Sentiment Analysis of Drug Reviews

Rudrani Bhadra
Department of Mathematics
University of Waterloo
Waterloo, Canada
r2bhadra@uwaterloo.ca

***Abstract*—Nowadays, there are many online review sites and forums that contain a lot of information in which users state their preferences, experiences and opinions over multiple product domains. This information can be collected and mined to obtain valuable insights using data analysis and data mining methods. In this project, online user reviews in the pharmaceutical field are examined. Online user reviews in this field contain information related to various aspects such as effectiveness of drugs and side effects. This makes analysis very interesting but also very challenging.**

In this work, the Drug Review data set has been obtained from the UCI Machine Learning Repository. First, sentiment analysis is performed using various models to predict the sentiments concerning overall user satisfaction. Then, the performance of cross-domain sentiment analysis using classification models is also investigated.

***Index Terms*—Sentiment Analysis, Classification, Data Analysis, Word Cloud, Feature Extraction, Machine Learning**

I. Introduction

In online platforms such as blogs, discussion forums, user review web sites and social networking sites, there are huge amounts of user-generated content. Therefore, many researchers have been studying effective algorithms for sentiment analysis of such content. Sentiment analysis is the interpretation and classification of sentiments, opinions and subjectivity within text data using text analysis techniques. It allows businesses and companies to identify customer sentiment towards products, brands or services in online conversations and feedback. It is considered to be a very challenging problem since user reviews are described in various ways using natural language [16], [17].

For sentiment analysis, most research is done on general domains such as electronic products, movies and restaurants but not extensively on health and medical domains. Studies have shown that usually users are often looking for reviews from patients having similar problems on the internet [8]. Sentiment analysis of drug reviews will be useful not only for patients to decide which drugs they should buy or take, but also for pharmacies and clinicians to obtain valuable summaries of public opinion and feedback. Sentiment analysis can also highlight patients' misconceptions and dissenting opinions about certain drugs.

In this project, different models have been trained to classify and predict the sentiment of drug reviews according to the ratings. It is based on the work [1] and it has been extended by implementing more models and analysing the accuracies according to the different number of features and other parameters taken.

II. Background and Related Work

There are several works done on drug review sentiment analysis and most of them generally have either of these two approaches: a machine learning approach or a natural language processing approach. Xia et al. [2] developed a topic classifier from data collected from patients. They applied several polarity classifiers, one per topic. Na et al. [3] implemented a clause-level sentiment analysis algorithm by considering multiple review aspects as overall satisfaction, effectiveness, side effects and condition. Gräßer et al. [1] gathered user comments and ratings from both Drugs.com and Drugslib.com using an automatic web crawler and presented a method to predict user ratings. The data set was divided into three categories (positive, negative and neutral) based on user ratings. They applied a n-grams approach to represent the reviews and used a logistic regression model for classification. Apurva [4] used the k-means clustering algorithm to cluster the unlabeled drug reviews and grouped drugs with similar usage and benefits. Chen et al. [5] implemented sentiment classification of drug reviews by employing fuzzy rough feature selection. The work [6] analyzed the drug review data set and other online reviews data sets. They proposed deep learning methods and a novel scalable windowing approach for pairwise-similarity search to improve search efficiency. Their main objective is to improve accessibility of relevant information via social media. K et al. [7] collected data from two different websites livewell.pk and kaymu.pk and used a lexicon based approach for sentiment classification according to positive, negative or neutral type of polarity. Gopalkrishna et al. [8] analysed user drug satisfaction by using supervised learning methods. They studied three levels of polarity, classified them and compared SVM with neural network methods. In work [9], aspect based sentiment analysis of user reviews is studied on oncological drugs in which words were identified and overall sentiments derived utilizing a lexical resource. In [10], a literature review is presented on cross-domain sentiment analysis. Goeuriot et al. [11] made a sentiment lexicon related to health and used it for polarity classification of patient drug reviews. Gonzalez and Nikfarjam [12] implemented an information extraction system to extract mentions of adverse drug reactions from drug reviews by using association rule mining.

There has also been some research works done on opinion mining in the medical domain using machine learning techniques. In work [13], supervised machine learning methods were used to analyze sentiments and opinions related to health in texts written by users which were posted online on a general medical forum. Wang et al. [14] proposed a method by combining rule based classifiers and machine learning and used it for opinion classification in suicide notes. Ali et al. [15] used machine learning algorithms and subjectivity lexicons to analyse the sentiment of messages posted on forums about hearing loss.

