PROJECT PROPOSAL

Sentiment Analysis on Drug Reviews

Members : Alekhya Majeti **|** Somesh Kale **|** Heet Detroja **|** Rohan Bhosale

Description:

Drug reviews provides valuable information like the side effects of the drugs, how that particular drug helped someone cure the disease or whether that drug was helpful or not, etc. Based on the past experience or past reviews given by the people who took the drug, we can help classify which reviews are positive and negative using sentimental analysis techniques on the Drug review Dataset [1]. This Dataset includes drug names, conditions for which the drug is used, patient reviews for that particular drug, rating of the drug, the date of the review and also the number of User counts who found that drug useful. Before taking any drug, this might help a person to understand whether that particular drug has a positive or negative side effects and can relate the same situation with the past experience of that person. We will be using different libraries and approaches and will be comparing this approaches which might help us gain information about how that particular library is stemming or lemmatizing a particular word to reduce its inflectional forms into a common base form like ‘colors’ into ‘color’, ‘different’ into ‘differ’ and also whether that particular word has a positive, negative or neutral meaning. Also, using different classification models in our approach will help us understand which model is performing better.

Dataset :

The dataset that we are going to use for this project is the Drug review dataset [1]. We’ve taken it from the UCI repository. This dataset has over 200,000 instances with 6 attributes. This dataset gives info from the patient reviews on some specific drugs with the related conditions as well. It also gives a star rating from patients out of 10. The data is mainly from the online pharmaceutical review sites. The data is split into two categories train and test sets. Train has 75% of the total data and remaining is the test data.
These websites give information about various kinds of drugs, their side effects, dosage, reviews, drug class.

The sample drugs that are in this dataset are Debrox, Adderall, Aspirin, Levofloxacin and so on.

Approach Overview:

First, we started exploring all the attributes, beginning from unique ID. We compared the unique ID with the length of the train data to check whether one patient has written multiple reviews.

So, the attributes that are labelled in this dataset are as follows:

- Drug name
    - Name of the drug.
- Condition
    - Name of the condition.
- Review
    - Patient review.
- Date
    - Date of the review entry.
- Usefulcount
    - Number of users who found the review useful.
- Rating
    - Patient rating on a scale from 1 to 10.
- Patient ID
    - Patient with a unique ID.

So to begin with the project we have to prepare the data that we are planning to feed the model. So, for this we have to preprocess the dataset. In this phase, we will be cleaning the data and check for some empty or missing datasets. After that, we have to do some text splitting, cleaning the text by removing the stop words, tokenization, POS tagging. After these steps, we are splitting the data into train and test datasets for our model.

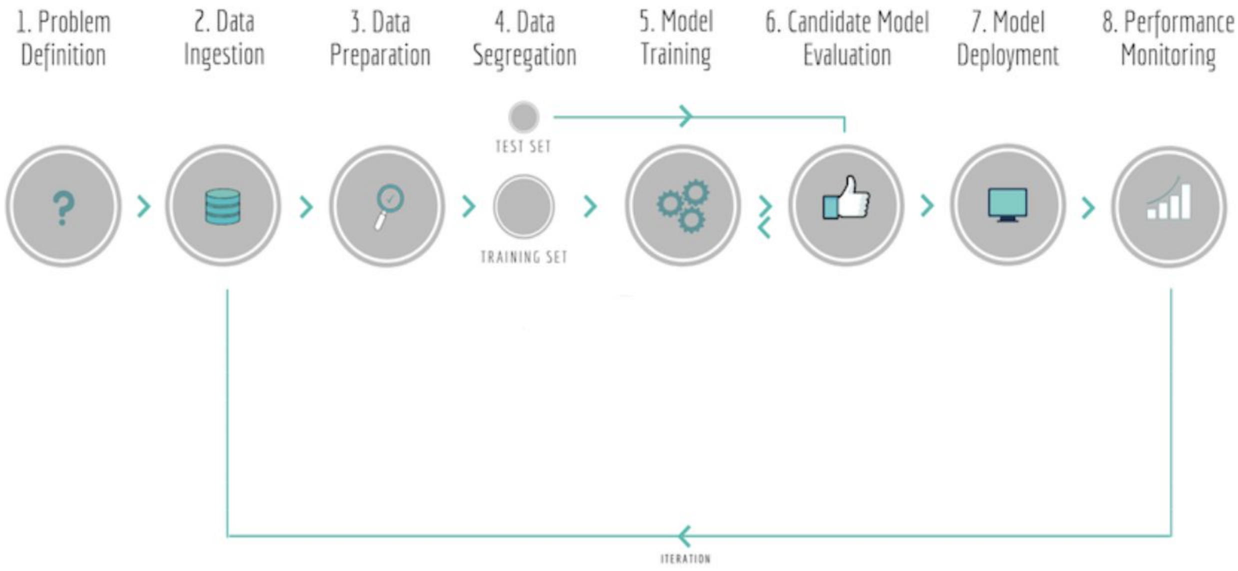
After this phase, we will be choosing different models for the project to train and test the data. As we are a team of four we are planning to split the work by each team member taking couple of models and training the dataset. Finally, we will be bringing the work together.

