MINING SERENDIPITOUS DRUGS FROM REVIEWS USING MACHINE LEARNING TECHNIQUES

B. Prathyusha
Asst.Professor
Department of Information Technology
Malla Reddy Engineering College for Women
(UGC-Autonomous)
Maisammaguda,Hyd-500100,Telangana,India.

Rupali Sharma
Student
Department of Information Technology
Malla Reddy Engineering College for Women
(UGC-Autonomous)
Maisammaguda,Hyd-500100,Telangana,India.

Y. Yamini
Student
Department of Information Technology
Malla Reddy Engineering College for Women
(UGC-Autonomous)
Maisammaguda,Hyd-500100,Telangana,India.

R. Amitha Bhavana Reddy
Student
Department of Information Technology
Malla Reddy Engineering College for Women
(UGC-Autonomous)
Maisammaguda,Hyd-500100,Telangana,India.

Abstract— The word serendipity means 'happy accident'. Making discoveries by accident has contributed a lot to medical history. Serendipitous drug use is when a patient takes a prescription for a separate known indication and unintentionally experiences relief from comorbid illnesses or symptoms. The discovery of numerous new medication indications has benefited greatly from serendipity throughout history. Drug-repositioning hypotheses might be created and validated if patient-reported serendipitous drug usage in social media could be computationally discovered. We looked into deep neural network models for social media mining of accidental drug use.

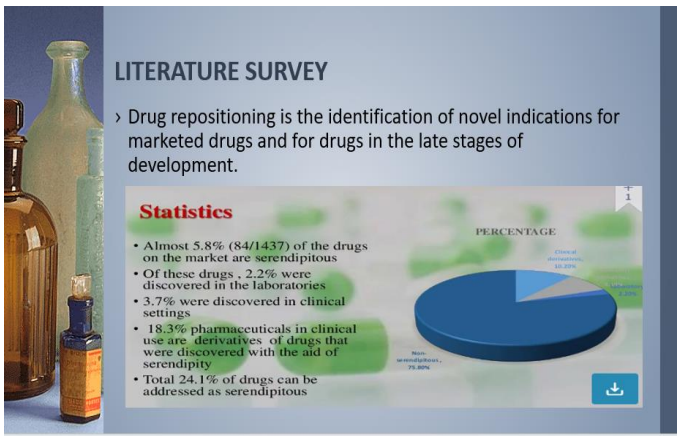
We contrasted Regression and KNN with our support vector machine, random forest, RNN, and LSTM algorithms. We used machine learning and natural language processing techniques in a web application to mine social media and data reviews for accidental drug use. An essential algorithm is sentiment analysis. We decided to employ Natural Language Processing for our project since it can be used to identify sentiment in text. Upon reviewing reviews of various pharmaceuticals that have been rated on a scale of 1 to 10 and have been reviewed as texts. This data set was collected from the UCI machine learning repository, which contained the train and test data sets (divided as 75–25%). In general, we categorize the drug's numerical rating into three categories: positive (7–10), negative (1-4), or neutral (4-7). We chose to look into how the ratings of the drugs are affected by the inclusion of different words in reviews for ailments with many reviews for drugs that are used to treat those conditions. Our main goal was to construct supervised machine learning classification algorithms that use textual reviews to predict the rating class. Last but not least, we used machine learning and natural language processing techniques to mine data reviews for drug usage.

I. INTRODUCTION

Finding new indications for medications that are already on the market or that are still in the early stages of development is known as drug repositioning. Repositioned medications can then be more easily accessible to patients with diseases that are not being appropriately treated and more cost-effective for pharmaceutical corporations because some of the time and money associated with preclinical research can be saved. These advantages are of interest to biomedical researchers, who have examined various computational methods to generate and verify drug repositioning hypotheses by assessing chemical and biological data, literature, and electronic health records. 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. These data cover various medication outcomes such as effectiveness, adverse effects due to medication, adherence, and cost. Recent research has examined this new data source primarily for pharmacovigilance purposes. In social media posts, some patients have also mentioned that comorbid diseases or symptoms unexpectedly improved while they were taking a certain drug for a common or known indication. We refer to these events as serendipitous drug usage. An example of serendipitous drug usage: a patient reported that her symptoms of irritable bowel syndrome were alleviated when taking sulfasalazine, which was prescribed for rheumatoid arthritis. Such information could help generate and verify drug-repositioning hypotheses if these statements could be computationally detected.