III. Data set and Data Analysis

A. Data set

The data set is called 'Drug Review Dataset (Drugs.com) Data Set' which is obtained from the UCI Machine Learning Repository. The data set was made by crawling online pharmaceutical review sites, Drugs.com [1]. The data set contains 215,063 records. As shown in Table I, the data set contains user ratings from one to ten showing overall user satisfaction, reviews on specific drugs along with related conditions, date and also the number of users who found the review helpful.

TABLE I
Data set attributes

Attribute	Type	Description
drugName	categorical	name of drug
condition	categorical	name of condition
review	text	patient review
date	date	date of review entry
usefulcount	numerical	number of users who found reviews helpful.

B. Exploratory Data Analysis

The first step is to understand how the user ratings are distributed. As shown in Fig 1, rating 10 is the most frequent rating that the users gave which has 68,005 samples, followed by ratings nine and one which have 36,708 and 28,918 samples respectively. Hence, it can be concluded that the drug users tend to give extreme ratings. Also it shows that if the users are not satisfied, they usually give a rating one instead of two or three. The overall user ratings' mean is 6.99 and the standard deviation is 3.28.

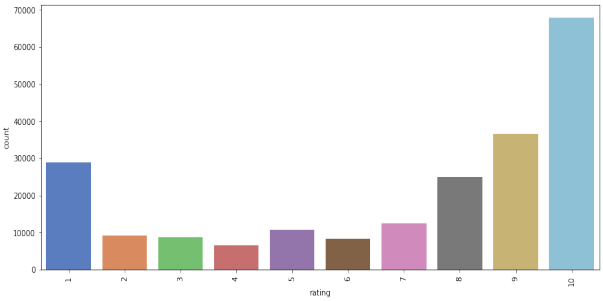

Fig. 1. Distribution of counts for different ratings

After that, the next step is to see if different dates effect user ratings. As shown in Fig 2, it is observed that the differences in average ratings are large. Year 2008 has the highest average rating, 8.93 and year 2017 has the lowest average rating, which is 6.04. However, there is no obvious effect of years to ratings. Similarly in Fig 3 and Fig 4 of reviews has similar views with respect to months and days respectively.

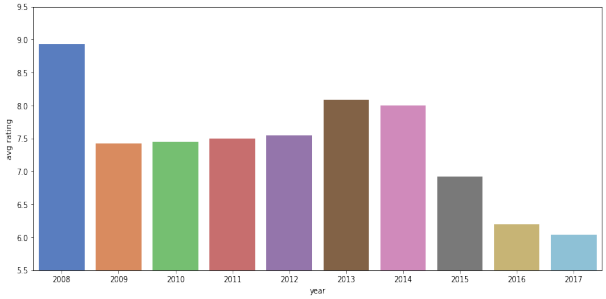

Fig. 2. Average user rating for each year

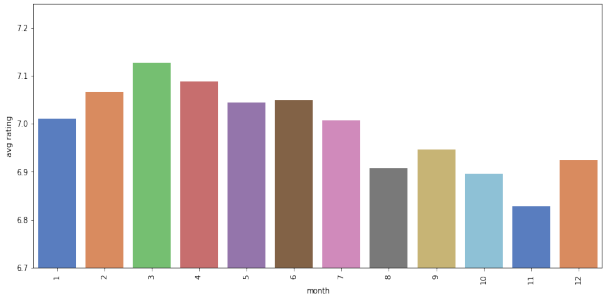

Fig. 3. Average user rating for each month

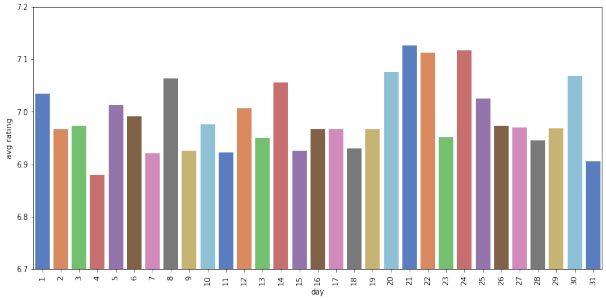

Fig. 4. Average user rating for each day

Fig 5 shows the word cloud. It is found that some words are general words like 'first', 'took', etc which are not very useful in prediction. Words such as 'prescribed', 'diagnosed' can be useful in the predictive model. There also some medical words like 'depressants' and 'antibiotics' and also some meaningless words like 've'. To make common words less influential to the models, bag-of-counts and TF-IDF scores can be used to represent the word features. To simplify the training process, words which are rare can be ignored and the most frequent

words can be used which occur in the data set. Some attributes of the data set found after analysis are shown in Table II.

