SYNOPSIS ON SENTIMENT ANALYSIS OF RESTAURANT REVIEWS

ABSTRACT

Sentiment analysis (also known as opinion mining or emotion AI) is the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine. Sentiment Analysis is the domain of understanding these emotions with software, and it's a must-understand for developers and business leaders in a modern workplace. Today we use natural language processing, statistics, and text analysis to extract, and identify the sentiment of words into positive, negative, or neutral categories.

INTRODUCTION

Natural language processing (NLP) refers to the branch of computer science—and more specifically, the branch of artificial intelligence or AI—concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.

NLP combines computational linguistics—rule-based modeling of human language—with statistical, machine learning, and deep learning models. Together, these technologies enable computers to process human language in the form of text or voice data and to 'understand' its full meaning, complete with the speaker or writer's intent and sentiment.

NLP drives computer programs that translate text from one language to another, respond to spoken commands, and summarize large volumes of text rapidly—even in real time. There's a good chance you've interacted with NLP in the form of voice-operated GPS systems, digital assistants, speech-to-text dictation software, customer service chatbots, and other consumer conveniences. But NLP also plays a growing role in enterprise solutions that

help streamline business operations, increase employee productivity, and simplify mission-critical business processes.

NLP enables computers to understand natural language as humans do. Whether the language is spoken or written, natural language processing uses artificial intelligence to take real-world input, process it, and make sense of it in a way a computer can understand. Just as humans have different sensors -- such as ears to hear and eyes to see -- computers have programs to read and microphones to collect audio. And just as humans have a brain to process that input, computers have a program to process their respective inputs. At some point in processing, the input is converted to code that the computer can understand.

METHODOLOGY

Program Statement:

Given a corpus of reviews for a restaurant on a food delivery app, analyse the sentiment behind each review. Classify the reviews into positive and negative and use it to give a rating (out of 5) to the restaurant.

1.DATA COLLECTION

Data Collection is one most important and crucial aspects of the Sentiment Analysis application. Due to the wide adoption of machine learning models, simply having large datasets on a domain specific task does not ensure superior performance. The performance of the model depends on the quality of dataset and labelling/annotation. As ML models learn from the data they are trained with, automatic predictions are likely to mirror the human disagreement identified during annotation. As a result, having a proper guideline to annotate data is also of utmost importance.

Using API **provided by social media platform which allows to collect data in a streaming fashion. Example: Twitter API to extract tweets by hashtags, NewsAPI to extract news by category from different news publishers.**

Using Web scrapers **that crawl up web data and collect specified information. It extracts data from webpage(HTML document).**

Using a Web browser plugin **with which users can extract information from any public website using HTML and export the data to the desired file format.**

Using existing open-source repositories of data that are cleaned and compiled which can be used directly**. Example: Rotten Tomatoes, IMDB movie review, Yelp, Amazon product review, Twitter tweets on Kaggle and from other websites.**

2.DATA PRE-PROCESSING

Data cleaning is an important technique for data pre processing tool. It is a process of Data Mining techniques. It removes the bad errors data and reduced unnecessary information of data. The missing of data are also included in data cleaning techniques. The presence of noise data may affect the intrinsic characteristic of a classification problem.

Stemming: **Stemming is a process of removing inflectional words which is affixes, for example playing-play, studies-study. Stemming works on some particular language mainly English and Spanish.**

Lemmatization: **Lemmatization takes the consideration of morphological analysis of the words. It reduces inflected words properly with the root words belongs to the sentences. It also called as lemma which is the set of words in dictionary form, citation form and canonical form.**

Data reduction**: The data reduction represents the original data that reduced to obtain a set of techniques to one way or another way those data needed to the distinction of data preparation to approximately suit the input data of DM task.**

3.DATA ANALSING

CountVectorizer **creates a matrix in which each unique word is represented by a column of the matrix, and each text sample from the document is a row in the matrix. The value of each cell is nothing but the count of the word in that particular text sample.**

In the l8th century, Reverend Thomas Bayes developed a method known as Naive Bayes that used probability and opportunity approaches. The workings of the Naive Bayes algorithm can be seen in Equation . Naive Bayes calculates future probability predictions from data or experiences that have been given, based on the opportunity point of view . One characteristic
of the Naive Bayes Classification is the existence of independent input variables which assume the presence of an articular feature from a class that is mutually independent of other features .