II. RELATED WORK

Numerous investigations and tests using a variety of sample data were carried out to determine how sentiment analysis could be used to identify fraudulent reviews. The websites of items and studies are scraped for numerous product reviews. In our work, we have chosen to make use of the misleading opinion spam dataset. We have taken the information out of the dataset and put it in a list. After that, we developed the Frame using the labels that went with each review's sentiment analysis. It is decided if the polarity is positive or negative. We have also designated if the model is True or Deceptive. Later, the class converter into 0 and 1s for polarity can be used. Naive Bayes classification, Vector Machine, and Decision Tree were our primary methods of choice. An algorithm for binary (two-class) and multiple-class classification issues is called Naive Bayes. There has been a lot of background information about applying machine learning, particularly deep learning techniques, for sentiment analysis. Emojis have been studied and trained to be utilized in tweet sentiment classification (1). SVM, Naive Bayes, RNN, and ANNs were used to train these models. The Glove representation was utilized to turn the emojis into a score, which was then used for sentiment analysis.



III. EXISTING SYSTEM

Many Serendipitous drug usage in social media can be valuable information for drug discovery and development, but they need to be manually verified to exclude false-positive cases, i.e. when patients inaccurately describe their medication outcomes. Our approach to sentiment analysis was the contextual meaning of certain keywords and how important are certain words in sentiment detection. To check the effectiveness of using Sentiment Analysis which could detect the sentiment of the review and hence be in agreement with the rating classification. We explored machine-learning methods to identify serendipitous drug usage from patient health forums (social media). We collected drug-review posts from WebMD and designed information filters to eliminate noise in the data. We constructed machine-learning features from n-grams, outputs from information-filtering tools, medical knowledge, and other information from the drug-review posts. We used machine-learning algorithms, namely. These studies typically used word embedding trained by unsupervised learning algorithms such as word2vec to construct features from texts. These features were then classified by using convolutional filters or recurrently connected neurons.

DISADVANTAGES:

1. The algorithms used in the system have shown less accuracy and precision values.
2. Although convolution filters are good at processing data in the matrix or grid representation, they capture only sequential patterns in a local area and sometimes miss long-range dependencies between words in the same sentence.
3. We cannot differentiate between fake and genuine reviews.

IV. PROPOSED SYSTEM

Natural language processing (NLP) and machine-learning methods were explored to identify serendipitous drug usage from patient health forums (social media). We collected drug reviews and designed information filters to eliminate noise in the data. This web application takes only drug-review comments and the drug name as inputs, making it adaptive to broader sources of patient-generated health data. The NLP and machine-learning methods are integrated into an automated workflow. We had chosen to use a data set that is more on par with the medical healthcare industry and we wanted to investigate the use of NLP algorithms (in particular Sentiment Analysis). We developed Serendipity, an easy-to-use, web-based software application, to effectively extract serendipitous drug usage from reviews given by users which provides an analysis of drug usage.

PHASE 1: DATA EXPLORATION

The train test split for the train and test files was (75-25%) samples. There were roughly N samples total in the data set. Each sample has the following fields: the drug name, the condition for which it is used, a user's review of the medicine, the user's rating of the drug, the date the drug was reviewed, and a useful count, which shows how many users considered the sample beneficial. Pre-processing of data: Reviews, polarity class, and class are the three columns that make up the data frame that we have produced. The column named reviews gives the text or reviews customers whereas our class shows whether it is deceptive or True. And the polarity class shows whether the polarity is positive or negative.

PHASE 2: Text to Numeric Data Representation

To encode the review texts into numeric data, we used certain pre-training algorithms such as Term Frequency Inverse Document Frequency and also Count Vectorizer. We used TF-IDF embedding to calculate the matrix of numeric values for each word t within each review text. If, The term frequency tf(t, d) calculates the proportion of times that the term t ∈ V (d) appears in document d. The vocabulary V (d) = P t n(t, d) is constructed by document d. Thus, if a word w 0 does not appear in a document d 0, the term frequency of(t 0, d0) in this case would be zero. The idea of the term frequency is essentially the same as Count Vectorizer.

PHASE 3: Training Models

We decided to try the following algorithms to investigate the accuracy of sentiment detection using the above numeric representation techniques. We used algorithms such as Neural Networks: Recurrent Neural Networks with Long Short Term Memory (LSTM) and also other state-of-the-art machine learning classification algorithms such as Support Vector Machines (SVM), and also Random Forests (RF). Our next aim was to identify which type of machine-learning algorithm would yield the best results.