TABLE II
DRUG REVIEW DATASET BASIC ATTRIBUTES

Attribute	Value
Length of Dataset	215,063
Average rating	6.99
Unique conditions	916
Unique uni-grams in reviews	88,495
Unique bi-grams in reviews	1,225,707

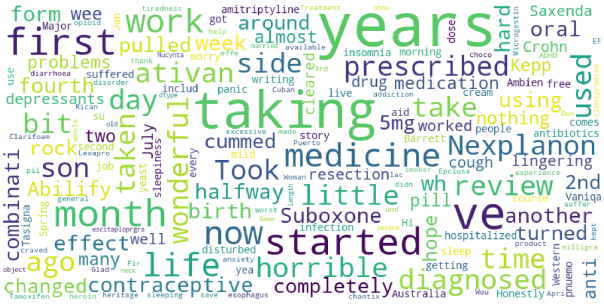

Fig. 5. Word cloud for drug review dataset

IV. METHODOLOGY

A. Data set Preprocessing

First, the user ratings were converted to three classes which represents the polarity of a user's sentiment regarding the medication: reviews with rating ≥ 7 were labelled as positive, reviews with rating ≤ 4 were labelled as negative, reviews with rating between 4 and 7 were labelled as neutral [1].

Then, for all the prediction tasks, a N-grams approach was applied to all user reviews. Single tokens (uni-grams) as well as two or three adjacent tokens (bi-grams, tri-grams) were used to derive the features for classification. Based on the total collection of occurring N-grams, each review can be represented as a sparse vector of token counts or token frequencies.

Initially, all reviews were converted to lowercase. All special characters, punctuation and numbers were removed. Stop words were removed from the reviews except stop words such as 'aren't', 'didn't', etc as they are important in analysing negative reviews.

B. Feature Extraction

The preprocessed reviews were tokenized on spaces to obtain the overall vocabulary and a feature space representation of each review. Here, Bag-of-words and TF-IDF are used for feature extraction.

Bag-of-words. In the Bag-of-words model, the number of times each word appears in a document is counted and then the frequency that each word appears in a document out of all the words in the document is calculated. In this experiment, not only uni-grams but also bi-grams and tri-grams are counted.

TF-IDF. In the TF-IDF model, the TF-IDF score of each word in the review is calculated. The following equations are used to calculate TF-IDF:

$$IDF(t, D) = log(N/DF(t, D)) \quad (1)$$

$$TF\text{-}IDF(t, d, D) = TF(t, d) * IDF(t, D) \quad (2)$$

where TF(t,d) = number of times a word t appears in document d, DF(t,D) = number of documents D that contain word t, N = total number of documents.

C. Evaluation Metric

For supervised classification of an unbalanced data set, accuracy/AUC is not optimal. However, in the next section it will be mentioned that a roughly balanced balanced data set will be sampled from the original data set to achieve a relatively balanced one. Thus, the accuracy score would be the main evaluation metric for the data set.

D. Data Sampling

Since the data set is an imbalanced one, it was found that after preprocessing, there are around 140,000 positive reviews, 58,000 negative reviews and 20,000 neutral reviews. Since the number of positive reviews is huge compared to the negative and neutral reviews, 60,000 positive reviews are randomly sampled from the 140,000 positive reviews and paired with the negative and neutral reviews. The numbers of the labels are shown in Fig 6. Even after sampling, the count of positive and negative labels are considerably unbalanced with respect to the count of neutral labels. Since the distribution is disproportionate, classification errors were penalized with a weight inversely proportional to the class frequency during training [1].

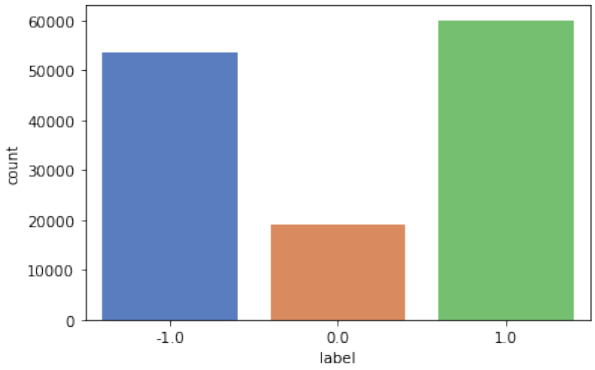

Fig. 6. Number of labels in data set

E. Model Selection

In order to predict user sentiment by using the extracted feature representations, the following classification models were used: logistic regression, multinomial naive bayes, random forest and gradient boosting.

