

***Sentimental Analysis in Twitter Using Machine Learning Techniques.***

CSC354 - Machine Learning

Members:

**Muhammad Asif Khan SP20-BCS-130**

**Hamza Chaudhary SP20-BCS-082**

**Shahmeer Hameed SP20-BCS-061**

**Luqman Kareem Butt SP20-BCS-031**

***Abstract***

Sentiment analysis deals with identifying and classifying opinions or sentiments expressed in source text. Social media is generating a vast amount of sentiment rich data in the form of tweets, status updates, blog posts etc. Sentiment analysis of this user generated data is very useful in knowing the opinion of the crowd.

Twitter sentiment analysis is difficult compared to general sentiment analysis due to the presence of slang words and misspellings. The maximum limit of characters that are allowed in Twitter is 140.

Knowledge base approach and Machine learning approach are the two strategies used for analyzing sentiments from the text. In this project, we try to analyze the twitter posts about politics and celebrity tweets using Machine Learning approach. By doing sentiment analysis in a specific domain, it is possible to identify the effect of domain information in sentiment classification. We present a new feature vector for classifying the tweets as positive, negative and extract peoples' opinion and identify if the person is depressed or not.

# **INTRODUCTION**

Twitter has emerged as a major micro-blogging website, having over 100 million users generating over 500 million tweets every day. With such large audience, Twitter has consistently attracted users to convey their opinions and perspective about any issue, brand, company or any other topic of interest. Due to this reason, Twitter is used as an informative source by many organizations, institutions and companies.

On Twitter, users are allowed to share their opinions in the form of tweets, using only 140 characters. This leads to people compacting their statements by using slang, abbreviations, emoticons, short forms etc.

Along with this, people convey their opinions by using sarcasm and polysemy. Hence it is justified to term the Twitter language as unstructured. In order to extract sentiment from tweets, sentiment analysis is used. The results from this can be used in many areas like analyzing and monitoring changes of sentiment with an event, sentiments regarding a particular brand or release of a particular product, analyzing public view of government policies etc.

A lot of research has been done on Twitter data in order to classify the tweets and analyze the results. In this paper we aim to review of some research in this domain and study how to perform sentiment analysis on Twitter data using Python. The scope of this project is limited to that of the machine learning models and we show the comparison of efficiencies of these models with one another.

## Twitter Sentiment Analysis

The aim while performing sentiment analysis on tweets is basically to classify the tweets in different sentiment classes accurately. In this field of research, various approaches have evolved, which propose methods to train a model and then test it to check its efficiency. Performing sentiment analysis is challenging on Twitter data, as we mentioned earlier.

Here we define the reasons for this:

* **Limited tweet size:**

With just 140 characters in hand, compact statements are generated, which results sparse set of features.

* **Use of slang:**

These words are different from English words, and it can make an approach outdated because of the evolutionary use of slangs.

* **Twitter features:**

It allows the use of hashtags, user reference and URLs. These require different processing than other words.

* **User variety:**

The users express their opinions in a variety of ways, some using different language in between, while others using repeated words or symbols to convey an emotion.

All these problems are required to be faced in the pre-processing section. Apart from these, we face problems in feature extraction with less features in hand and reducing the dimensionality of features.

# **METHODOLOGY**

In order to perform sentiment analysis, we are required to collect data from the desired source (here Twitter). This data undergoes various steps of pre-processing which makes it more machine sensible than its previous form.

## Tweet Collection

Tweet collection involves gathering relevant tweets about the particular area of interest. The tweets are collected using Twitter’s streaming API or any other mining tool (for example we used Twint in our project), for the desired time period of analysis. The format of the retrieved text is converted as per convenience (for example excel file in this case).

The dataset collected is imperative for the efficiency of the model. The division of dataset into training and testing sets is also a deciding factor for the efficiency of the model. The training set is the main aspect upon which the results depends.

Text

Description automatically generated with medium confidence

## Pre-processing of tweets

The preprocessing of the data is a very important step as it decides the efficiency of the other steps down in line. It involves syntactical correction of the tweets as desired. The steps involved should aim for making the data more machine readable in order to reduce ambiguity in feature extraction. Below are a few steps used for pre-processing of tweets -

1. Removal of re-tweets.
2. Converting upper case to lower case.
3. Twitter feature removal like URLs, Hashtags, Mentions, Punctuations and Numbers, etc.

