PROJECT ON SENTIMENT ANALYSIS OF TWEETS

TEAM MEMBERS:

Tushar Nirmal

Sonakshi Khosla

Preksha Dhoot

SthitaPragyan Patnaik

CONTENT:

I. INTRODUCTION:

Data analysis is the process of applying organised and systematic statistical techniques to describe,recap,check and condense data. It is a multistep process that involves collecting, cleaning, organizing and analyzing. Data mining is like applying techniques to mold data to suit our requirement. Data mining is needed because different sources like social media, transactions, public data, enterprise data etc. generate data of increasing volume, and it is important to handle and analyze such big data.

Sentiment analysis, also referred to as opinion mining, is a sub machine learning task where we determine the sentiment of a given document.The objective of the proposed analysis, 'Sentiment Analysis', is the analysis of the enormous amount of data easily available from social media. Using machine learning techniques and natural language processing we can extract the subjective information of a document and try to classify it according to its polarity such as positive, neutral or negative. It is a really useful analysis since we could possibly determine the overall opinions and categorizing them helps in moulding the data according to our need.

In this project we tried to classify tweets from Twitter into positive,negative or neutral sentiment by building a model based on probabilities. Twitter is a microblogging website where people can share their feelings quickly and spontaneously by sending tweets limited by 140 characters. You can directly address a tweet to someone by adding the target sign "@" or participate to a topic by adding an hashtag "#" to your tweet. Because of the usage of Twitter, it is a perfect source of data to determine the current overall opinion about anything.

The project would heavily rely on techniques of "Natural Language Processing" in extracting significant patterns and features from the large data set of tweets and on "Machine Learning" techniques for accurately classifying individual unlabelled data samples (tweets) according to whichever pattern model best describes them.

II. VARIOUS MODELS USED FOR ANALYSIS :

RULE BASED APPROACH:

Rule Based Approach is the most basic approach that involves a list of positive, negative and neutral words the presence of which defines whether a sentence is positive or negative. This approach uses rules of thumb/heuristics to determine sentiments and uses Linguistics and Communications research to analyze sentiments.

SUPERVISED LEARNING:

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs.In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way.

STOCHASTIC GRADIENT DESCENT:

Stochastic Gradient Descent (SGD) is a simple yet very efficient approach to fitting linear classifiers and regressors under convex loss functions such as (linear) Support Vector Machines and Logistic Regression. The advantages of Stochastic Gradient Descent are:
1. It is easier to fit into memory due to a single training sample being processed by the network. It is computationally fast as only one sample is processed at a time
2. For larger datasets it can converge faster as it causes updates to the parameters more frequently.

LOGISTIC REGRESSION:

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

SUPPORT VECTOR MACHINE:

SVM is a supervised machine learning algorithm which can be used for classification or regression problems. It uses a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs. After giving an SVM model sets of labeled training data for each category, it is able to categorize new text. It is mostly used in classification problems.

RANDOM FOREST CLASSIFIER:

The random forest is a classification algorithm consisting of many decision trees. It uses bagging and features randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

III. DIFFICULTIES FACED:

Method to be chosen:

The method to be chosen was to based on the accuracy the model gives and the time complexity for each model. For this we had to run all the processes on each model and find the accuracy and time taken by each model. In the end, the model which gave the best figures for both was chosen. The final model chosen and submitted was Random Forest Classifier.

Text not getting properly cleaned up:

Cleaning the text is the first and the most important step in sentiment analysis. If the text is not well cleaned, the accuracy for sentiment analysis drops significantly and leads to an inaccurate and inefficient model, which is of no use. As a result, cleaning the text becomes important. The text is cleaned by removing unwanted data, like special characters, RT mentions. @mentions, hashtags, unwanted words like articles, proper nouns etc. Various methods were used for cleaning text. First such characters were counted and then the tweets were converted to lowercase. The spellings were corrected and all the repetitions of the unwanted words were removed. Then the verb forms were converted to the base forms and the text was then cleaned and ready to be trained.