Logistic regression is used since it was the model which was used in work [1] and hence it can be compared against the other models chosen.

Multinomial naive bayes is chosen since it is the fastest model that can be computed and its accuracy is comparable to other models.

Random forest is selected as one of the classifiers as it builds trees in parallel to each other which are not correlated to each other and applies the general technique of bootstrap aggregating to tree learners. This leads to better model performance because it decreases the variance of the model without increasing the bias.

Gradient boosting is selected due to its effectiveness. Like other boosting methods, gradient boosting combines weak learners into a strong single learner in an iterative manner.

Since text reviews usually contain a lot of noise and meaningless words, the decision tree will be prone to overfitting easily and will not be suitable for this predicting task.

V. Experimental Results

A. In-domain Sentiment Analysis

In this section, the main aim was to predict the overall patient satisfaction by employing different classification-based sentiment analysis.

To calculate the accuracy of each method, the data set is split into training and test data sets. 75% of the data set is used for training the model and the remaining 25% is used for testing.

TABLE III
Data Description for Model Evaluation

Train	Test	Rating	Label	%
108978	36326	rating ≤ 4	-1	37
		$4 <$ rating < 7	0	22
		rating ≥ 7	1	41

The data set is divided into two stream approaches for training: TF-IDF and bag-of-words count. They are used to process the data and feed them into the classification models.

Table IV shows the experimental results of various models with different combination of features. It is observed that the random forest model outperforms the logistic regression as it does not assume linear relationships between the variables. It performs better than the multinomial naive bayes as there is no underlying naive assumptions about the data. It also beats the gradient boosting model in terms of accuracy. For this, the reason is that, because of the limitation of computing time and computing resources, the parameters of gradient boosting model are not fine-tuned. In addition to direct comparison on the solution quality between models, some experiments are performed within each model.

1) Interpretation of parameters in Logistic Regression: When the logistic regression model is trained using TF-IDF representation, I was interested to know the most important words (features) in each class and to check if the model is reasonable. As shown in Fig 7, it can be observed that the most important words in the negative class are words like 'not recommend', 'worst', 'never' and 'stay away'. As seen in Fig 8, the most important words in the neutral class are words such as 'hope', 'not sure', 'yet'. Fig 9 shows the most important words in the positive class such as 'love', 'great', 'amaz'. So the top important features in each class are reasonable.

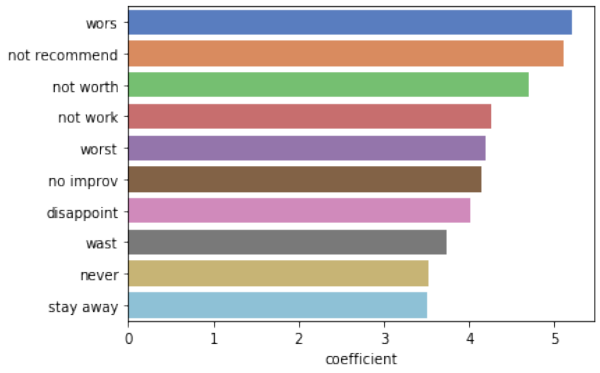

Fig. 7. Line chart of top-ten most important features for negative class in logistic regression with 30000 of TF-IDF features

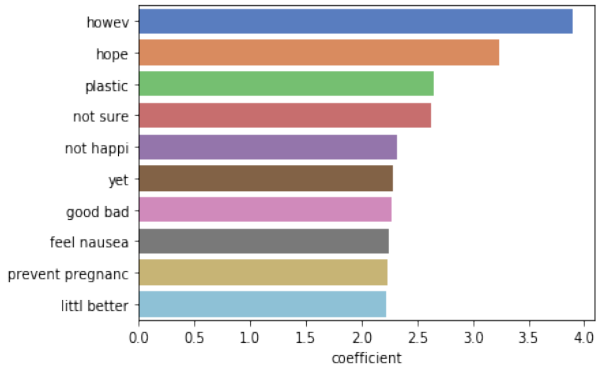

Fig. 8. Line chart of top-ten most important features for neutral class in logistic regression with 30000 of TF-IDF features