## Feature Extraction

Various methodologies for extracting features are available in the present day. Term frequency-Inverse Document frequency is an efficient approach. TF-IDF is a numerical statistic that reflects the value of a word for the whole document (here, tweet).

Scikit-learn provides vectorizers that translate input documents into vectors of features. We can use library function TfidfVectorizer(), using which we can provide parameters for the kind of features we want to keep by mentioning the minimum frequency of acceptable features.

Table

Description automatically generated with low confidence

# **EXPERIMENTATION FOR MODEL VALIDATION**

The collected dataset is divided in two– training set and testing set. The training set is used to train the classifier (machine learned model) while the testing set is the one on which the experimentation is performed. The ratio of training and testing dataset can vary as per to applications. This method selects 80% for training set and 20% for testing.

The Word Cloud of the dataset provides us with the information about how many times a word is used in our dataset which is provided below:

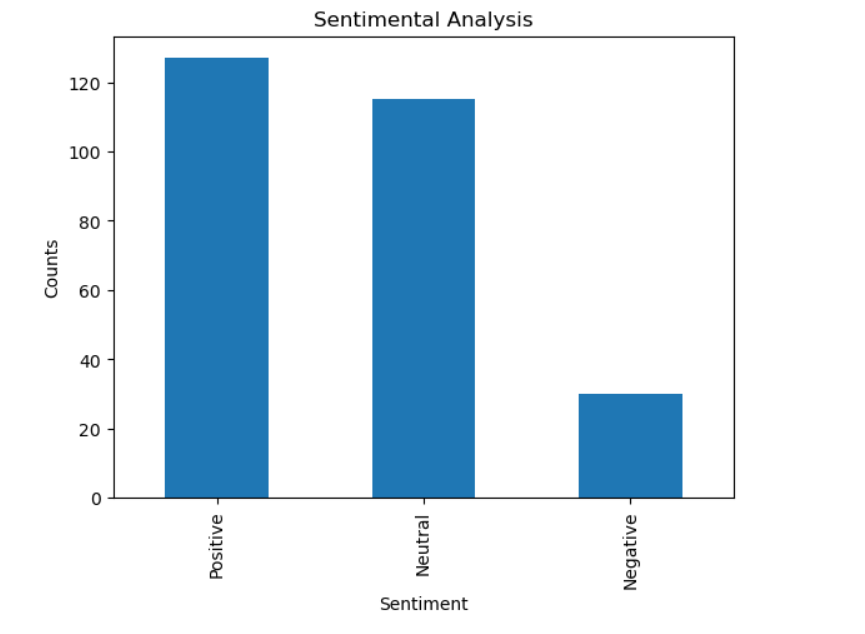
Text

Description automatically generated

Chart, scatter chart

Description automatically generatedThe Gaussian Naïve Bayes model which is chosen for experimentation is trained using the training set of data. Then this same trained model is used to classify new data, by which we can check its accuracy. It predicts if the user is depressed or not based on the tweets in the dataset.

After that, we use TextBlob to identify the polarity and subjectivity of the tweets. The polarity is then used to analyze if the tweet is positive, negative, or neutral. In this case, the positive tweets were 46.7%, negative tweets were 42.3%, and the neutral tweets were 11.0%.



Based on the value counts we plotted this graph.

# **CONCLUSION**

Twitter sentiment analysis comes under the category of text and opinion mining. It focuses on analyzing the sentiments of the tweets and feeding the data to a machine learning model in order to train it and then check its accuracy, so that we can use this model for future use according to the results.

It comprises of steps like data collection, text pre-processing, sentiment detection, sentiment classification, training and testing the model. This research topic has evolved during the last decade with models reaching the efficiency of almost 85%-90%.

But it still lacks the dimension of diversity in the data. Along with this it has a lot of application issues with the slang used and the short forms of words. Many analyzers don’t perform well when the number of classes are increased.

Also, it’s still not tested that how accurate the model will be for topics other than the one in consideration. Hence, sentiment analysis has a very bright scope of development in future.