A Survey on Analysis of Drug Usage over Social Media Network Using Machine Learning Techniques

Mrs. M. Praveena[1], Ms. J. Yeswanthi[2], Ms. K. Susmitha[3], Ms. G. SriSowmya[4], Ms. VVL Anjali[5], Ms. Ch. Keerthi[6]

[1]Assistant Professor, Dept Of CSE, Qis Institute of Technology, Ongole, Prakasam (Dt)
[2]Student, Dept Of CSE, Qis Institute of Technology, Ongole, Prakasam (Dt)
[3]Student, Dept Of CSE, Qis Institute of Technology, Ongole, Prakasam (Dt)
[4]Student, Dept Of CSE, Qis Institute of Technology, Ongole, Prakasam (Dt)
[5]Student, Dept Of CSE, Qis Institute of Technology, Ongole, Prakasam (Dt)
[6]Student, Dept Of CSE, Qis Institute of Technology, Ongole, Prakasam (Dt)

Abstract

Recent growth in the data evolution Machine Learning plays a vital role in analyzing the data and extracting the results accordingly. This paper is based on the research carried out on bulk amount of labeled data from drug reviews posted on social media. Sentiment analysis is performed to determine the positive, negative and neutral data. Stop words and unwanted words are identified by NLP techniques. In order to perform machine learning on text, the data is transformed into vector representations such that we can apply numeric machine learning. Representing data numerically gives us the ability to perform meaningful analytics and also creates the instances on which machine learning algorithms operate. Hence feature extraction or more simply, vectorization is performed by computing term frequency and inverse document frequency. Different classification models are applied to the data and the model performance is evaluated on the basis of metrics such as accuracy, precision, recall, f1 score, support, confusion matrix. The best classifier is then used to test which type of review is the best to train model.

Keywords:

Machine Learning, accuracy, f1 score, precision, recall, support, confusion matrix.

I. INTRODUCTION

In the past decade, fast-growing social media websites have reached a critical mass of patient discussions about diseases and drugs primarily in the form of unstructured, casual human language. This data covers various medication outcomes such as effectiveness, adverse effects due to medication, adherence, and cost. Such information could be helpful for generating and verifying drug-repositioning hypotheses if these statements could be computationally developed using sentiment analysis and different classification models. Generally there are two main approaches for sentiment analysis: a machine learning approach (or a statistical text mining approach) and a linguistic approach (or a natural language processing approach). Since clauses are quite short and do not contain many subjective words, the machine learning approach generally suffer from data sparseness problem. Also the machine learning approach cannot handle complex

grammatical relations between words in a clause.

Classifiers

Supervised learning is built to make prediction, given an unforeseen input instance. A supervised learning algorithm takes a known set of input dataset and its known responses to the data (output) to learn the regression/classification model. A learning algorithm then trains a model to generate a prediction for the response to new data or the test dataset. Supervised learning uses classification algorithms and regression techniques to develop predictive models.

Naive Bayes's:
Naive Bayes is a type of Classification technique, which based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other function. Naive Bayes model is accessible to build and particularly useful for extensive datasets. Multinomial Naive Bayes classification algorithm tends to be a baseline solution for sentiment analysis task. The basic idea of Naive Bayes technique is to find the probabilities of classes assigned to texts by using the joint probabilities of words and classes.

Logistic Regression:
Logistic regression falls under the category of supervised learning; it measures the relationship between the dependent variable which is categorical with one or more than one independent variables by estimating probabilities using a logistic/sigmoid function. Logistic regression can generally use where the dependent variable is Binary or Dichotomous. It means that the dependent variable can take only two possible values like "Yes or No", "Living or Dead". Logistic Regression is a good baseline model for us to use for several reasons:

(1) They're easy to interpret,
(2) Linear models tend to perform well on sparse datasets like this one, and
(3) They learn very fast compared to other algorithms.

