**Mini Project Report on**



**A TWITTER DATA SENTIMENTAL ANALYSIS**



**Submitted in partial fulfillment of the requirement for the award of the degree of**

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE & ENGINEERING**

**Submitted by:**

**Harshita Mehboobani** **2017330**

***Under the Mentorship of***

**Dr. A. Suresh Kumar**



**Department of Computer Science and Engineering**

**Graphic Era (Deemed to be University)**

**Dehradun, Uttarakhand**

**January 2023**



**CANDIDATE’S DECLARATION**

I hereby certify that the work which is being presented in the project report entitled **“A Twitter Data Sentimental Analysis Using Machine Learning”** in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineeringof the Graphic Era (Deemed to be University), Dehradun shall be carried out by the under the mentorship of **Dr. A. Suresh Kumar**, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Student Name: Harshita Mehboobani    University Roll No. : 2017330

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Description** | **Page No.** |
| Chapter 1 | Introduction | 1 |
| Chapter 2 | Literature Survey | 2 |
| Chapter 3 | Methodology | 3-4 |
| Chapter 4 | Result and Discussion | 5-6 |
| Chapter 5 | Conclusion and Future Work | 7-8 |
|  | References |  |

**Chapter 1**

**Introduction**

**1.1 Motivation for work**

Organisations primarily work over customers satisfaction, customer reviews about their

products. Shifts in sentiments on social platforms have been shown to correlate with shifts

observed in stock markets. Recognizing customer grievances thereby resolving them leads to

customer satisfaction as well as reliability of a business. Hence there exists a necessity of an

unbaised automated system to analyse customer reviews regarding any issue.

Sentiment classification gives some answers into what the major problemsare, from the

perspective of customers. As sentiment analysis can be automated, decisions can be made

based on a significant amount of data rather than plain intuition that isn’t always correct.

**1.2 Problem Statement**

Generating statistical data regarding views out of analysis of user’s opinions from the tweets,

which can act as an inference to understand how users feel thereby refining users

experiences. Despite the availability of software to extract data concerning a person’s opinion

on a particular product or service, organizations and other data workers still face problems

regarding the data extraction. With the rapid growth of the World Wide Web, people are

using social media platforms such as Twitter which generates huge volumes of opinionated

texts in the form of tweets which is available for the sentiment analysis. This translates to a

large volume of information from a human viewpoint which make it complicated to extract a

line, read them, analyse tweet by tweet, review them and organize them into an

comprerhensive format in a timely manner.

**Chapter 2**

**Literature Survey**

**2.1 EXISTING SYSTEM:**

The existing system, make use of a knowledge base approach to classify the tweets into either positive, negative or neutral. But, employing this existing approach results in less precision of the classification.

DISADVANTAGES OF EXISTING SYSTEM:

In Existing System, they have employed Lexicon based approach to compute the sentiment of the data coming from twitter which resulted in lower accuracy rate.

Also, there is a lot of overhead while computing the sentiment of a sentence, because for each word this method retrieves the sentiment from a predefined word dictionary(generally SentiWord)

**2.2 PROPOSED SYSTEM:**

In the proposed system, we try to analyze the sentiments of the tweets about automated products like mobiles, laptops etc using Data Mining approach.

By doing data sentimental analysis in a specific domain, it is possible to detect the effect of domain information in sentiment classification.

In proposed system we do a comparative analysis on finding the sentiment using two different Machine Learning algorithms Naïve Baye's Method and Support Vector Machine(SVM).

ADVANTAGES OF PROPOSED SYSTEM:

In proposed system we have computed Data Mining Techniques which resulted in enhancing the accuracy rate for finding the sentiment of data.

The overhead on the supervised learning algorithms has been drastically reduced because of absence of the predefined datasets to find out the sentiment of each word, thereby directly increasing the efficiency.

We have used WordCloud and Pie Chart to represent the final sentiment visually which could help the user to apprehend the sentiment more easily.

