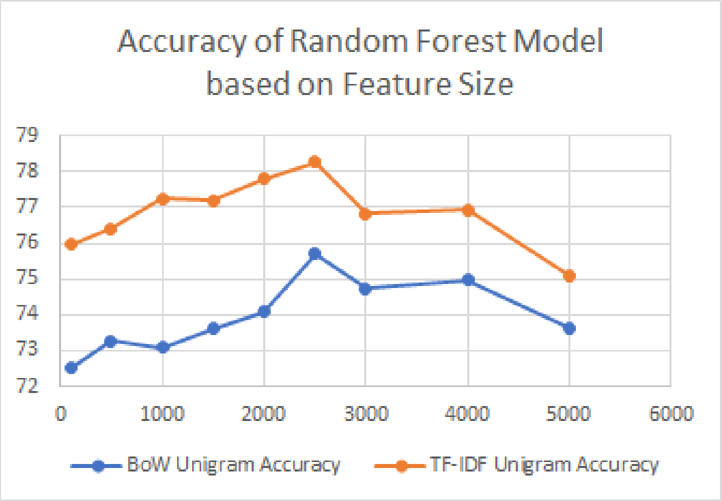


Figure 7: Accuracy of the Random Forest model based on Feature set size

for both the models was 2500 and appropriate parameters are set for the models with values obtained using grid search technique. The reviews are vectorized and split into training and test data set. Different machine learning models are run for analysis. The details of the results obtained for each of the models are tabulated as shown in Table 1. In the results from Table 1, it can be seen that the XGBoost algorithm outperformed other models in each case, independent of the feature set taken for vectorization. It can further be seen that, the unigram model with tf-idf vectorization had the best performance using the XGBoost Algorithm. From the analysis, it can also be clearly seen that the model performance is much better that the base model i.e. logistic regression. The accuracy of XGBoost was found to be 79.5%, while the precision 79.786%, recall of 79.5% and f1 score 79.453%. Random Forest model also performed equally well and it had an accuracy of 78.255% and a recall score of 80.642%. These two models, performed much better than the base model, logistic regression used in the study Gräßer et al. (2018).

6.2 Case Study 2 - Cross-Domain Data Analysis

Second case study is based on cross domain analysis. Here the feature set is obtained from one dataset is used as a base to vectorize the reviews from the other data set, that is the feature set obtained from reviews from WebMD.com is used to vectorize the reviews from Drugs.com and vice versa. In each case the data is split in the ratio as per model parameters and tested for model performance.

The results of cross-domain data analysis using Drugs.com feature set is tabulated in Table 2 and the results for cross-domain data analysis using the feature set from WebMD.com is given in Table 3. The main purpose of performing a cross-domain data analysis is to check the transferability of models across domains or websites. From the tables, it can be seen that the performance of the models using the feature set from Drugs.com are slightly less accurate as compared to the model performance using the feature set from WebMD.com. The best model in Table 2 being XGBoost model using Continuous Bag-of-Words Unigram features with an Accuracy of 77.116%, Recall of 77.116%, Precision of 77.140%, F1 score of 77.102% and in case of Table 3, again the XGBoost model using Tf-

Table 1: In-Domain Data Analysis

Feature Vector being used for training the models.				
Machine Learning Algorithm	BoW – Unigram features	BoW – Unigram & Bigram features	Tf-Idf – Unigram features	Tf-Idf – Unigram, Bigram Features

Logistic Regression	accuracy = 76.043
	precision = 76.423
	recall = 76.043
	f1 = 75.994
		accuracy = 76.156
		precision = 76.518
		recall = 76.156
		f1 = 76.111
			accuracy = 75.122
			precision = 75.349
			recall = 75.122
			f1 = 75.097
				accuracy = 74.577
				precision = 74.933
				recall = 74.577
				f1 = 74.527

kNN	accuracy = 51.560
	precision = 68.740
	recall = 51.560
	f1 = 37.757
		accuracy = 52.029
		precision = 68.392
		recall = 52.029
		f1 = 38.468
			accuracy = 60.259
			precision = 68.988
			recall = 60.259
			f1 = 55.121
				accuracy = 58.944
				precision = 68.396
				recall = 58.944
				f1 = 52.943