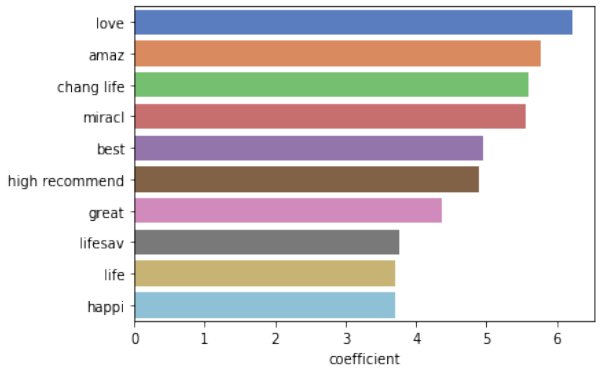

Fig. 9. Line chart of top-ten most important features for positive class in logistic regression with 30000 of TF-IDF features

TABLE IV
IN-DOMAIN SENTIMENT ANALYSIS: MODEL COMPARISON

Model	Features	Fixed Parameters	Avg Training Accuracy%	Avg Testing Accuracy%.
Logistic Regression	Bag-of-words (uni- & bi- & tri-gram)	-	92.95	78.16
	TF-IDF (uni- & bi- & tri-gram)	-	82.63	73.82
Multinomial Naive Bayes	Bag-of-words (uni- & bi- & tri-gram)	-	76.63	69.36
	TF-IDF (uni- & bi- & tri-gram)	-	73.71	71.62
Random Forest	Bag-of-words (uni- & bi-gram)	N=10	99.36	81.15
		N=50	99.92	85.01
		N=100	99.92	85.51
	TF-IDF (uni- & bi-gram)	N=10	99.37	81.11
		N=50	99.92	84.97
		N=100	99.92	85.50
Gradient Boosting	Bag-of-words (uni- & bi-gram)	W=10000	71.81	69.76
	TF-IDF (uni- & bi-gram)	W=10000	72.71	70.14

2) Bag-of-words Versus TF-IDF using Logistic Regression: When training the logistic regression model, I wanted to see whether different textual representation models would give different solutions. The number of features was varied from 10000 to 55000 and the changes in accuracy is observed in Fig 10. It can be observed that as the number of features increase, the training accuracy for both bag-of-words and TF-IDF keeps increasing which is a case of overfitting. However, the maximum value of the test accuracies in this case are around 80% and after 73% for bag-of-words and TF-IDF respectively.

When training the logistic regression model, it is observed that the training and test accuracies when using the bag-of-words model are much higher than using the TF-IDF model as seen in Table IV.

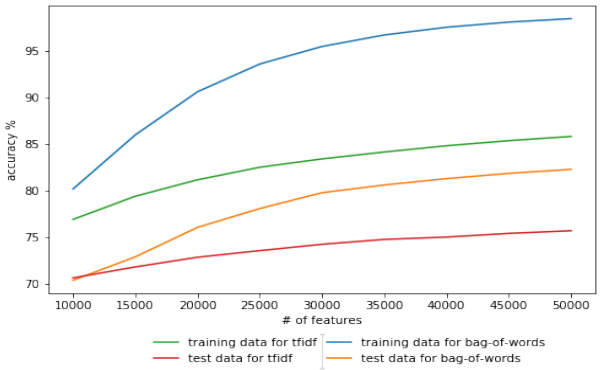

Fig. 10. Line chart of experiment on different textual representation models with logistic regression

3) Bag-of-words Versus TF-IDF using Multinomial Naive Bayes: When training the multinomial naive bayes model, I wanted to see whether different textual representation models would give different solutions. The number of features is varied from 10000 to 55000 and the changes in accuracy is observed in Fig 11. It can be observed that as the number of features increase, the training accuracy for bag-of-words model keeps increasing which is a case of overfitting. The training accuracy for the TF-IDF model does not increase as steeply as the bag-of-words model. The maximum value of the test accuracies in both cases are around 70% and 72% for bag-of-words and TF-IDF respectively.

When testing the multinomial naive bayes model, it is observed that the test accuracy when using the TF-IDF model is slightly higher than using the bag-of-words model as seen in Table IV.

