https://doi.org/10.48009/3_iis_2020_59-65

Issues in Information Systems
Volume 21, Issue 3, pp. 59-65, 2020

CONTEXT BASED SENTIMENT ANALYSIS APPROACH USING N-GRAM AND WORD VECTORIZATION METHODS

Parthasarathi Tumu, University of St Thomas, part6550@stthomas.edu
Vaghindra Manchenasetty, University of St Thomas, vaghi.manchenasetty@stthomas.edu
Manjeet Rege, University of St Thomas, rege@stthomas.edu

ABSTRACT

Consumer reviews are key indicators for product credibility and central to almost all product manufacturing companies to align and alter the products to the needs of customers. Using Sentiment analysis approach, these reviews can be analyzed for positive, negative and neutral feedback. There are many techniques designed to do Sentiment analysis and opinion mining in the past on drug reviews to study their effectiveness and side-effects on the people. In this paper, an approach is presented which is a combination of context-based sentiment analysis using N-gram and tf-idf word vectorization method to find the sentiment class – positive, negative, neutral and use this sentiment class in Naïve Bayes and Random Classifiers to predict user review emotion. Our validation process involved measuring the model performance using quality metrics. The results showed that the proposed solution outperformed conventional sentimental analysis techniques with an overall accuracy of 89%.

Keywords: Sentiment Analysis, N-gram, Naïve Bayes, Tf-idf, Random classifiers

INTRODUCTION

In today's world with so many variations of same product available in the market, user reviews and ratings are becoming pivotal for consumers making decisions about buying and using a product. The user reviews sprawling across internet, provide plethora of information for researchers and product manufacturing companies. Opinion mining and Sentiment analysis is key in these research activities. The research involves mining user reviews to gain insights into people's choices, beliefs and opinions of products to a certain degree called sentiments.

Machine learning has permeated nearly all fields and disciplines of study. Natural Language Processing and Sentiment analysis is one of the major disciplines of research to identify, extract, and make use of subjective (text) information. Text collected from open-ended questions can be valuable feedback to a Medical research institution.

N-gram method is a group of n words or characters (for pieces of grammar) which follow one another. So an n of 4 for the words from sentence would be like "I don't like movie". This method is used to create an index of how frequently words follow one another. Higher the number, and output will be total copy of the original. If it is too low, and the output will be messy. To understand the user review reviews input data set divided in to two based review score. we classified 1 ~ 5 as negative, and 6 ~ 10 as positive, and we checked through 1 ~ 5 grams which corpus best classifies emotions**.**

Our goal is to predict the sentiment of a given review (Bakshi, R. K., Kaur, N., Kaur, R., & Kaur, G, 2016). The obvious starting question for such an approach is how we can convert the raw text of the review into a data representation that can be used by a numerical classifier. Vectorization is the widely used process for text to numerical conversion. By vectorizing the 'review' column, we can allow widely varying lengths of text to be converted into a numerical format which can be processed by the classifier. This is achieved via the Term-Frequency Inverse Document-Frequency (TF-IDF) method.TF-IDF method handles noise words in the reviews which do not provide much insight or add any value to the sentiment of the review. In fact, their high frequency tends to obfuscate words that provide significant insight into sentiment.

Random Forests (Will Koehrsen, 2017) or Random decision forests are ensemble learning for classification. In random forests [fig (1)], several decision trees are built during training. Each tree is built from the sample obtained

from the training using bootstrap. The output of the random forests will be the mode of the classes of individual decision trees. This averaging process solves the problem of overfitting in decision trees.