4.DATA PREDICTION

The data can be bested predicted by using Confusion Matrix, Accuracy, Precision, Recall, F1-Score.
Accuracy is a measure for the closeness of the measurements to a specific value, while *precision* is the closeness of the measurements to each other, i.e. not necessarily to a specific value. To put it in other words: If we have a set of data points from repeated measurements of the same quantity, the set is said to be accurate if their average is close to the true value of the quantity being measured. On the other hand, we call the set to be precise, if the values are close to each other. The two concepts are independent of each other, which means that the set of data can be accurate, or precise, or both, or neither.

A *Confusion matrix* is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives us a holistic view of how well our classification model is performing and what kinds of errors it is making.

For a binary classification problem, we would have a 2 x 2 matrix as shown below with 4 values:

	Predicted Review is Fake (Negative / 0)	Predicted Review is True(Positive / 1)
Review is Fake (0)	TN	FP
Review is Genuine (1)	FN	TP

True positive (TP): **The algorithm-predicted value is matched with the reality that news is fake. We can conclude that the algorithm has correctly classified and news shared in the social network is fake.**

False negative (FN): **The predicted output is a false negative, where news is incorrectly classified that the news shared is negative even though the news is genuine.**

True negative (TN)**: Predicted output is a true negative when the algorithm-predicted value is matched with the reality that news is genuine. We can conclude that the algorithm has appropriately classified.**
False positive (FP): **News is inaccurately classified that shared news is genuine news even though it is fake news.**
Negative (N): **A 0 value is used to represent a negative case, which means the news is genuine.**
Positive (P): **A value of 1 is used to represent a positive case, which means the news is fake.**

Once the confusion matrix was constituted, the performance of the data classification algorithms was compared by doing the comparative analysis using parameters classification accuracy, classification error, sensitivity or recall, specificity, precision,

Now we will find the *precision (positive predictive value)* in classifying the data instances. Precision is defined as follows:

Precision = $TP / (TP+FP)$

Precision **should ideally be 1 (high) for a good classifier. *Precision*** **becomes 1 only when the numerator and denominator are equal i.e *TP* = *TP* +*FP*, this also means *FP* is zero. As *FP*** **increases the value of denominator becomes greater than the numerator and *precision*** **value decreases (which we don't want).**

Now we will introduce another important metric called recall. Recall is also known as *sensitivity* **or *true positive rate*** **and is defined as follows:**

Recall = $TP / (TP+FN)$

Recall **should ideally be 1 (high) for a good classifier. *Recall* becomes 1 only when the numerator and denominator are equal i.e *TP* = *TP* +*FN*, this also means *FN* is zero. As *FN* increases the value of denominator becomes greater than the numerator and *recall* value decreases (which we don't want).**

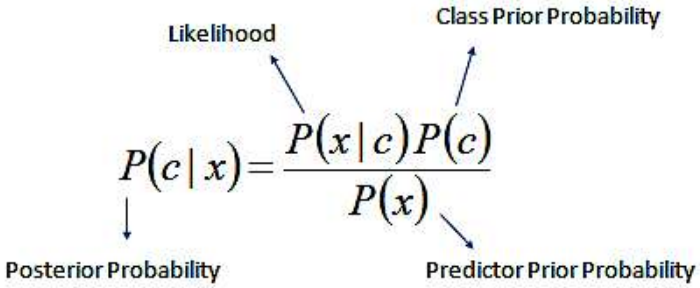
So ideally in a good classifier, we want both *precision* and *recall* to be one which also means *FP* and *FN* are zero. Therefore we need a metric that takes into account both *precision* and *recall*. *F1-score* is a metric which takes into account both *precision* and *recall* and is defined as follows:

$$\text{Fl-Score} = 2 * \frac{Precision * Recall}{Precison + Reca}$$

ALGORITHM

1.Multinomial Naïve Bayes Theorem

In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive)independence assumptions between the features. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum likelihood training can be done by evaluating a closed form expression (mathematical expression that can be evaluated in a finite number of operations), which takes linear time. It is based on the application of the Baye's rule given by the following formula:



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

2. Binomial Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. Logistic Regression models the data using the sigmoid function.