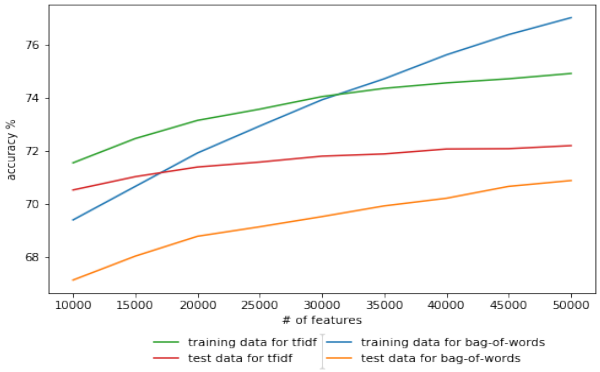

Fig. 11. Line chart of experiment on different textual representation models with multinomial naive bayes

4) Random Forest and Feature Importance: Fig 13 and Fig 14 show the most important textual representation features when taking the TF-IDF model. It can be observed that the most useful features are uni-grams such as 'love', 'great' and 'no' and have a strong sentiment associated with them. On the other hand, most of the useless features are bi-grams and do not contain heavy tendency.

5) Bag-of-words Versus TF-IDF using Random Forest: As shown in Fig 12, the number of bag-of-words and TF-IDF features does not have a major effect on the solution quality. Nevertheless, the number of trees in the forest does, which improves the test accuracy by around 4.0. When the number of trees is taken to be 10, the test accuracy is found to be around 81.5%. When the number of trees is 50 and 100, the test accuracies are around 85%.

When training the random forest model, it is observed that the training and test accuracies when using the bag-of-words model and the TF-IDF model are not much different as seen in Table IV.

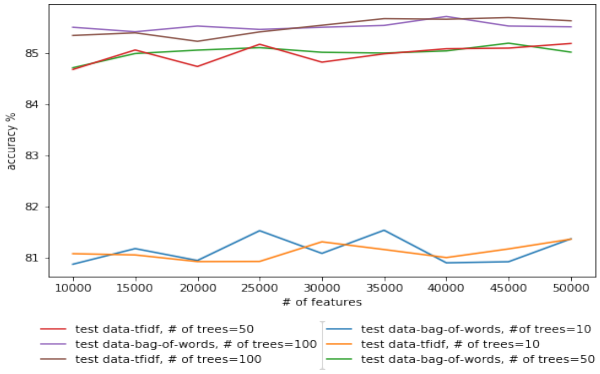

Fig. 12. Line chart of experiment on different textual representation models with random forest

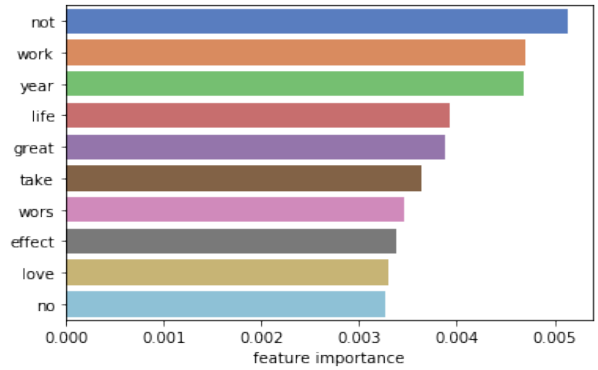

Fig. 13. Line chart of top-ten most important features in random forest with 30000 of TF-IDF features

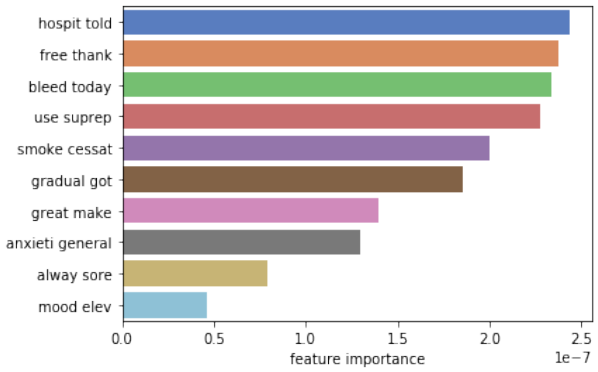

Fig. 14. Line chart of top-ten least important features in random forest with 30000 of TF-IDF features