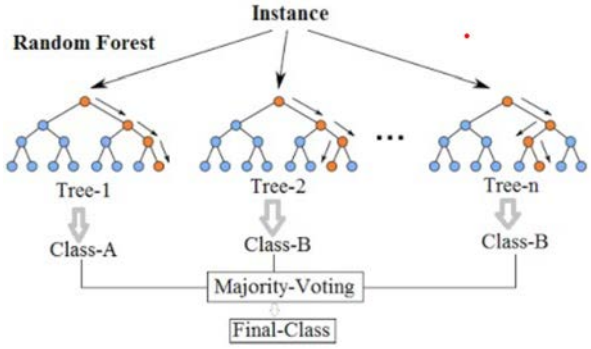

Figure 1. Random Forest Machine Learning Model

RELATED WORK

The ongoing research work related to the Sentiment Analysis are given in this section. Angulakshmi, ManickaChezian, (2014), propose 3 step process for feature extraction, Extracting initial Opinion, Conjunction-Based Opinion, Double Propagation Extraction. Also highlights having importance of domain specific opinion words will improve the model efficiency to correctly classify the user review sentiment. Kherwa, Sachdeva, Mahajan (2014), provides insights into sentimental analysis system that automatically extract the related features of a given review to be analyzed, and create sentiment summary for each feature talked in a review. This information is very useful for product manufactures, quickly analyze the product performance and user needs. Rana, Cheah (2015), suggest rule based hybrid method. It finds sequential patterns and Normalized Google Distance to obtain aspects of a review. Aspect based opinion mining main focus is on extracting aspects from customer reviews and rank the aspects as positive or negative. Kumar, Sachdeva, (2014), explains an automated process to collect users reviews for a given product and create summarized dashboard with user sentiment scores. Zhan, Wang and Zhan, (2020), suggest that public opinion related to the focal enterprise is spread widely on the Internet, but it cannot be easily collected and managed accurately. This research aims to employ natural language processing (NLP), sentiment analysis and data mining technologies to build a public opinion analysis system to serve enterprises' need of online public opinion detection. K. Patel *et al*. (2020) discuss that with the improvement and higher efficient algorithms in machine learning and deep learning, the prediction of various critical applications in computer vision become reality. Facial recognition-based sentiment analysis getting popular now. Analysis based computer vision using Deep learning is most fast-moving research area.

PROPOSED METHOD

In this work, we propose a method to identify how sentiment plays into rating and usefulness of review and predicting the sentiment or rating based on the review. The sequence of processes followed to achieve the goal is shown in fig (2).

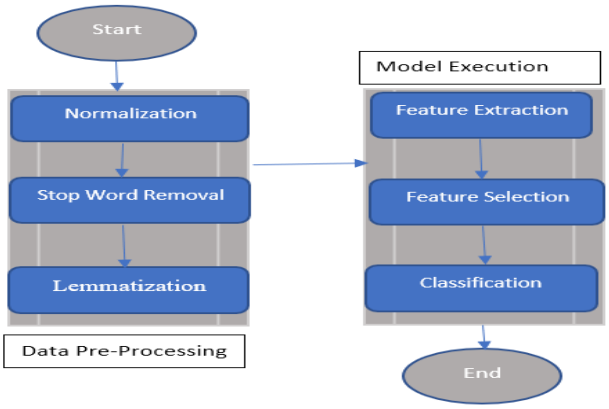

Figure 2. Proposed Algorithm approach

a) Dataset

The UCI ML Drug Review dataset provides patient reviews on specific drugs along with related conditions and a 10-star patient rating system reflecting overall patient satisfaction. The data was gathered by crawling online pharmaceutical review sites. This data was published in a study on sentiment analysis of drug experience over multiple facets, ex. sentiments learned on specific aspects such as effectiveness and side effects.

b) Data set characteristics and Preprocessing:

Data preprocessing is the major step in data exploration and predictive analysis. The performance and success of a model is directly related to the quality of data supplied to it. The current dataset [table (1)] has 7 features and 160K observations(records) in the training data set.

Table 1. Dataset Features

Features	Description
uniqueID	Unique ID
drugName	Name of drug
condition	Name of condition
review	Patient review
rating	10 star patient rating
date	Date of Review of Entry
usefulCount	Number of users who found review useful

- Missing Values:

  Form the analysis it is noticed that, missing values for ‘condition ‘ is 0.5% .No missing values for other features. Considering the size of input data, very low impact of missing values, so decided to remove all rows with missing values.

- Clean Up erroneous data:

  The feature “Condition” has erroneous text (ex: '') for about 0.6% of the data. Since this constitutes insignificant number of rows and also doesn’t provide any value to the data analysis, decided to remove these records.

- Convert all HTML Symbols in to Words, convert to Lower case, remove punctuation and numbers :

Some of the reviews has html tags like ''s'. For sentiment analysis it is important have only text in the reviews. So using HTMLParser converted special characters in to plain text.

- Apply Lemmatization:

  The process of grouping together the inflected forms of a word so they can be analyzed as a single item, identified by the word's lemma, or dictionary form.

c) Data Exploration:

The user ratings are rated from 1 to 10. Most of the reviews fall in category of ratings ranging from 1-2 and 8-10. This distribution illustrates that people generally write reviews for drugs they really like (or those that they really dislike). There are fewer middle ratings as compared to extreme ratings. Birth Control is most reviewed condition.

Correlation between Rating and usefulness of the review: Based on the data people found reviews with higher scores to be more useful. In the sense that reviews with high ratings received more 'useful' tags than reviews with low ratings. The number of drugs per condition: Data suggests that the number of drugs for top eight conditions is roughly about 100 for each condition. There are few conditions where there is only one drug prescribed. Considering the recommendation system, it is not feasible to do a recommendation for those conditions. Therefore, we will analyze only the conditions that have at least 2 drugs per condition.

Understanding User Reviews:
Using Word Cloud to show insights into what words are most important for a high and a low rated review. This is important for user sentiment analysis – differentiating between positive and negative reviews.