SVM	accuracy = 75.179
	precision = 75.415
	recall = 75.179
	f1 = 75.152
		accuracy = 75.291
		precision = 75.503
		recall = 75.291
		f1 = 75.269
			accuracy = 73.487
			precision = 73.782
			recall = 73.487
			f1 = 73.446
				accuracy = 72.980
				precision = 73.309
				recall = 72.980
				f1 = 72.929

Random-Forest	accuracy = 75.691
	precision = 76.318
	recall = 75.691
	f1 = 75.514
		accuracy = 76.564
		precision = 77.422
		recall = 76.564
		f1 = 76.346
			accuracy = 78.255
			precision = 80.642
			recall = 78.255
			f1 = 77.778
				accuracy = 78.199
				precision = 80.634
				recall = 78.199
				f1 = 77.711

XGBoost	accuracy = 77.828
	precision = 77.832
	recall = 77.828
	f1 = 77.824
		accuracy = 76.851
		precision = 76.877
		recall = 76.851
		f1 = 76.844
			accuracy = 79.500
			precision = 79.786
			recall = 79.500
			f1 = 79.453
				accuracy = 78.749
				precision = 78.859
				recall = 78.749
				f1 = 78.714

Table 2: Cross-Domain Data Analysis using Drugs.com Feature set.

		Drugs.com Feature Set			
Review Data from WebMD.com	**Machine Learning Algorithm**	BoW – Unigram features	BoW – Bigram features	Tf-Idf – Unigram features	Tf-Idf – Bigram Features
	Logistic Regression	accuracy = 72.489 precision = 73.132 recall = 72.489 f1 = 72.268	accuracy = 72.912 precision = 73.485 recall = 72.912 f1 = 72.721	accuracy = 72.235 precision = 72.422 recall = 72.235 f1 = 72.161	accuracy = 72.517 precision = 72.802 recall = 72.517 f1 = 72.412
	kNN	accuracy = 53.330 precision = 66.557 recall = 53.330 f1 = 39.447	accuracy = 51.637 precision = 66.146 recall = 51.637 f1 = 37.740	accuracy = 58.521 precision = 63.940 recall = 58.521 f1 = 53.630	accuracy = 57.167 precision = 65.822 recall = 57.167 f1 = 51.119
	SVM	accuracy = 72.602 precision = 72.776 recall = 72.602 f1 = 72.534	accuracy = 71.924 precision = 72.125 recall = 71.924 f1 = 71.844	accuracy = 71.501 precision = 71.739 recall = 71.501 f1 = 71.404	accuracy = 70.711 precision = 70.992 recall = 70.711 f1 = 70.591
	Random-Forest	accuracy = 71.979 precision = 72.683 recall = 71.979 f1 = 71.800	accuracy = 72.000 precision = 72.673 recall = 72.000 f1 = 71.830	accuracy = 74.265 precision = 76.583 recall = 74.265 f1 = 73.748	accuracy = 73.968 precision = 76.347 recall = 73.968 f1 = 73.427
	XGBoost	accuracy = 74.859 precision = 74.856 recall = 74.859 f1 = 74.856	**accuracy = 77.116** **precision = 77.140** **recall = 77.116** **f1 = 77.102**	accuracy = 76.383 precision = 76.602 recall = 76.383 f1 = 76.326	accuracy = 75.762 precision = 75.851 recall = 75.762 f1 = 75.745