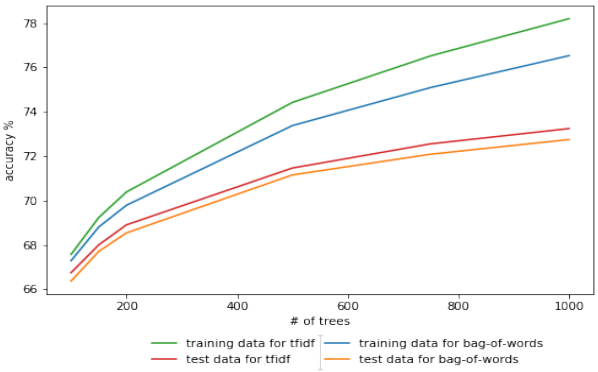

Fig. 15. Line chart of experiment on different textual representation models with gradient boosting

6) Gradient Boosting: As shown in Fig 15, the number of trees in the gradient boosting algorithm does have an effect on the solution quality. Here, the number of features is kept fixed at 10000, while the number of trees is varied in order to analyse its effect on the accuracy. It is observed that the number of trees to fit in the gradient boosting can improve the accuracy by quite a bit. There are still ways to improve the gradient boosting model. The number of trees to fit the model can be increased if more computing resources and computing time are available.

When training the gradient boosting model, it is observed that the training and test accuracies when using the bag-of-words model are not much different than using the TF-IDF model as seen in Table IV.

B. Cross-domain Sentiment Analysis

In this section, the main aim is to study the performance of models built on data from one medical condition (source domain) and evaluate on data related to other conditions (target domain) [1], [10]. Here, overall user satisfaction models were trained on drug review subsets related to one specific condition. Then the performance of these trained domain models were evaluated on other condition related subsets. The domains (subsets of the conditions) were selected by taking the five most frequent conditions present in the data set. These were Birth Control, Depression, Pain, Anxiety and Acne.

For this problem, logistic regression model and the entire data set is used for this analysis. The results are summarized in Table V. The results show that the selected domain has a notable impact on the performance of the classifier when it is applied to data from other domains. It is obvious from the analysis done that in-domain training and testing analysis clearly outperforms the performance of cross-domain models. This shows and emphasizes the hypothesis of domain specific vocabulary [1]. From Table V, it is observed that a specific domain trained model performed the best when it was given its own domain to be tested. However, the model trained on Acne data seems to generalize better on other domain data as compared to models trained on other domain data. It also can be seen that there are some combinations which show better performances than others for e.g. Depression and Anxiety, Pain and Anxiety, Acne and Anxiety, Acne and Pain, as compared to combinations like Depression and Birth Control. This can be due to the fact that certain side effects, expressions and domain specific vocabulary used by patients are common among certain conditions. Moreover, medical fields dealing with conditions like Depression and Anxiety are closely related. From drugs concerning Depression and

TABLE V
CROSS-DOMAIN SENTIMENT ANALYSIS

Train Data	Test Data					Avg Train Accuracy %
	Depression	Anxiety	Pain	Birth Control	Acne	
Depression	84.05	71.63	70.61	49.67	67.34	68.66
Anxiety	68.69	84.84	77.25	52.36	70.84	70.80
Pain	68.72	77.88	83.80	52.36	70.84	70.72
Birth Control	45.67	34.76	38.84	82.98	38.83	48.22
Acne	68.69	77.88	75.65	52.30	82.24	71.35
Avg Test Accuracy %	67.16	69.40	69.23	57.93	66.02	65.95

Anxiety, 33 drugs are applied in both conditions whereas for Birth Control and Anxiety there is no overlap. If the problem is reduced to a binary classification task instead of a three class problem, it would increase the accuracy substantially.

VI. CONCLUSION

Within this project, the application of machine learning based sentiment analysis of patient generated drug reviews has been studied. Several classification models were trained using simple lexical features such as uni-grams, bi-grams and tri-grams extracted from the user reviews.

In in-domain sentiment analysis, the model which performed the best overall is random forest, followed by logistic regression. Although it achieved decent results, it can be improved in a number of ways:

1. More hyperparameter tuning can be done for all the base classifiers by using either grid search or other cross validation methods. Due to time and resource limitations, it was decided to analyse each model by varying the parameters instead of explicitly tuning each model, which requires a lot of computational resources given the length of the data set.
2. Applying more powerful machine learning models such as deep learning approaches can further improve this solution.
3. Here, only the reviews were taken as features. Other attributes of the data set such as year, month and date can be taken alongside with each text review.
4. In the bag-of-words model, the word count is dependent on the review length so it might be better to remove the effect of text length by dividing the words count by the text length.