Word cloud using n-gram method:

Used word cloud data[fig(3)] visualization method to view the key word from positive and negative reviews. Based results, n-gram =5 gave us best results to understand the sentiment on both positive and negative review.

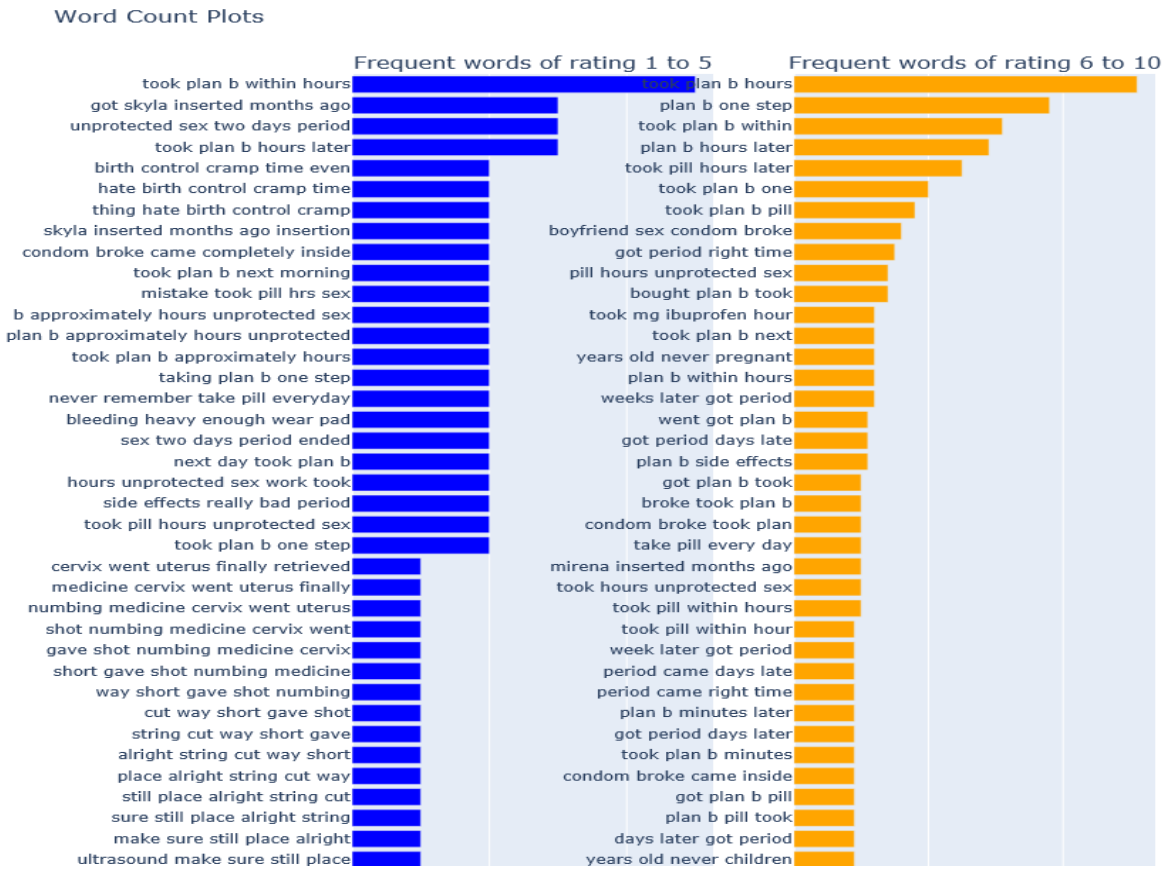

Figure 3. 5-gram model output

MACHINE LEARNING MODEL EXECUTION AND RESULTS

Our goal is to predict the sentiment of a given review. A new target variable 'sentiment 'is created. All reviews rated between 6 to 10 are considered as favorable and flagged as "Positive". Reviews rated between 1 to 5 considered as negative and flagged as "Negative". Based on target variable we considered this as classification model.

Naïve Bayes Classification:
The simplest type of model we can attempt to fit on this data is the Naive Bayes classifier. Naive Bayes on a binarized version of the rating column which attempts to identify which reviews are favorable.
Results with Naive Bayes classifier: (Accuracy: 83%)