**Chapter 3**

**Methodology**

The Twitter data sentimental analysis is an emerging field that needs much more care. We used Tweepy, an API to stream live tweets from Twitter. User can choose a keyword of his choice and tweets containing that keyword are collected and stored into a csv file. Then we make a labeled dataset of the collected tweets using textblob and setting up the sentiment fields accordingly. Hence our training data set is ready without preprocessing. Next we perform cleaning of our dataset to clean, remove unwanted text, characters out of the tweets. Training classifier is done by fitting the training data to the learner or classifier ,there after predicting the results over unseen test.

This in turn provides us with the accuracy with which the classifier had predicted the outcomes. There after we present our results in form of a pictorial representation which is a great way to showcase results because of its easiness to comprehend information out of it.

**Tools Used**

►Twitter API

* Twitter is the best source of data when it comes to public opinion. It comprises some sensitive information. Not anyone has access to this data. As of 20th October 2018, one need to fill developers form to get the access of this API.

►Python: Specifically, following libraries of the python are used:

* Tweepy: To interact with the Twitter API.
* TextBlob: To do textual and sentimental analysis.
* Matplotlib: To plot the results in a Pie chart to showcase the final sentiment.

**3.1 Description:**

To be able to extract Twitter data programmatically in order to access tweets we require to create and register an app on twitter developers website for authentication and thereafter we can access data by using Twitter API.

**2.2 Registering App:**

Create a new app <https://apps.twitter.com/> to register the twitter app. On registering the twitter app consumer key and consumer\_secret\_key are received. Get access\_token and access\_token\_secret from the configuration page of the app. These will be used to get access to twitter on behalf of our application. The authentication tokens must be kept private as they are more likely to get misused. Creating a separate config file and keeping these tokens in that file would be a good practise.

**3.3. Accessing Data :**

Twitter offers REST API's in order to connect with their service. We will use one python library to access the twitter REST API's called Tweepy. It provides wrapper methods to easily access twitter REST API. Below command can be used to install Tweepy.

pip install tweepy.

**3.4 Storing Data:**

Now we extract all tweet data from personal profile and it is then stored for analysing. Tweepy library offers simple cursor interface in order to iterate through all the tweets and then it stores them into a file.

**3.5 Preparing Data:**

Before we start analyzing the twitter data, it's vital to comprehend the structure of the tweet as well as pre-process the data to eliminate non-useful terms called stop-words. Preprocessing of data in data analysis is the very important step. Preprocessing is in the simple term means to take in the data and formulate the data for optimal output considering our need.

**3.5.1 Tokenizing the tweet** Tokenization is one of the most basic, yet most important, steps in text analysis. The purpose of tokenization is to split a stream of text into smaller units called tokens, usually words or phrases. We will use python NLTK library to tokenize the tweets. Even NLTK Sbrary needs some preprocessing steps to correctly tokenize @mentions and #hashtags. We use regular expressiom to provide exceptions for mentions and hashtags.

**3.5.2 Removing Stop-Words** Stop-word elimination is one crucial step that should be considered during the pre-processing phases. Stop-words are most popular and common words of any language. While their use in the language is important, they don't usually convey a particular meaning, especially if taken out of context. This is the case of articles, conjunctions, some adverbs, etc. which are commonly called stop- words. Some libraries provide default stop words for different langages NLTK library provides default stop-words for English language,

**3.6 Fitting Data to Classifier and predicting test data:**

Trained data is fitted to a suitable classifier on feature extraction , then once the classifier is trained enough then we predict the results of the test data using the classifier, then comparing the original value to the value returned by the classifier.