Table 3: Cross-Domain Data Analysis using WebMD.com Feature set

		WebMD.com Feature Set			
Review Data from Drugs.com	**Machine Learning Algorithm**	BoW – Unigram features	BoW – Bigram features	Tf-Idf – Unigram features	Tf-Idf – Bigram Features
	Logistic Regression	accuracy = 74.482 precision = 74.491 recall = 74.482 f1 = 74.475	accuracy = 74.300 precision = 74.312 recall = 74.300 f1 = 74.290	accuracy = 78.197 precision = 78.259 recall = 78.197 f1 = 78.177	accuracy = 78.867 precision = 78.949 recall = 78.867 f1 = 78.843
	kNN	accuracy = 51.096 precision = 70.502 recall = 51.096 f1 = 36.446	accuracy = 50.974 precision = 70.652 recall = 50.974 f1 = 36.369	accuracy = 62.607 precision = 72.160 recall = 62.607 f1 = 57.588	accuracy = 59.379 precision = 69.459 recall = 59.379 f1 = 53.179
	SVM	accuracy = 76.918 precision = 76.947 recall = 76.918 f1 = 76.905	accuracy = 78.319 precision = 78.344 recall = 78.319 f1 = 78.309	accuracy = 81.060 precision = 81.375 recall = 81.060 f1 = 80.998	accuracy = 80.938 precision = 81.054 recall = 80.938 f1 = 80.911
	Random-Forest	accuracy = 77.752 precision = 80.080 recall = 77.752 f1 = 77.409	accuracy = 77.661 precision = 79.982 recall = 77.661 f1 = 77.316	accuracy = 79.123 precision = 82.841 recall = 79.123 f1 = 78.619	accuracy = 79.580 precision = 83.041 recall = 79.580 f1 = 79.128
	XGBoost	accuracy = 79.233 precision = 79.234 recall = 79.233 f1 = 79.233	accuracy = 78.076 precision = 78.075 recall = 78.076 f1 = 78.075	**accuracy = 81.425** **precision = 81.843** **recall = 81.425** **f1 = 81.370**	accuracy = 80.207 precision = 80.318 recall = 80.207 f1 = 80.172

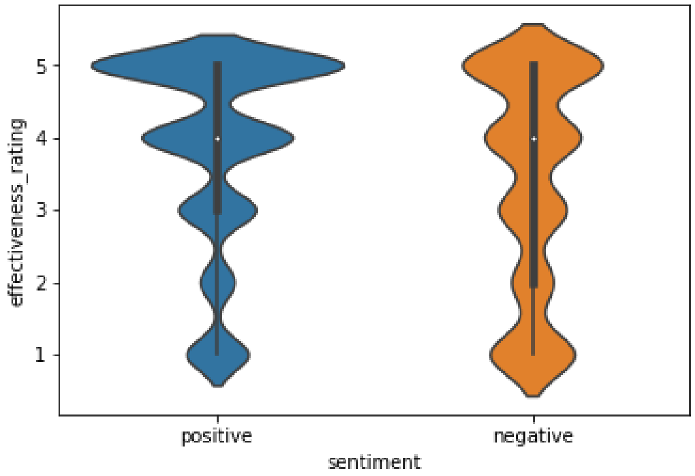


Figure 8: Distribution of Sentiment based on Reviews across effectiveness rating

idf Unigram features with an Accuracy of 81.425%, Precision 81.843%, Recall 81.425%, F1 81.370%. In order to check if there a statistical differnce between the two models, t_score is calculated using the formula given in Section 3.4. The value of t_score is 8.38, which shows that there is high statistical significant difference between the two models with 99% confidence level.

Moreover, the models trained and tested using the features from WebMD.com performed slightly better as compared to the models trained and tested using the features from Drugs.com, one of the reasons for the models to perform in such a manner is as due to limited size of training data set for Drugs.com and differing data properties across the domains. Another reason for this would be the features obtained are less familiar due to limited data size resulting in more of zeroes being embedded to the vectors as compared to the features obtained from WebMD.com reviews.

6.3 Case Study 3 - Correlation with Sentiment Analysis

Most of the analysis on textual data is usually performed on sentiment of the text. It can be thought of that the effectiveness prediction is similar to that of predicting the sentiment of the sentence. This case study shows the correlation between the sentiment and the effectiveness of the drug review.