Linear SVM:
A Support Vector Machine is a type of Classifier, in which a discriminative classifier formally defined by a separating hyperplane. This classifier works trying to create a line that divides the dataset leaving the larger margin as possible between points called support vectors. The algorithm outputs an optimal hyperplane which categorizes new examples. In two dimensional space, this hyperplane is a line dividing a plane into two parts wherein each class lay on either side.

II RELATED WORK

Serendipity— A Machine-Learning Application for Mining Serendipitous Drug Usage From Social Media: Serendipitous drug usage refers to the unexpected relief of comorbid diseases or symptoms when taking medication for a different known indication. Historically, serendipity has contributed significantly to identifying many new drug indications. If patient-reported serendipitous drug usage in social media could be computationally identified, it could help generate and validate drug-repositioning hypotheses. We investigated deep neural network models for mining serendipitous drug usage from social media. We used the word2vec algorithm to

construct word-embedding features from drug reviews posted in a WebMD patient forum. We adapted and redesigned the convolutional neural network, long short-term memory network, and convolutional long short-term memory network by adding contextual information extracted from drug-review posts, information-filtering tools, medical ontology, and medical knowledge. We trained, tuned, and evaluated our models with a gold-standard dataset of 15714 sentences (447 [2.8%] describing serendipitous drug usage). Additionally, we compared our deep neural networks to support vector machine, random forest, and AdaBoost.M1 algorithms. Context information helped to reduce the false-positive rate of deep neural network models. If we used an extremely imbalanced dataset with limited instances of serendipitous drug usage, deep neural network models did not outperform other machine-learning models with n-gram and context features. However, deep neural network models could more effectively use word embedding in feature construction, an advantage that makes them worthy of further investigation. Finally, we implemented natural-language processing and machine-learning methods in a web-based application to help scientists and software developers mine social media for serendipitous drug usage.

Mining Social Media Data for Understanding Drugs Usage: This study carried out in the area of data mining depends for managing bulk amount of data with mining in social media on using composite applications for performing more sophisticated analysis using cloud platform. Enhancement of social media may address this need. The objective of this study is to introduce such type of tool which used in social network to characterised drug abuse. This paper outlined a structured approach to analyse social media in order to capture emerging trends in drug abuse by applying powerful methods like cloud computing and Map Reduce model. This study described how to fetch important data for analysis from social network as Twitter, Facebook, and Instagram. Then big data techniques to extract useful content for analysis are discussed.

Social networks mining for analysis and modeling drugs usage: This study presents approach for mining and analysis of data from social media which is based on using Map Reduce model for processing big amounts of data and on using composite applications for performing more sophisticated analysis which are executed on environment for distributed computing based cloud platform. We applied this system for creation characteristics of users who write about drugs and to estimate factors that can be used as part of model for prediction drug usage level in real world. We propose to use social media as an additional data source which complement official data sources for analysis and modeling illegal activities in society.

A Survey on Mining Social Media Data for Understanding Drug Usage: The study on this review paper presents the complicated usage of prescribed drugs which perform under the area of data mining for managing high volume of data and usage of complex

function for performing more refined analysis using cloud platform. The aim of this study is to understand the innovative and extensive frame that characterize drug abuse using social media. The concept of this study paper is an analytical approach to analyze social media by applying powerful techniques such as cloud computing and Map Reduce model. for acquiring the drug abuse emerge trends. This paper describes how to capture important data to evaluate from networks like Twitter, Facebook, and Instagram. Also, big data techniques are used for analysis of data content.