ADVANTAGES:

- The system has more speed for drug health informatics detecting in a fast way.
- It calculates an exact search result for drugs in social media.
- Better accuracy and precision values that are reliable.
- Identification of fake reviews is easy.

MODULES:

SERVICE PROVIDER: In this module, the Service Provider has to log in using a valid username and password. After the login is successful, he can do some operations such as Login, Browse and Train & Test Data Sets, and View Trained and Tested Accuracy in the form of a Bar Chart, Pie chart, and line graph. View Trained and Tested Accuracy Results, View Predicted Tweet Account Type Details, Find Tweet Account Type Ratio, Download Predicted Data Sets, View Tweet Account Type Ratio Results, and View All Remote Users.

VIEW AND AUTHORIZE USERS: The admin can view the registered user's list in this module. In this, the admin can view the user's details such as user name, email, address, and reviews, and authorize the users.

REMOTE USER: In this module, there are n numbers of users are present. Users should register before doing any operations. Once user registers, their details will be stored in the database. After registration is successful, he has to login in using an authorized username and password. Once Login is a successful user will do some operations like Register And log in, Add drug details, and give reviews.

RESULTS

From our results, we are deducing the following conclusions.

• In general, we noted that neural networks in particular did better prediction on test data sets than the other machine learning algorithms.

• Deep learning algorithms capture more significant features for classifying to predict the sentiment within the review.

• SVM and Logistic Regression also have a pattern in common. They perform similarly in every model with SVMs having an edge over Logistic Regression.

• The reason why we find it is because SVM can provide a better algorithm for classifying since it does margin classification while Logistic Regression classifies based on probability of likelihood which does not provide to be better.

ALGORITHM	ACCURACY	PRECISION
SVM	92.4%	91%
Random Forest	64%	66%
RNN with LSTM	94%	92%

Fig.1: Result Values

CONCLUSION

We created a web-based solution for those who are interested in mining reviews for random drug use. It offers a GUI that allows users to input data and see the results of analytics. We are drawing the following inferences from our findings. In general, we found that compared to other machine learning techniques, neural networks performed particularly well in terms of prediction on test data sets. Deep learning algorithms gather more important features for classification and sentiment prediction in review evaluations. The use of SVM and Logistic Regression share a similar trend. They all perform similarly, with SVMs outperforming Logistic Regression. As opposed to Logistic Regression, which classifies based on probability of likelihood and does not offer a better classifier based on the relevant features (words present) utilized in TFIDF and CV, SVM can provide a better technique for classification. In addition to the aforementioned trends, we also found that random forest algorithm models generally outperformed the other algorithms. We may infer this performance because, unlike neural networks or SVMs, the random forest is a decision tree classifier that likely does not incorporate all the essential information utilized for classification.

FUTURE WORK

These initiatives just represent the first stage of software development. There are numerous methods to enhance the present implementation. The drug information database can first be expanded by including more medications, or by integrating with additional databases. Second, the GUI can offer extra options to modify the behavior of the application, such as modifying the sentiment and semantic difference filters' thresholds or including or excluding particular prediction models. Finally, we identified 2 broad user categories without additional verification. The following phase is to more precisely define users, produce business or real-world use cases for user research, and potentially conduct customer interviews. Furthermore, we can use tools like Python, R or R studio, Statistical Software, and sentiment classification algorithms to identify fraudulent reviews.

REFERENCES

[1] 1. [1] 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.

[2] 2. [2] 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

[3] 3. [3] 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.

[4] 4. [4] 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.

[5] 5. Dane Hankamer, David Liedtka. Twitter Sentiment Analysis with Emojis. 2019

[6] 6. .Aliaksei Severyn, Alessandro Moschitti.Twitter Sentiment Analysis with Deep Convolutional Neural Networks. 2015

[7] 7. .Janata Wehrmann et al. A Multi-Task Neural Network for Multilingual Sentiment Classification and Language Detection on Twitter. 2018

[8] 8. Simon Provoost et al. Validating Automated Sentiment Analysis of Online Cognitive Behavioral Therapy Patient Texts- An Exploratory Study. 2019

[9] Rakibul Hassan and Md. Rabiul Islam “Detection of fake online reviews using semi-supervised and supervised learning” 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)