As it can be seen in Figure 8, we can see that the sentiment of the review is spread across all the ratings independent of sentiment. From this we can clearly say that there is very little correlation between the sentiment of the review against the effectiveness of the review. Similarly, in Figure 9 the effectiveness of a review is clearly categorized in to two types depending on effectiveness rating provided by the user, that is the ratings from 1 to 3 is categorized as not completely effective and ratings 4 to 5 is very effectiveness.

To prove this further, a Pearson correlation co-efficient was obtained between sentiment and effectiveness of the reviews and it was seen that the value of co-efficient was found to be 0.144 as shown in the Figure 10, which further suggests that there is no correlation between sentiment and effectiveness.

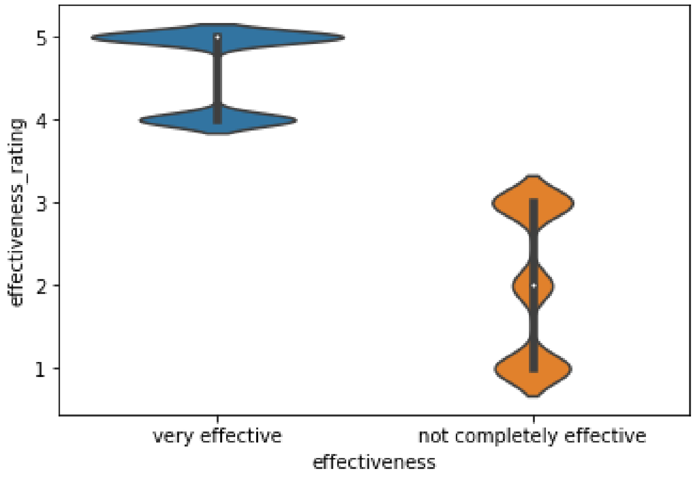
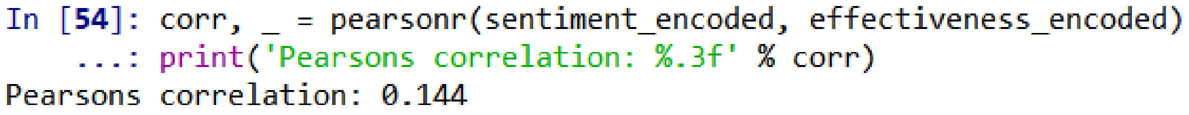


Figure 9: Distribution of Drug Effectiveness based on Reviews across effectiveness rating



```
In [54]: corr, _ = pearsonr(sentiment_encoded, effectiveness_encoded)
    ...: print('Pearsons correlation: %.3f' % corr)
Pearsons correlation: 0.144
```

Figure 10: Correlation between Sentiment and Drug Effectiveness based on Reviews

6.4 Discussion

The main aim of this research was to develop a model that can effectively classify the review submitted in online public forums based on the drug effectiveness. Case study one was on classification of drug review based on effectiveness of the drugs. Four different feature sets were used for classifying the model and the results for each feature set is as tabulated in Table 1. It can be seen that the Feature set that contained the Tf-Idf values for unigram and bigram features performed the best among other models. In this feature set, the XGBoost model had the best performance among the applied machine learning models. It outperformed the baseline model i.e. Logistic Regression model used by Gräßer et al. (2018). The accuracy of the model was found to be 80.049%, which is nearly 7% more than the baseline model. Figure 11 gives a clear picture of the values obtained from the models using the Tf-Idf Unigram feature set.

Case study two was performed to check the transferability of the models across domain by using feature sets of one domain on train and test data from other domains. This is mainly useful to show that the model behaves well and will have a very good impact towards predicting the effectiveness of the review. Similar to case study one, four different feature sets were developed and the same was used to train and test the different machine learning models. The different models used in this study were, Logistic Regression as the base model, KNN, SVM, RandomForest and XGBoost model. The results for this case study are as tabulated in Table 2 and Table 3. It was seen that the feature set obtained from WebMD.com reviews when used to train and test the models had better performance in classification as compared to the feature set used from Drugs.com reviews. This is mainly because the data set taken from WebMD.com was large as compared to Drugs.com and the reviews in WebMD.com were more structured as compared to the reviews from Drugs.com. The XGBoost model with Tf-Idf Unigram features taken from WebMD.com performed better than other models in this use case. The model had the following performance metrics values, Accuracy of 81.425%, Precision 81.843%, Recall 81.425%, F1 81.370%. It was also seen that this model was statistically different from others models in predicting using a measure known as t_score. The values of t_score was found to be 8.24 which signifies high statistical difference with 99% confidence level. Again the following model performed better than the state of art model used by Gräßer et al. (2018). The results for the Tf-Idf model using feature set from WebMD.com is also visualised as shown in Figure 12.