In the phase of training the model, we are going to take the training data to train the model. We are planning to use the training data incrementally enhance each model’s ability to anticipate the result.

The next step is to ensure the training is complete and check whether our model is good or not. This phase includes the testing dataset. After using the test dataset to our trained model, we are planning to evaluate the results. This will give us an estimate of how our model is going to perform in the real world.

After this, we will visualise the result and try to improve the result by further tuning the dataset.

The image below shows the architecture of our project.



Literature Survey:

Sentiment Analysis is used to identify the phrase in a text that contains some sentiment, it is an information extraction task. Here, drug reviews are used as text on which the analysis is done, it may contain many entities but it is necessary to find the entity which is directed towards the sentiment. Sentiment are classified as positive or negative. There are few challenges mentioned in [2]:

- **Identifying Sarcasm**

  A sentence may have implicit sentiment even without the presence of any sentiment bearing words.

  For example : How can anyone eat this food ?

  The example does not contain any negative sentiment bearing word although it is a negative sentence.

- **Domain Dependency**

  The polarity of words may change depending on the domain.

  For example: The movie was extremely scary.

  The jungle is pretty scary after dark.

  Here the first sentence is positive for a horror movie but it uses a negative sentiment, the second sentence uses the same negative sentiment in a negative way.

- **Thwarted Expectations**

  Some texts may contain positive words in the start but would refute it in the end with negative words.

  For example:

  > The dish was presented in an aesthetically pleasing way, it was mouth watering. The food had a savoury aroma to it. However, the taste did not hold up.

  The overall sentiment is negative due to the crucial last sentence, but it gives a positive orientation due to the presence of positive words.

- **Pragmatics**

  Pragmatic of user opinion must be detected for the statement as it can change the sentiment thoroughly.

  For Example : India CRUSHED Australia in the last test match.

  The car was crushed under the tree which fell due to heavy rain.

  Capital words can be used to denote sentiments.The first example gives positive sentiment while the second example gives a negative sentiment.

- **World Knowledge**

  World Knowledge needs to be incorporated in the system for detecting sentiments.

  For example: Inception and lunch. What an afternoon.

  One has to know about Inception here to find out the sentiment.

- **Subjectivity Detection**

  Subjectivity is to differentiate between opinionated and non-opinionated text.

  For Example: I hate love stories.

  I do not like the movie “I hate love stories”.

  The first example gives an objective fact whereas the second example shows the opinion about a movie.

- **Entity Identification**

  A sentence may contain multiple entities. It is important to find the entity towards which the opinion is directed.

  For Example: AT & T is better than Vodafone.

  The example is positive for AT & T but negative for Vodafone.

Sentiment Analysis can be used for a wide range of reviews and opinion forums regarding user experiences and preferences over multiple product domains. This information can behoove the industry by obtaining valuable insights on their products. Online review sites contain information related to multiple aspects of a product such as effectiveness of drugs and side effects which makes automatic analysis interesting but also challenging. Here analyzing the sentiments of various drug reviews can help with decision making and improve monitoring public health by revealing collective experience.

Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning is a paper on analysis of drug reviews by applying in-domain analysis where one model was used on two datasets to classify overall patient satisfaction reviews. For cross-domain analysis, performance of models built on one condition, i.e. the source domain was studied and evaluation of the model on data related to other condition, i.e. the target domain was done. These domain models were then evaluated on other condition related subsets. Finally, cross-data analysis was done by studying the transferability of the trained models among different data sources, overall patient satisfaction models were trained on both associated training dataset and evaluated on drug reviews from other independent data source test set. The evaluation of these different domains and models were displayed to show their performance [3].

Sentiment Analysis of Tweets using Machine Learning Approach is another paper that shows the use of machine learning for sentiment analysis by using a hybrid classifier of KNN and SVM to process the tweet features. After preprocessing the data a list of adjectives was used for feature generation which had positive and negative score plus the overall rating of an adjective. Feature generation classification is done then the hybrid approach in which prediction probability of both the classifier is used on the data. The evaluation of the hybrid classifier is then shown compared to individual KNN and SVM classifiers[8].

Sentiment Analysis on Twitter by using Machine Learning Technique is a paper where tweets are categorized in two parts opinions and opinionless, and then each tweet is compared to a database of positive, negative and average words, tweets containing these words were listed in opinion category rest were discarded.Finally estimating the probability of positive tweets naive bayes, SVM and maximum entropy were used and compared for better probability result.This paper shows using machine learning strategies are less difficult and more efficient[5].