As labeled data sets for building classification models are hard to find or are available in an unstructured format, various approaches for model portability are still being investigated. Although in-domain classification training and testing showed good classification results, the performance of models trained on one specific condition and tested on another condition varies among different domains. However, conditions which belong to similar medical fields and are partly treated with equal medications, also show higher potentials for model transferability. In this case, cross-data sentiment analysis (training and testing classifiers on data from different sources) can be further studied and analysed by taking different data sets in the same domain as done in [1].

REFERENCES

[1] F. Gräßer, S. Kallumadi, H. Malberg, and S. Zaunseder, "Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning", in DH'18: 2018 International Digital Health Conference, April 23–26, 2018, Lyon, France. ACM, New York, NY, USA, pp. 121–125, 2018.

[2] L. Xia, A. L. Gentile, J. Munro, and J. Iria, "Improving Patient Opinion Mining through Multi-step Classification", in Proceedings of the 12th International Conference on Text, Speech and Dialogue (TSD '09). Springer-Verlag, Berlin, Heidelberg, pp. 70–76, 2009.

[3] J. C. Na and W. Y. Min Kyaing, "Sentiment Analysis of User Generated Content on Drug Review", Journal of Information Science Theory and Practice, pp. 6–23, Mar 2015.

[4] B. Apurva, "Grouping of Medicinal Drugs Used for Similar Symptoms by mining Clusters from Drug Benefits Reviews", Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM), Amity University Rajasthan, Jaipur - India, February 26-28, 2019.

[5] T. Chen, P. Su, C. Shang, R. Hill, H. Zhang, and Q. Shen, "Sentiment Classification of Drug Reviews Using Fuzzy-rough Feature Selection", in 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, pp. 1–6, 2019.

[6] I. Inuwa-Dutse, M. Liptrott, and I. Korkontzelos, "A deep semantic search method for random tweets", Online Social Networks and Media, vol. 13, 100046, 2019.

[7] K. Mahboob and F. Ali, "Sentiment Analysis of Pharmaceutical Products Evaluation Based on Customer Review Mining", Journal of Computer Science & Systems Biology, 11(3), pp. 190–194, 2018.

[8] V. Gopalakrishnan and C. Ramaswamy, "Patient opinion mining to analyze drugs satisfaction using supervised learning", Journal of Applied Research and Technology, 15(4), pp. 311–319, 2017.

[9] A. Mishra, A. Malviya, and S. Aggarwal, "Towards Automatic Pharmacovigilance: Analysing Patient Reviews and Sentiment on Oncological Drugs", in 2015 IEEE International Conference on Data Mining Workshops, pp. 1402–1409, 2015.

[10] T. Al-Moslmi, N. Omar, S. Abdullah, and M. Albared, "Approaches to Cross-Domain Sentiment Analysis: A Systematic Literature Review", IEEE Access 5, pp. 16173–16192, 2017.

[11] L. Goeuriot, J. C. Na, W. Y. Min Kyaing, C. Khoo, Y. K. Chang, Y. L. Theng, et al, "Sentiment lexicons for health-related opinion mining", in Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, ACM, pp. 219–226, 2012.

[12] M. Sokolova and V. Bobicev, "Sentiments and opinions in health-related web messages", in RANLP, pp. 132–139, 2011.

[13] A. Nikfarjam and G. H. Gonzalez, "Pattern mining for extraction of mentions of adverse drug reactions from user comments", in Proceedings of AMIA Annual Symposium, pp. 10192–1026, 2011.

[14] W. Wang, L. Chen, M. Tan, S. Wang and A. P. Sheth,"Discovering fine-grained sentiment in suicide notes", Biomedical Informatics Insights, 5(1), pp. 137–144, 2012.

[15] T. Ali, M. Sokolova, D. Schramm and D. Inkpen, "Opinion learning from medical forums", in RANLP, pp. 18–24, 2013.

[16] M. K. Das, B. Padhy and B. K. Mishra, "Opinion Mining and Sentiment Classification: A Review", International Conference on Inventive Systems and Control, 2017.

[17] R. Hu, L. Rui, P. Zeng, Lei Chen and X. Fan, "Text Sentiment Analysis: A Review", 2018 IEEE 4th International Conference on Computer and Communications, pp. 2283–2288, 2018.