The third case study was to check how correlated is the sentiment of the review to that of finding the drug effectiveness for the same review. The coefficient of Pearson also known as Pearson's 'r' value, provides a linear correlation between two variables X and Y which in our case is the sentiment of the review and the effectiveness of the drug for the same review. The Pearson correlation varies from -1 to +1, where -1 suggest there is negative correlation between the two variables and a value of +1 indicates positive correlation. While a 'r' value of 0 means there is no correlation between the two variables. Upon calculating the correlation between the sentiment and effectiveness of the review, the 'r' values obtained was 0.144 which proves that there very little correlation between sentiment and effectiveness and finding the two based on the review are two different entities.

Finally, there are numerous ways in which neutral class problem can be handled during classification. On one hand, under lexicon-based analysis the neutrality of the sentences are taken into consideration to detect neutral opinion or in certain cases these sentences are filtered out focusing mainly on positive or negative opinionated classes. On the other hand, in case of binary classification, emphasis is more on subjectivity of the sentences and not on objectivity (neutrality) of a sentence. Hence depending on the type of classification under consideration, the neutral class sentences can either be considered a part of the dataset or can be completely ignored. In this research, focus is laid more on the subjectivity of drug review analysis and not on objectivity of a review, hence the neutral class (i.e reviews with ratings 3) sentences are considered.

Overall it is seen that the XGBoost model performed better than other models. It can also be seen that the feature set obtained using the Tf-idf model with its respective parameters helped in producing vectors that better suited in classifying the reviews based on effectiveness as compared to the continuous bag-of-words method of vectorization.

7 Conclusion and Future Work

In this research work, different machine learning models are used to classify the reviews, provided by the patients on online public medical forums, based on the effectiveness. The models were trained using different feature sets such as Unigram and Bigram Bag-of-Words, Unigram and Bigram Tf-Idf features that were extracted from the reviews. Promising classification results were obtained with XGBoost classifier, an ensemble classifier based on gradient boosting performing the best compared to other models. Also different approaches were followed for testing the performance validity of the models, namely in-domain classification analysis and cross-domain classification analysis. The cross-domain analysis, i.e. using the feature set of one domain to train and test the model based on another domain, helped in testing the transferability of models. Good