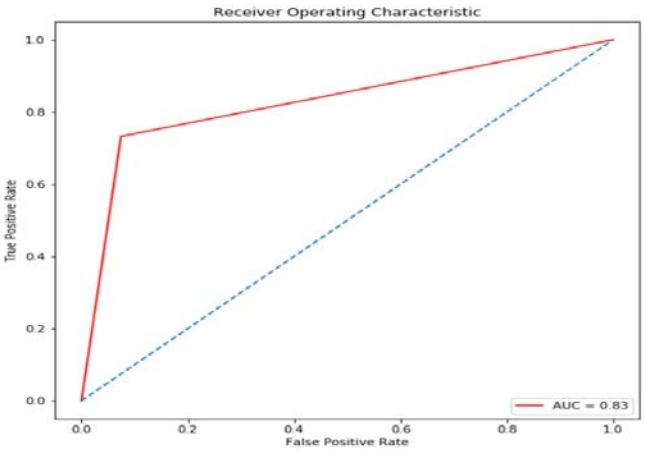

Figure 4. Naive Bayes classifier model performance measurement

Random Forest Classification(RFC):
RFC is made up of 100 Decision Trees (Pedregosa et al ,2020) .We arrived at 100 Decision Trees after multiple tests at 25,50, and 75. We have seen a significant increase in accuracy from 25 to 75 Decision Trees. Results with Random forest Classifier: (Accuracy: 89%)

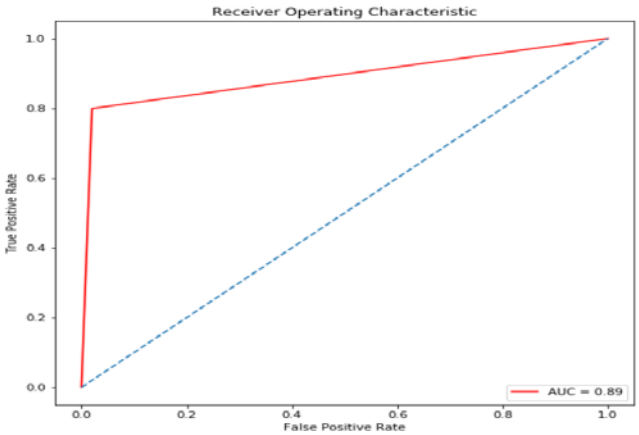

Figure 5. Random-Forest model performance measurement

CONCLUSION

We set the topic as recommending the right medicine for the patient's condition with reviews and proceeded with the data preprocessing, data exploration, and modeling. In the data exploration section, we looked at the forms of data using visualization techniques and statistical techniques. We also looked for n-grams that can best represent emotions, and the relationship with date and rating. The next step was to preprocess the data according to the topic we set, such as removing the condition that has only one drug for recommendation. In the process of modeling, we used Classification models to predict the sentiment.

LIMITATIONS

Process though outperforms compared other sentiment analysis methods, we wanted to outline few limitations with this approach in this paper:

a) Sentiment analysis using sentiment word dictionary has low reliability when the number of positive and negative words is small. For example, if there are 0 positive words and 1 negative word, it is classified as negative. Therefore, if the number of sentiment words is 5 or less, we could exclude the observations.

b) To ensure the reliability of the predicted values, we normalized useful Count and multiplied it to the predicted values. However, useful Count may tend to be higher for older reviews as the number of cumulated site visitors increases. Therefore, we should have also considered time when normalizing useful Count.

c) If the emotion is positive, the reliability should be increased to the positive side, and if it is negative, the reliability should be increased toward the negative side. However, we simply multiplied the useful Count for reliability and did not consider this part. So, we should have multiplied considering the sign of useful Count according to different kinds of emotion

REFERENCES

G.Angulakshmi, Dr.R.ManickaChezian, (2014) "An Analysis on Opinion Mining: Techniques and Tools." International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 7.

PoojaKherwa, ArjitSachdeva, Dhruv Mahajan (2014) "An approach towards comprehensive sentimental data analysis and opinion mining". IEEE International Advance Computing Conference (IACC).

Toqir Ahmad Rana, Yu-N Cheah (2015) "Hybrid Rule-Based Approach for Aspect Extraction and Categorization from Customer Reviews", 9th International Conference on IT in Asia (CITA'15)

Prashast Kumar, ArjitSachdeva, (2014)"An approach towards feature specific opinion mining and sentimental analysis across e-commerce websites".

G. Zhan, M. Wang and M. Zhan, (2020) "Public Opinion Detection in an Online Lending Forum: Sentiment Analysis and Data Visualization," IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)

K. Patel *et al*. (2020), "Facial Sentiment Analysis Using AI Techniques: State-of-the-Art, Taxonomies, and Challenges," in *IEEE Access*, vol. 8

Scikit-learn: (2020)Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

Will Koehrsen (2017), https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d

L. Jiang, H. Zhang and Z. Cai, (2009)"A Novel Bayes Model: Hidden Naive Bayes," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 10, pp. 1361-1371.

Bakshi, R. K., Kaur, N., Kaur, R., & Kaur, G. (2016). Opinion mining and sentiment analysis. In 2016 3rd *International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 452-455). IEEE