An Efficient Sentiment Analysis and Summarization Using Unsupervised and Supervised Method for Amazon Review Dataset is a paper which uses text summarization where most striking highlights of a content are extricated and arranged into a short abstract for one record. Here both unsupervised and supervised methods are used to find positive,negative and neutral sentiments. The supervised learning used SVM algorithm on the amazon review dataset while for unsupervised learning used ANN algorithm on the same dataset. Both supervised and unsupervised algorithms performances were compared which gave SVM 96% accuracy while the ANN algorithm gave 100% accuracy. [7]

Project Timeline:

Task ID	Task	Start Date	End Date	Duration (Days)
1	Project dataset selection	Sept 5	Sept 15	11
2	Analysis of Data	Sept 15	Sept 22	8
3	Planning	Sept 22	Sept 28	7
4	Project Proposal	Sept 28	Oct 3	6
5	Data Preprocessing	Oct 3	Oct 28	26
6	Libraries	Oct 28	Nov 1	5
7	Mid Progress Report	Nov 1	Nov 7	7
8	Model Training	Nov 7	Nov 18	12
9	Model Evaluation and Training	Nov 18	Nov 28	11
10	Final Report and Poster	Nov 28	Dec 5	8

Team Members Roles

Task	Somesh Kale	Alekhya Majeti	Heet Detroja	Rohan Bhosale
Dataset Selection	✓	✓	✓	✓
Analysis of Data	✓	✓	✓	✓
Planning	✓	✓	✓	✓
Project Proposal	✓	✓	✓	✓
Data Preprocessing			✓	✓
Libraries	✓	✓	✓	✓
Model Selection	✓	✓	✓	✓
Mid-progress Report	✓	✓	✓	✓
Model Training	✓	✓		
Model Evaluation and Tuning	✓	✓	✓	✓
Final Report	✓	✓	✓	✓

We will be working together on almost every step like Selecting Dataset, Analysis of Data, Planning of implementation, usage of libraries etc, Model Selection where will be implementing different classification models like decision trees, random forest and naive bayes, Model Evaluation and Final Report. In Data pre-processing, steps like replacing null values, normalization if required, dropping irrelevant columns or features, Splitting the dataset into training and testing sections and also cleaning of data like Tokenization and cleaning a particular word into its noun or common form will be done. We will be implementing different libraries and will be comparing these implementations for our final approach.

Questions This Project Will Answer

- Which machine learning algorithm gives us the best results for our project?
- What features are best suited for drug review classification?
- What type of keywords are more useful for Drug review analysis?

Things That We Expect to Learn

We will learn how to build and deploy a machine learning model in real life scenario. This includes the following:

1. Learning about data cleaning process and sanitizing the text into its normal form.

2. The ability to build a well-organized training and testing dataset based on the selected features and also eliminating the irrelevant features in our dataset which will not affect our model performance.

3. The ability to use different machine learning models for sentiment analysis and comparison of these models for evaluating the performance of each model.

4. This project will help us learn about different libraries that we will be trying and will also help us learn the ability to tune a model .

5. The ability and an opportunity to work on a real world problem.

Is Our Idea Novel?

Different approaches have been proposed and implemented for sentiment analysis not in particular for medicine, but for multiple applications like Social media comments, movie reviews etc. This project will present a good analysis to doctors, and patients whether the particular drug is having a positive or a negative effect based on the reviews given by the people who used the drug.

References:

[1] UCI Machine Learning Repository: Drug Review Dataset (Drugs.com) Data Set, https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29

[2] Sentiment Analysis - A Review. (2015). International Journal of Science and Research (IJSR), 4(12), 1842–1845. doi: 10.21275/v4i12.nov152437

[3] Gräßer, F., Kallumadi, S., Malberg, H., & Zaunseder, S. (2018). Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. Proceedings of the 2018 International Conference on Digital Health - DH 18. doi: 10.1145/3194658.3194677

[4] Gautam, Geetika, and Divakar Yadav. "Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis." 2014 Seventh International Conference on ContemporaryComputing (IC3), 2014, doi:10.1109/ic3.2014.6897213.

[5] Upadhyay, N., & Singh, P. A. (2016, May). "Sentiment Analysis on Twitter by using Machine Learning Technique." International Journal for Research in Applied Science & Engineering Technology (IJRASET).

[6] Sood, Akriti, et al. "An Initiative to Identify Depression Using Sentiment Analysis: A Machine Learning Approach." Indian Journal of Science and Technology, http://www.indjst.org/index.php/indjst/article/view/119594.

[7] R, Vanitha. "An Efficient Sentiment Analysis and Summarization Using Unsupervised and Supervised Method for Amazon Review Dataset." International Journal of Emerging Technologies and Innovative Research JETIR, JETIR, http://www.jetir.org/view?paper=JETIRBI06016.

[8] Gupta, Ankita, et al. "Sentiment Analysis of Tweets Using Machine Learning Approach." International Journal of Computer Science and Mobile Computing, 4 Apr. 2017.