Stopwords :

Removing a list of common words that add no particular meaning to the sentiment of the data is important as they do not help in the analysis. At the same time,using a generic list of stopwords can have a negative impact on sentiment analysis performance, as removing some common stop words like "don't", "not", "couldn't" etc. can change sentiment of a sentence. We tried to use a list of words that do not affect the sentiment of a particular text/sentence.

RAM crash:

On using lemmatization in our code,we faced runtime error as the RAM kept crashing even though not a lot of RAM was really being used. So we had to try a different approach to apply it. We also faced the same error in one of our models,for which we used a different approach as well.

TIME COMPLEXITIES:

Time complexity of an algorithm quantifies the amount of time taken by an algorithm to run as a function of the length of the input. Similarly, Space complexity of an algorithm quantifies the amount of space or memory taken by an algorithm to run as a function of the length of the input. Depending on the time taken in yielding the result we chose for our model. Lesser the time taken by the algorithm better would be the efficiency of the code.

ATTRIBUTE MISMATCH:-

Attribute Mismatch was encountered when there was a mismatch in the number of features in the training and testing data set. This was resolved by applying the tfidf transform function that converts the texts into an array that further helps in the processing.

INCORRECT ANALYSIS OF TWEETS:-

Tweets were incorrectly analysed due to the incorrect processing of tweets. When the tweets were converted into numbers based on their priorities, due to improper cleaning they were wrongly assigned with the numerical values. This yielded in the wrong analysis of the tweets. This was resolved by applying proper cleaning methods and improving the present ones. For eg: Various special characters along with the hyperlinks were removed using better regex functions and by using better lemmatization code each word was converted to its base form for better processing.

IV. FINAL MODEL:

Counting the number of special characters:-

Special characters, counted using word frequencies, which calculates how many times a word occurred in a particular string. For this, we created a variable which stores the frequency value found using a function from the Pandas library which is called, value_counts, which results in the unique value for word frequencies. By this, we can calculate the number of times a special character has occurred and we can remove it easily.

Converting tweets to lowercase:-

The first pre-processing step which we will do is transform our tweets into lower case. This avoids having multiple copies of the same words.Tweets are converted to lowercase using a lowercase function, which is hard coded. The text is split and manually converted into the lowercase. This is done to bring a uniformity into the text so that it can easily be analyzed.

Removal of multiple spaces, URL, RT, special characters:-

For removing the spaces characters RT and URL the regular expression library is used which is an in built library in python. It analyses the regular expressions and with the help of the code provides, removes all the unwanted occurrences of characters and rt and url.

Removal of stop words:-

A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query. NLTK (Natural Language Toolkit) in python has a list of stopwords stored in 16 different languages. We used this lost in the english language to remove all the occurrences of stopwords efficiently.

Removing commonly occurring and rarely occurring words:-

Remove rarely occurring words from the text. Because they’re so rare, the association between them and other words is dominated by noise. You can replace rare words with a more general form and then this will have higher counts. For this, we created a variable which stores the frequency value found using a function from the Pandas library which is called, value_counts, which results in the unique value for word frequencies. By this, we can calculate the number of times a special character has occurred and we can remove it easily.

Spelling correction:-

Usually when people tweet they commit various spelling mistakes, thus these need to be rectified before they can be used for sentiment analysis. Thus we used the TextBlob which is a python library for processing textual data. It provides a simple API for diving into common

natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more. Hence, the tweets were now spelled correctly.

Tokenization:-

Tokenization basically refers to splitting up a larger body of text into individual words.This method is used to analyse the tweets word by word. To do this we used TextBlob library's another inbuilt function TextBlob.word() that splits the given tweet into words and helps in analyzing the overall sentiment of the tweet. This step plays a vital role in analysing the sentiment of the tweets.