**Chapter 4**

**Result and Discussion**

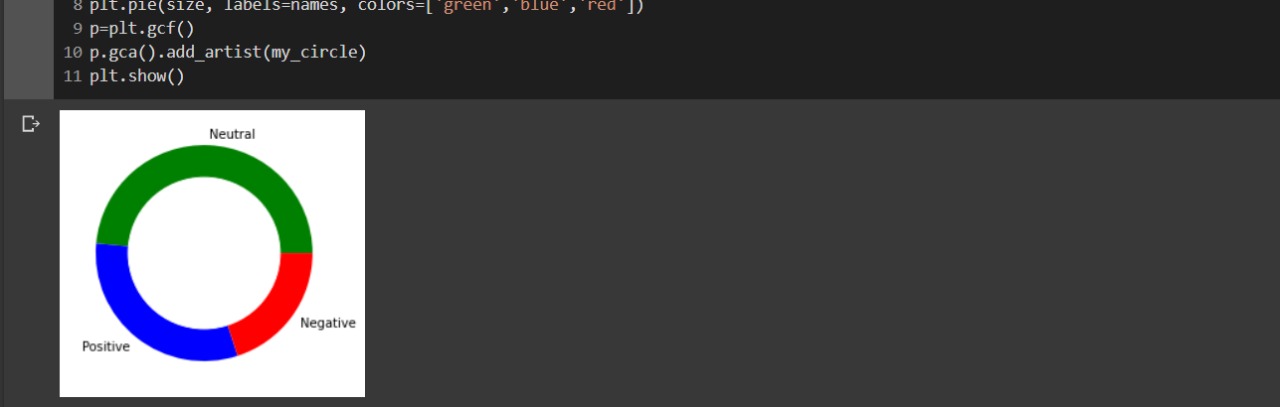
**4.1 Result Analysis:**

Here the accuracy of different classifiers are shown among which the best

classifier with highest accuracy percent is the chosen. Some factors such as polarity-scores,

f-score,mean,variance etc., also accounts for consideration of the classifiers.

**4.2 Visual Representation:**

* We have used WordCloud and Pie Chart to represent the final sentiment visually which could help the user to apprehend the sentiment more easily.Pictorial representation is the best way to convey information without much efforts.Thus it is chosen  
    
    
  For a labelled dataset of the collected tweets when the user inputs keyword ‘LOCKDOWN’  
    
    
    
  **PIE CHART :**   
    
    
    
    
  Final Sentiment is NEUTRAL

**WORDCLOUD :**



**Chapter 5**

**Conclusion and Future Work**

**5.1 Conclusion:**

* The field of sentiment analysis is an exciting new research direction due to large number of real-world applications where discovering people's opinion is important in better decision-making.
* Recently, people have started expressing their opinions on the Web that increased the need of analyzing the opinionated online content for various real-world applications.
* A lot of research is present in literature for detecting sentiment from the text. Still, there is a huge scope of improvement of these existing sentiment analysis models. Existing sentiment analysis models can be improved further with more semantic and commonsense knowledge.

**5.2 Future Scope :**

* Data Pre-Processing using more parameters to get best sentiments.
* Updating Dictionary for new Synonym and Antonyms of already existing words.
* Web-Application can be converted to Mobile Application.
* Multi-lingual support: Due to the lack of multi-lingual lexical dictionary, it is current not feasible to develop a multi-language based sentiment analyser.
* Analysing sentiments on emoji/smileys.

**References**

  [1] Alexander Pak, Patrick Paroubek. 2010, Twitter as a Corpus for Sentiment Analysis and

Opinion Mining.

[2]<https://www.slideshare.net/RaviKumar3589/new-sentiment-analysis-of-tweets-using-python-by-ravi-kumar>

[3] Jin Bai, Jian­Yun Nie. Using Language Models for Text Classification.

[4] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau.

Sentiment Analysis of Twitter Data.

[5] <https://www.slideshare.net/AntaraBhattacharya12/python-report-on-twitter-sentiment-analysis>

[6] <https://marcobonzanini.com/2015/03/17/mining-twitter-data-with-python-part-3-term-frequencies/>