Sentimental Analysis of Twitter Data

**Project Report**

**by**

**Rohit Saha (893572115)**

**CPSC 481-01 (20712)**

**Spring, 2019**

**Professor: Dr. Wenlin Han**

**Department of Computer Science**

**California State University, Fullerton**

**May 9, 2019**

**Table of Contents**

|  |  |
| --- | --- |
| Topics | Page Number |
| 1. Abstract | 4 |
| 1. Introduction | 5 |
| 1. Review of the Literature | 9 |
| 1. Methodology | 13 |
| 1. Results and Discussion | 18 |
| 1. Conclusion and Implications | 22 |
| 1. Recommendations | 22 |
| 1. References | 24 |

**List of Figures**

1. Figure 1: Architecture of Sentimental Analysis …..……………………...………………6
2. Figure 2: Naive Bayes - ML algorithm for classification  ……………………………....10
3. Figure 3: Gaussian Naive Bayes classifier ………………………………………………10
4. Figure 4: Kernel machines are used to compute non-linearly separable functions into a higher dimension linearly separable function …………………………………………...11
5. Figure 5: Typical Natural Language Processing Data Flow …………………………….11
6. Figure 6: Observation for Individual SVM’s Coordinates ………………………………12
7. Figure 7: Sample dataset from Kaggle …………………………………………………..14
8. Figure 8: System Architecture of Data Preprocessing and Data Visualization ………….16
9. Figure 9: System Architecture of the Classifiers ………………………………………...16
10. Figure 10: Schematic representation of methodology for the sentiment calculation ........17
11. Figure 11: Sentimental Analysis of Twitter Data ………………………………………..20
12. Figure 12: Consolidated data from all the Classifiers: NLP, Naïve Bayes and SVM .......20
13. Model Comparison for NLP, Naïve Bayes and SVM…………………………………....21

**Abstract**

Social Media and social networking especially Twitter is the new trendsetters. This marked the beginning of Big Data. Analysis of data to extract knowledge from the World Wide Web repository is a tedious and challenging process. This paper focused primarily on how the data from Twitter website could be analyzed for knowledge gathering through sentimental analysis. For our project we have used the data from twitter using the Twitter API which could be classified into positive, negative or neutral. The knowledge from people’s sentiment could help to extract reviews and customer feedback to improve business establishments. With the help from various Machine Learning Algorithms, Data Mining process analysis of historical data to determine sentiments of the people based on the posts, tweets, comments and likes shared on social media could be evaluated. The primary task being analysing and predicting messages based on human sentiment. We have used various algorithms to carry out this activity such as Naïve Bayes Classifier (NBC), Natural Language Processing (NLP)and SVM (Support Vector Machine). The collected dataset has been first cleaned and trained, and the required information through sentimental analysis has been gathered from the dataset. Furthermore, we are analysing the Twitter data in the real time basis. This would help us understand people’s sentiment on a particular topic and help us obtain useful insights from the same. We will be able to get the people’s sentiment for any latest topic and understand the reaction of common people and take future decision accordingly.

Keywords: Sentimental Analytics, SVM, NLP, Algorithms, Twitter API.

**1. Introduction**

**1.1 Background:**

Sentiment analysis, also refers as opinion mining, is a sub machine learning task where we want to determine which is the general sentiment of a given document. 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 determine the overall opinion about a selling-objects, or predict stock markets for a given company like, if most people think positive about it, possibly its stock markets will increase, and so on. Sentiment analysis is actually far from to be solved since the language is very complex (objectivity/subjectivity, negation, vocabulary, grammar) but it is also why it is very interesting to working on. In this project I choose to try 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 a 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 a 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.

Using tweets as data makes it challenging. Tweets contain emoticons, misspelled words, slangs which makes data extraction for sentimental analysis difficult. Certain ways to ascertain value from the data has been devised to format the information and obtain maximum knowledge from the same (M. F. Çeliktuğ, 2018). The methods proposed in my research helps to learn the data abnormalities and then analyze sentimentally for knowledge gathering. This sentimental analysis would help business organizations extract data from public reviews and comments and help them decide future strategies to better the company returns.