Lemmatization:-

Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma. While writing tweets people use various tenses and word forms that makes their analysis a bit tedious. Hence its advice to convert every word to their base form. Hence lemmatization is an important step in the cleaning of tweets.

Bag of words:-

This process is used to convert the tweets into a group or bag of words which is then used to calculate the frequency of the words. The frequency of words is a useful aspect when we consider sentiment analysis as we can remove the maximum occurring words and the rarely occurring words so that the intensity of the data set is reduced and it becomes more efficient.

Words to Numbers:-

This method is applied to convert words into numbers. This is done on the basis of the importance of the words, for example a word that plays a greater role in the sentiment analysis is assigned with a greater numerical value as compared to a word that contributes less in the sentiment analysis. The computer understands numbers hence this method plays an important role classifying the text as positive, neutral or negative.

MODELS:

STOCHASTIC GRADIENT DESCENT:

Stochastic Gradient Descent (SGD) is a simple yet very efficient approach to fitting linear classifiers and regressors under convex loss functions such as (linear) Support Vector Machines and Logistic Regression. The advantages of Stochastic Gradient Descent are:

- It is easier to fit into memory due to a single training sample being processed by the network.
- It is computationally fast as only one sample is processed at a time.
- For larger datasets, it can converge faster as it causes updates to the parameters more frequently.

LOGISTIC REGRESSION:

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

SUPPORT VECTOR MACHINE:

SVM is a supervised machine learning algorithm which can be used for classification or regression problems. It uses a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs. After giving an SVM model sets of labeled training data for each category, it is able to categorize new text. It is mostly used in classification problems.

RANDOM FOREST CLASSIFIER:

The random forest is a classification algorithm consisting of many decision trees. It uses bagging and features randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

V. RESULTS OBTAINED:

The results obtained include the accuracy we achieved with each model on splitting the given train dataset into train and test in the ratio of 80:20. This includes:

MODEL	ACCURACY(%)
SGD	60.26
LR	60.81
LR CV	61.89
SVM	58.28
RFC	61.68

VI. SCOPE OF IMPROVEMENT:

Improving accuracy:-

One of the ways to improve accuracy for logistic regression models is by optimising the prediction probability cutoff scores generated by your logit model. The Information Value package provides a way to determine the optimal cutoff score that is specific to your business problem. Even having more data is always a good idea. It allows the “data to tell for itself,” instead of relying on assumptions and weak correlations. Presence of more data results in better and accurate models.

Treat missing and Outlier values:-

The unwanted presence of missing and outlier values in the training data often reduces the accuracy of a model or leads to a biased model. It leads to inaccurate predictions. This is because we don’t analyse the behavior and relationship with other variables correctly. So, it is important to treat missing and outlier values well.

Feature Engineering:-

This step helps to extract more information from existing data. New information is extracted in terms of new features. These features may have a higher ability to explain the variance in the training data. Thus, giving improved model accuracy.

Feature Selection:-

Feature Selection is a process of finding out the best subset of attributes which better explains the relationship of independent variables with target variables.

Multiple algorithms:-

Hitting at the right machine learning algorithm is the ideal approach to achieve higher accuracy. But, it is easier said than done.Some algorithms are better suited to a particular type of data set than others. Hence, we should apply all relevant models and check the performance.

Algorithm Tuning:-

The objective of parameter tuning is to find the optimum value for each parameter to improve the accuracy of the model. To tune these parameters, you must have a good understanding of these meanings and their individual impact on the model. You can repeat this process with a number of well performing models.

Ensemble methods:-

This is the most common approach found majorly in winning solutions of Data science competitions. This technique simply combines the result of multiple weak models and produces better results. This can be achieved through many ways like Bagging (Bootstrap Aggregating or Boosting.

Cross Validation:-

This method helps us to achieve more generalized relationships.we must use cross validation technique. Cross Validation is one of the most important concepts in data modeling. It says, try to leave a sample on which you do not train the model and test the model on this sample before finalizing the model.