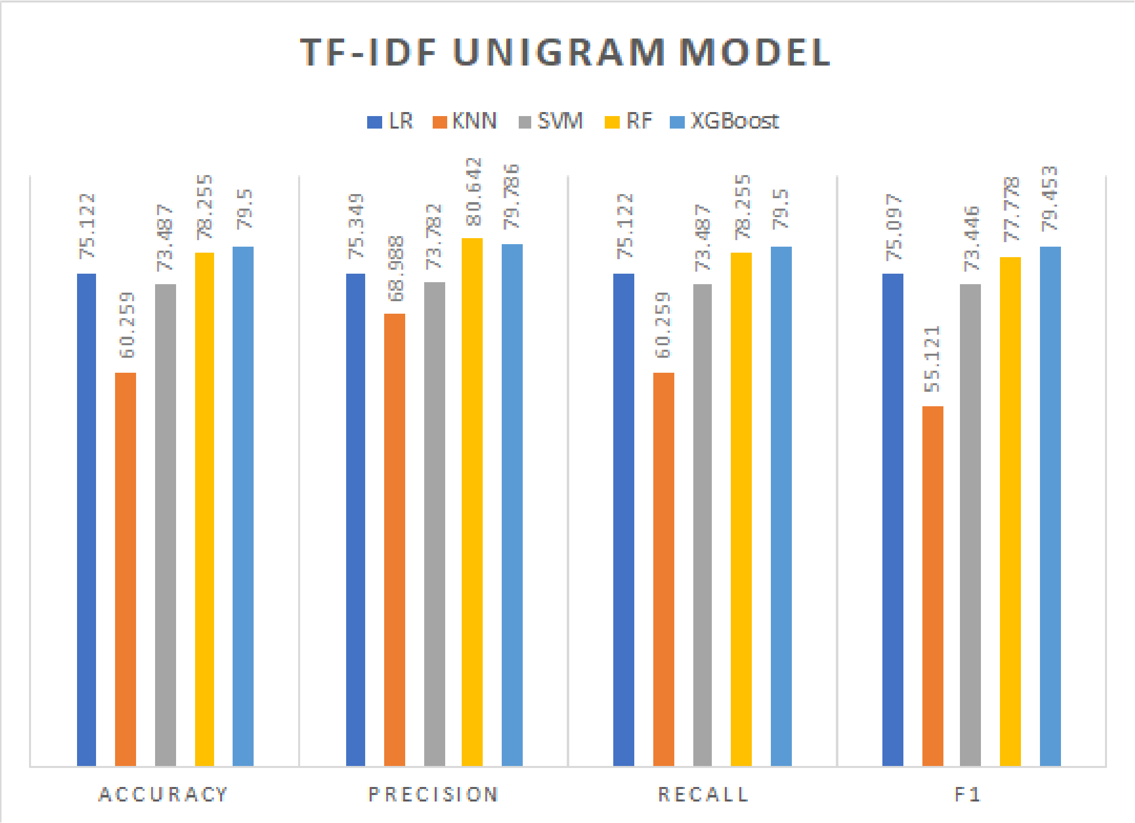


Figure 11: TF-IDF Unigram Features using all the reviews (Case study 1 Results)

classification results were obtained signifying that the model developed here can be used in future to find the effectiveness of a drug from a review in real world irrespective of the website in which the review is mentioned in.

Furthermore, correlation of sentiment analysis was compared to that of effectiveness-based analysis on the review data. It was seen that there very little correlation between sentiment of the review to that of effectiveness of the review, proving that the analysis for these approaches are different from one another.

Currently there has been very little research performed in the field of classification of reviews based on the effectiveness and therefore further improvements can be made on this model. Obtaining a labeled dataset of reviews is very rare, therefore there is a lot of potential in using the obtained feature set from this study as a part of unsupervised learning. Furthermore, the performance of classification can be improved by applying deep learning algorithms. In recent studies, promising results are seen in sentiment classification models which are developed using lexicons. Similar approach can be followed here by developing a lexicon that contain words that define the rate of efficacy. Therefore using such a lexicon, a word embedding vectorization methodology can be followed in classifying the reviews based on effectiveness.

Acknowledgement

I would like to thank my supervisor Dr. Pierpaolo Dondio for his sincere support and guidance throughout the course module and in assisting with this research study