It is a mathematical function having a characteristic that can take any real value and map it to between 0 to 1 shaped like the letter "S". The sigmoid function also called a logistic function.

$$Y = 1 / 1+e^{-z}$$

So, if the value of z goes to positive infinity then the predicted value of y will become 1 and if it goes to negative infinity then the predicted value of y will become 0. And if the outcome of the sigmoid function is more than 0.5 then we classify that label as class 1 or positive class and if it is less than 0.5 then we can classify it to negative class or label as class 0.

PROGRAM- MODEL

Step 1 **: Reading a Restaurant Review csv file and storing the data.**

Step 2 : **Sorting the reviews according to the Rate . Rates between 5-3 are considered as Positive (P) reviews .Whereas Rates between 2-0 are considered as Negative (N) reviews.**
For eg :
"The food was great super super delicious " ,Rate: 4/5 , Sentiment : P
"Pasta was not upto to mark. It was too spicy !! Could have been great " , Rate: 2/5 , Sentiment : N

Step 3 : **Data cleaning and pre processing. i.e. Removing punctuations, stemming, lemmatizing, removing stopwords and splitting the sentences in words.**
For eg : The above two sentences now become
Food great super delicious
Pasta mark spicy great

Step 4 : **Using CountVectorizer and .fit( ) and .transform( ) functions counting and the frequency of each words and storing it in a matrix form. .fit( ) function is used to learn a vocabulary dictionary of all tokens in the raw documents. .transform ( ) function is used to transform documents to document-term matrix.**
Lets consider that these many times each words were used . The words are tokenized with the number of tines they have occurred
Food – 1 Great – 2 Super – 4 Delicious - 10
Pasta – 6 Mark – 3 Spicy – 9

After .transform( )
(0,1) 1
(0,2) 1
(0,4) 2 – this means that the word "superb" has occurred twice in the first sentence
(0,10) 1
(1,6) 1
(1,3) 1
(1,9) 1
(1,2) 1

Step 5 : **The Dataset will now be trained and used for prediction by using train_test_split from sklearn.model_selection .**
Train_test_split is used to train some part of the data lets say 80% data is used to train the system and other 20% data is used for testing i.e. prediction
X_train is used to train the data set
X_test is used to test the data set
Y_train is used to set the labels to all the data in X_train
Y_test is used to set the labels to all the data in X_test.

Step 6 : **The Multinomial Naïve Bayes algorithm is then implemented by passing X_train and Y_train datas as parameters .**
The dataset is then tested by passing the remaining X_test dataset in classifier.predict( )function where the dataset is tested according to the conditions and the trained dataset and stored in Y_predict .

Step 7 : **Predicted dataset is checked by forming the confusion matrix of the Y_test(tested dataset) and Y_predict (predicted dataset)**
The confusion matrix is then further checked for its accuracy , recall and f1_scores .

In similar ways algorithms such as Logistic Regression , Guassian Naïve Bayes , SVC and KNeigbhorsClassifier can also be used for the Sentimental Restaurants Reviews .

CONCLUSION

Nowadays, sentiment analysis or opinion mining is a trending topic in machine learning. We are still far to detect the sentiments of corpus of texts very accurately because of the complexity in the English language and even more if we consider other languages too. In this project we tried to show the basic way of classifying restaurant reviews into positive or negative category using Naive Bayes as baseline and how language models are related to the Naive Bayes and can produce better results. We could further improve our classifier by trying to extract more features from the restaurant reviews, trying different kinds of features, tuning the parameters of the naïve Bayes classifier, or trying another classifier all together.