Data, tools and resources for mining social media drug chatter: Social media has emerged into a crucial resource for obtaining population-based signals for various public health monitoring and surveillance tasks, such as pharma co vigilance. There is an abundance of knowledge hidden within social media data, and the volume is growing. Drug-related chatter on social media can include user-generated information that can provide insights into public health problems such as abuse, adverse reactions, long-term effects, and multi-drug interactions. Our objective in this paper is to present to the biomedical natural language processing, data science, and public health communities data sets (annotated and unannotated), tools and resources that we have collected and created from social media. The data we present was collected from Twitter using the generic and brand names of drugs as keywords, along with their common misspellings. Following the collection of the data, annotation guidelines were created over several iterations, which detail important aspects of social media data annotation and can be used by future researchers for developing similar data sets. The annotation guidelines were followed to prepare data sets for text classification, information extraction and normalization. In this paper, we discuss the preparation of these guidelines, outline the data sets prepared, and present an overview of our state-of-the-art systems for data collection, supervised classification, and information extraction. In addition to the development of supervised systems for classification and extraction, we developed and released unlabeled data and language models. We discuss the potential uses of these language models in data mining and the large volumes of unlabeled data from which they were generated. We believe that the summaries and repositories we present here of our data, annotation guidelines, models, and tools will be beneficial to the research community as a single- point entry for all these resources, and will promote further research in this area.

III METHODOLOGY

The proposed paper follows the approach for analysis as shown in Figure 1. Implementation of the analysis was done using python software.

Reviews and tweets collection:

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

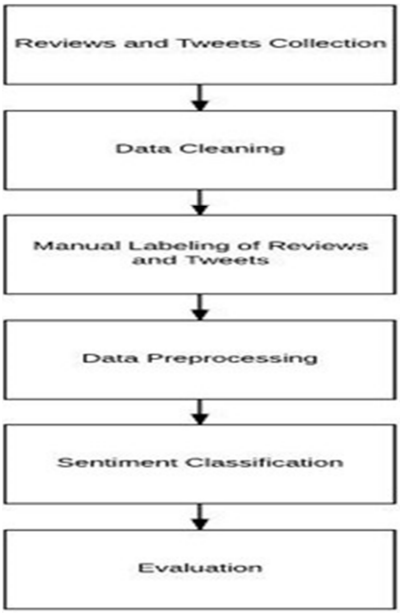


Fig: Implementation

Data cleaning:
Data is prepared for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. About 3500 reviews were chosen from the medications. To simplify the process of manually labeling the reviews, lengthy reviews were removed. Also, reviews containing irrelevant information such as questions and suggestions about the drug, blank reviews and repeated reviews were removed.

Data labeling: Labeled data is a group of samples that have been tagged with one or more labels. Labeling typically takes a set of unlabeled data and embedding each piece of that unlabeled data with meaningful tags that are informative. A primary step in enhancing any model is to set a training algorithm and validate these models using high-quality training data. Data labeling is important because the machine learning algorithm have to understand the data. Based on this only we will be able to train the model.

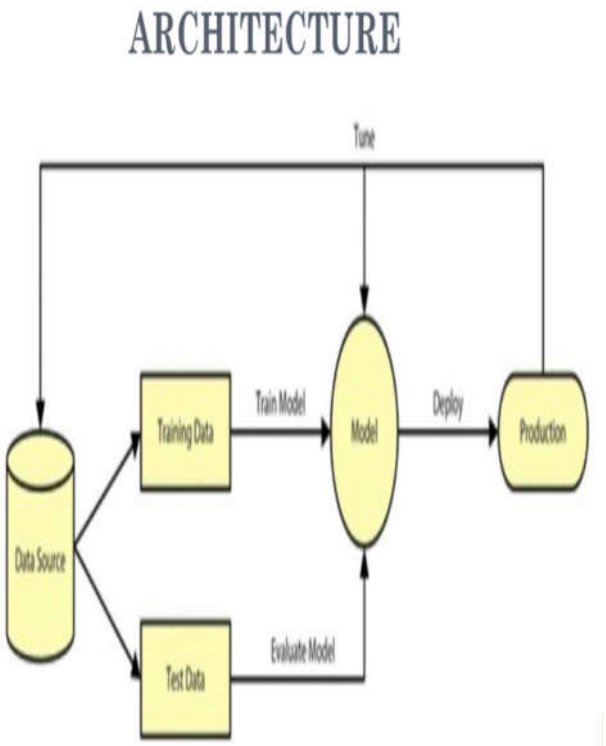
Data pre-processing: For the purpose of training the machine learning model, The data had to undergo several preprocessing steps. They were:

- **Nominal to Text:** In order to perform text processing, the data was converted from nominal to text values.
- **Transforming cases:** The entire reviews were converted to lower case to avoid am- biguity.
- **Tokenization:** The text in the reviews were broken down into constituent words with each word being termed as a 'token'. Tokenization removes unwanted symbols and punctuation marks which have no meaning. Also, tokenizing is done to construct a Term frequency - Inverse document frequency (TF-IDF) dictionary for constructing n grams. Term frequency is the number of a times a word appears in a document and inverse document frequency is how much information that a word provides and is generally calculated as logarithmic quotient of total number of documents by the number of documents containing the word.

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

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

- **Stemming:** The remaining words were converted to their root word or stemmed. For example, running will be converted into run, eating will be converted into eat. The purpose of stemming is to group similar words together by converting them into their root word. Stemming process reduces the size of vocabulary and hence decreases the redundancy of word occurrence.



Sentiment Classification: The text is analyzed and the underlying sentiment is classified as positive or negative or neutral. This is a common NLP task, which involves classifying texts or parts of texts into a pre-defined sentiment. Sentiment analysis can be used to categorize text into a variety of sentiments.

Evaluation: The performance of a machine learning model is evaluated to estimate the generalization accuracy of a model on future data. Model evaluation metrics are required to quantify model performance. The choice of evaluation metrics depends on a given machine learning task.

IV IMPLEMENTATION

Model accuracy and confusion matrix: Model accuracy in terms of classification models can be defined as the ratio of correctly classified samples to the total number of samples. As a performance measure, accuracy is inappropriate for imbalanced classification problems. For imbalanced classification problems, the majority class is typically referred to as the negative outcome and the minority class is typically referred to as the positive outcome. The confusion matrix provides more insight into not only the performance of a predictive model, but also which classes are being predicted correctly, which incorrectly, and what type of errors are being made. The simplest confusion matrix is for a two-class classification problem, with negative and positive classes. In this type of confusion matrix, each cell in the table has a specific and well- understood name

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Or for binary classification models, the accuracy can be defined as:

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

Predicted

Actual		Positive	Negative
	Positive	TP	FN
	Negative	FP	TN

Confusion Matrix

- **True Positive (TP) :** A true positive is an outcome where the model correctly predicts the positive class.
- **True Negative (TN) :** A true negative is an outcome where the model correctly predicts the negative class.
- **False Positive (FP) :** A false positive is an outcome where the model incorrectly predicts the positive class.
- **False Negative (FN**) : A false negative is an outcome where the model incorrectly predicts the negative class.
- **Precision and recall:** Precision is a metric that quantifies the number of correct positive predictions made therefore, calculates the accuracy for the minority class.It is calculated as the ratio of correctly predicted positive examples divided by the total number of positive examples that were predicted.

  Precision=TruePositives/(TruePositives+ FalsePositives)

  Recall is a metric that quantifies the number of correct positive predictions made out of all positive predictions that could have been made. Unlike precision that only comments on the correct positive predictions out of all positive predictions, recall provides an indication of missed positive predictions. Recall is calculated as the number of true positives divided by the total number of true positives and false negatives.

Recall=TruePositives/(TruePositives+ FalseNegatives)

F1 score: F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Support: The support is the number of samples of the true response that lie in that class. Support is the number of actual occurrences of the class in the specified dataset. Imbalanced support in the training data may indicate structural weaknesses in the reported scores of the classifier and could indicate the need for stratified sampling or rebalancing. Support doesn't change between models but instead diagnoses the evaluation process.