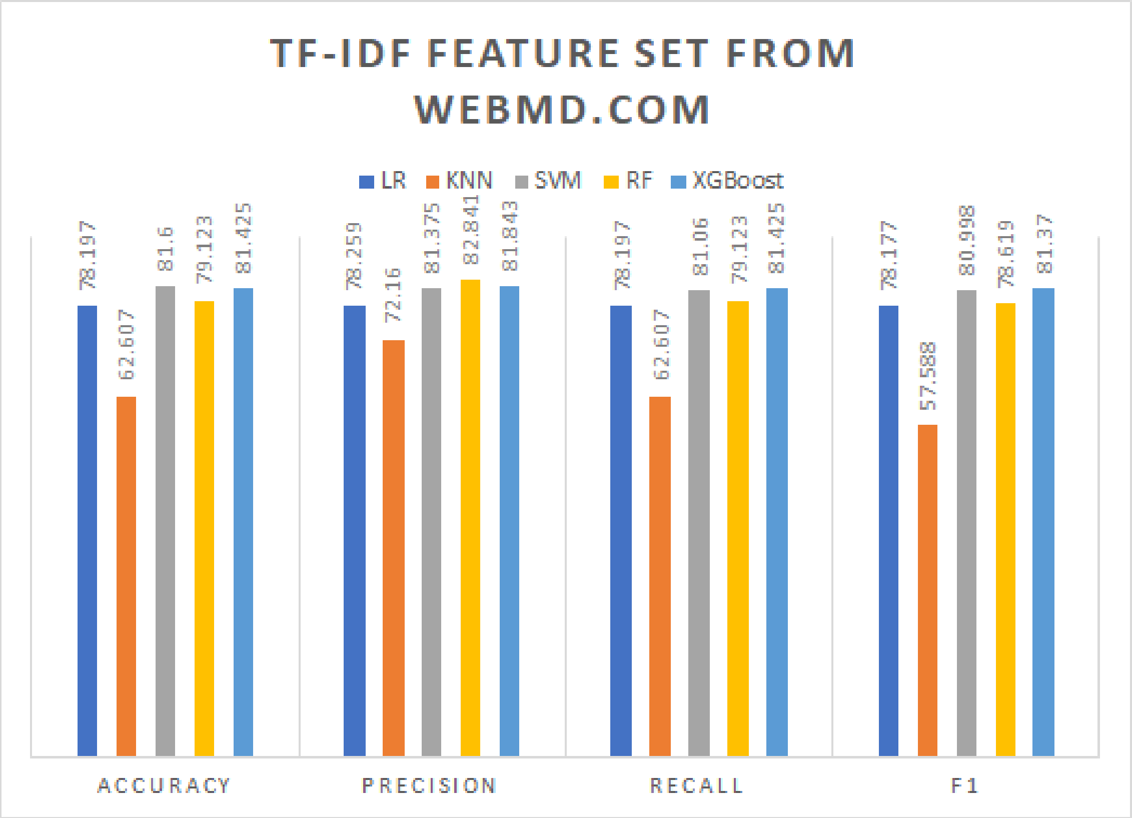


Figure 12: TF-IDF Unigram Feature set from WebMD.com reviews (Case study 2 Results)

with important guidelines and ideas. This study made use of open source software and ethically available data. I would like to thank everyone involved in making this data and software available for the research.

Finally, I would also like to thank my Parents for their support, patience and guidance through the course of my study without which it would have been impossible to complete this research.

References

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

Alimova, I., Tutubalina, E., Alferova, J. and Gafiyatullina, G. (2017). A machine learning approach to classification of drug reviews in russian, *2017 Ivannikov ISPRAS Open Conference (ISPRAS)* pp. 64–69.

Asghar, D. M., Khan, A., Ahmad, S. and Ahmad, B. (2013). Subjectivity lexicon construction for mining drug reviews, *Science International* **26**: 145–149.

Asghar, M. Z., Ahmad, S., Qasim, M., Zahra, S. R. and Kundi, F. M. (2016). Sentihealth: creating health-related sentiment lexicon using hybrid approach, *SpringerPlus* **5**.

Breiman, L. (2001). Random forests, *Machine Learning* **45**(1): 5–32.
URL: *https://doi.org/10.1023/A:1010933404324*

Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2011). SMOTE: synthetic minority over-sampling technique, *CoRR* **abs/1106.1813**.
URL: *http://arxiv.org/abs/1106.1813*

Dai, H., Touray, M., Jonnagaddala, J. and Syed-Abdul, S. (2016). Feature engineering for recognizing adverse drug reactions from twitter posts, *Information (Switzerland)* **7**(2).

Ebrahimi, M., Yazdavar, A., Salim, N. and Eltyeb, S. (2016). Recognition of side effects as implicit-opinion words in drug reviews, *Online Information Review* **40**: 1018–1032.

Fan, H. and Qin, Y. (2018). Research on text classification based on improved tf-idf algorithm.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). From data mining to knowledge discovery in databases, *AI Magazine* **17**(3): 37.
URL: *https://www.aaai.org/ojs/index.php/aimagazine/article/view/1230*

Goeuriot, L., Na, J.-C., Min Kyaing, W. Y., Khoo, C., Chang, Y.-K., Theng, Y.-L. and Kim, J.-J. (2012). Sentiment lexicons for health-related opinion mining, *Proceedings of the 2Nd ACM SIGHIT International Health Informatics Symposium*, IHI '12, ACM, New York, NY, USA, pp. 219–226.
URL: *http://doi.acm.org/10.1145/2110363.2110390*

Gopalkrishnan, V. and Chandrasekaran, D. (2017). Patient opinion mining to analyze drugs satisfaction using supervised learning, *Journal of Applied Research and Technology* **15**.

Gräßer, F., Kallumadi, S., Malberg, H. and Zaunseder, S. (2018). Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning, pp. 121–125.

Hossin, M. and Sulaiman, M. (2015). A review on evaluation metrics for data classification evaluations, *International Journal of Data Mining Knowledge Management Process* **5**: 01–11.

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

Kumar, M. R., Mani, T. T., Bodhanapu, S., Phaneendra, P. and Rahiman, F. (2011). Pharmacovigilance and its importance in drug regulation: An overview, *Der Pharmacia Lettre* **3**(2): 165–169.

Lazarou, J., Pomeranz, B. H. and Corey, P. N. (1998). Incidence of adverse drug reactions in hospitalized patients: A meta- analysis of prospective studies, *JAMA* **279**(15): 1200–1205.

Li, H., Liu, B., Mukherje, A. and Shao, J. (2014). Spotting fake reviews using positive-unlabeled learning, *Computación y Sistemas* **18**: 467–475.

Manek, A. S., Pallavi, R. P., Bhat, V. H., Shenoy, D. P., Mohan, M. C., Venugopal, K. R. and Patnaik, L. M. (2013). Sentrep: Sentiment classification of movie reviews using efficient repetitive pre-processing, *2013 IEEE International Conference of IEEE Region 10 (TENCON 2013)*, pp. 1–5.

Mishra, A., Malviya, A. and Aggarwal, S. (2015). Towards automatic pharmacovigilance: Analysing patient reviews and sentiment on oncological drugs, pp. 1402–1409.

Mohasseb, A., Bader-El-Den, M., Cocea, M. and Liu, H. (2018). Improving imbalanced question classification using structured smote based approach.

Mussa, D. and M. Jameel, N. (2019). Relevant sms spam feature selection using wrapper approach and xgboost algorithm, *Kurdistan Journal of Applied Research* **4**: 110–120.

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

Pirmohamed, M., James, S., Meakin, S. and Green, C. (2004). Adverse drug reactions as cause of admission to hospital: Authors' reply, *BMJ* **329**(7463): 460.

Sampathkumar, H., Chen, X.-W. and Luo, B. (2014). Mining adverse drug reactions from online healthcare forums using hidden markov model., *BMC medical informatics and decision making* **14**: 91.

Thabtah, F., Hammoud, S., Kamalov, F. and Gonsalves, A. (2019). Data imbalance in classification: Experimental evaluation, *Information Sciences* .
URL: *http://www.sciencedirect.com/science/article/pii/S0020025519310497*

Vimalraj, S. and Porkodi, D. R. (2018). A review on handling imbalanced data.

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

Whitehead, M. and Yaeger, L. (2010). Sentiment mining using ensemble classification models, *in* T. Sobh (ed.), *Innovations and Advances in Computer Sciences and Engineering*, Springer Netherlands, Dordrecht, pp. 509–514.

Wood, K. L. (1994). The medical dictionary for drug regulatory affairs (meddra) project, *Pharmacoepidemiology and Drug Safety* **3**(1): 7–13.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/pds.2630030105*

Zhang, J., Ye, Q., Zhang, Z. and Li, Y. (2011). Sentiment classification of internet restaurant reviews written in cantonese, *Expert Syst. Appl.* **38**: 7674–7682.