V CONCLUSION

In this paper, we studied the application of machine learning based sentiment analysis of patient generated drug reviews. We evaluated classification algorithms like naive Bayes's model, linear SVM model and logistic regression among which logistic regression showed better performance. Hence Logistic regression models were trained using simple lexical features such as unigrams, bigrams and trigrams extracted from the reviews. Depending on data source, promising classification results could be obtained. As labeled data sets for building classification models are rare or are only available in unstructured fashion, we investigated various approaches for model portability. Whereas in-domain (i.e. condition) training and evaluation shows very good classification results, the performance of models trained on one specific condition and tested on another condition, varies among domains. Cross-data evaluation, i.e. training and testing classifiers on data from different sources, was only

unsatisfactorily possible with the applied classifier and features. Therefore, we believe that employing more sophisticated features and applying more powerful machine learning models, e.g. deep learning approaches can improve the achieved results.

VI REFERENCES

1. Ali,F.,Kwak,D.,Khan,P., Islam, S.R., Kim,K.H. and Kwak, K. (2017). Fuzzy ontology- based sentiment analysis of transportation and city feature reviews for safe traveling, Transportation Research: Part C 77: 33 –48.
2. URL:http://ezproxy.ncirl.ie/login?url=http://search.ebscohost.com/login.aspx?direct=trueAuthTyplivescope=site.
3. Chan, S. W. and Chong, M. W. (2017). Sentiment analysis in financial texts. Decision Support Systems 94: 53 –64.
4. Compton,W.M.andVolkow,N.D.(2006). Abuseofprescriptiondrugsandtheriskof addiction,DrugandAlcoholDependence83(Supplement1):S4–7.DrugFormulation and Abuse Liability.
5. J. T. Dudley, T. Deshpande, and A. J. Butte, "Exploiting drug–disease relationships for computational drug repositioning," Briefings in Bioinformatics, vol. 12, pp. 303-311, 2011.
6. T. T. Ashburn and K. B. Thor, "Drug repositioning: identifying and developing new uses for existing drugs," Nature Review Drug Discovery, vol. 3, pp. 673-683, 2004.
7. L. Yao, Y. Zhang, Y. Li, P. Sanseau, and P. Agarwal, "Electronic health records: Implications for drug discovery," Drug Discovery Today, vol.16, pp. 594-599, 2011.
8. C. Andronis, A. Sharma, V. Virvilis, S. Deftereos, and A. Persidis, "Literature mining, ontologies and information visualization for drug repurposing," Briefings in Bioinformatics, vol. 12, pp. 357-368, 2011.
9. M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. Abbas, S. J. Hufeisen, et al., "Predicting new molecular targets for known drugs," Nature, vol.462, pp. 175-181, 2009.
10. P. Sanseau, P. Agarwal, M. R. Barnes, T. Pastinen, J. B. Richards, L. R. Cardon, et al., "Use of genome-wide association studies for drug repositioning," Nature Biotechnology, vol. 30, pp. 317-320, 2012.
11. A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan, "PREDICT: a method for inferring novel drug indications with application to personalized medicine," Molecular Systems Biology, vol. 7, p. 496, 2011.
12. J. D. Wren, R. Bekeredjian, J. A. Stewart, R. V. Shohet, and H. R. Garner, "Knowledge discovery by automated identification and ranking of implicit relationships," Bioinformatics, vol. 20, pp. 389-398, 2004.

Authors Profile

Mrs. M Praveena, M.Tech., is working as an Assistant Professor of CSE Department in **Qis Institute of Technology, Ongole, Prakasam (Dt).** She has 16 years Teaching Experience and her areas of interest are machine learning and Image Segmentation.

Ms. J Yeswanthi pursuing B Tech in computer science engineering from **Qis Institute of Technology, Ongole, Prakasam (Dt)**

Ms. K Susmitha pursuing B Tech in computer science engineering from **Qis Institute of Technology, Ongole, Prakasam (Dt)**

Ms. G SriSowmya pursuing B Tech in computer science engineering from **Qis Institute of Technology, Ongole, Prakasam (Dt)**

Ms VVL Anjali pursuing B Tech in computer science engineering from **Qis Institute of Technology, Ongole, Prakasam (Dt)**

Ms. Ch Keerthi pursuing B Tech in computer science engineering from **Qis Institute of Technology, Ongole, Prakasam (Dt)**