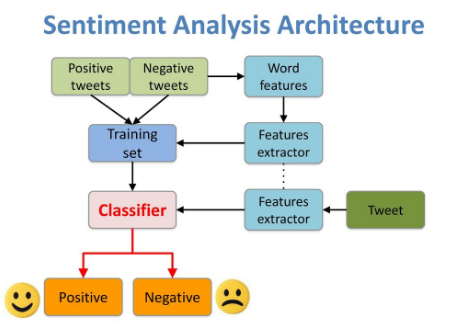


Figure 1: Architecture of Sentimental Analysis. (Min Yuh Day, 2016)

**1.2 Problem Statement:**

The main goal of this project is to find out the “positive”, “negative” and the “neutral” sentiment of any topic that the user will be enter using our GUI. Furthermore, we are analyzing the Twitter data in a real time basis (topic is entered by the user). This analyzation can further be used to predict any good or bad review about any topic, any person or any current or past affair. Users can use our application to get reviews about any college before taking admission, reviews about any item before buying or selling and several other useful activities.

* 1. **Project Goals and Benefits:**
* To identify the sentiment about any item, place, person or any topic.
* Capability of the system to analyze or evaluate a large amount of datasets.
* Statistical analysis and visualizations of historical data to the users.
* Compare various machine learning algorithms and find the most efficient one.
* Any user can use this system to find the sentiment of any chosen topic.
  1. **Relevance and Significance:**
* Based on the trained model, users can use this system. System would notify them about the sentiment of any chosen topics based on the twitter data.
* Very useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics.
* Various organizations can extract insights from social data and that can be used in various forms for selling or buying products.
* Organizations can predict the future by using the Sentimental Analysis.
* System can be used to get detailed review and analysis of any person, place, current affairs.
  1. **Assumption and Limitations:**
* Text Analysis, especially when there is precious little additional text to place it in context such as is the case with a 140 character tweet, will always prove lacking unless it is coupled with additional sets of data.
* The problem is keyword-based sentiment detection can’t understand situations like this:

“Oh, yeah, Fast Food Restaurant. I just LOVE the 30 minute wait for my food.”

We humans understand sarcasm. We understand the sentiment of this comment is clearly negative. Yet a machine would flag it as positive, possibly even very positive because of the all-caps LOVE.

* Day is the primary feature that is used to Sentimental Analysis.
  1. **Sentimental Analysis:**

Sentiment analysis also known as opinion mining  refers to the use of [natural language processing](https://en.wikipedia.org/wiki/Natural_language_processing), [text analysis](https://en.wikipedia.org/wiki/Text_analytics), [computational linguistics](https://en.wikipedia.org/wiki/Computational_linguistics), and [biometrics](https://en.wikipedia.org/wiki/Biometrics) to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to [voice of the customer](https://en.wikipedia.org/wiki/Voice_of_the_customer) materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from [marketing](https://en.wikipedia.org/wiki/Marketing) to [customer service](https://en.wikipedia.org/wiki/Customer_relationship_management) to clinical medicine.

**2. Review of the Literature**

“Sentimental Analysis involves a huge dataset which is initially fed to the training algorithm to extract the knowledge using the values of the necessary parameters using the algorithm. We can use several annotated datasets which has the main sentimental polarity classes. The dataset in any analysis algorithm is always divided into two parts consisting of the training dataset and the test data set.

The accuracy of a sentiment analysis system is, in principle, how well it agrees with human judgments. This is usually measured by variant measures based on [precision and recall](https://en.wikipedia.org/wiki/Precision_and_recall) over the two target categories of negative and positive texts. However, according to research human raters typically only agree about 80% of the time. Thus, a program which achieves 70% accuracy in classifying sentiment is doing nearly as well as humans, even though such accuracy may not sound impressive. If a program were "right" 100% of the time, humans would still disagree with it about 20% of the time, since they disagree that much about any answer.

On the other hand, computer systems will make very different errors than human assessors, and thus the figures are not entirely comparable. For instance, a computer system will have trouble with negations, exaggerations, [jokes](https://en.wikipedia.org/wiki/Joke), or sarcasm, which typically are easy to handle for a human reader: some errors a computer system makes will seem overly naive to a human. In general, the utility for practical commercial tasks of sentiment analysis as it is defined in academic research has been called into question, mostly since the simple one-dimensional model of sentiment from negative to positive yields rather little actionable information for a client worrying about the effect of public discourse on e.g. brand or corporate reputation.”

* 1. **Learning an Algorithm**

Here the text has been prepared so that the text can be used to train an algorithm. There are several algorithms available:

* + - * 1. **Naïve Bayes:** In Naïve Bayes classifier the class of the text is determined using the theory of probability (positive, neutral or negative). In Gaussian Naive Bayes, continuous values associated with each feature are assumed to be distributed according to a Gaussian distribution(Geeksforgeeks, 2017).

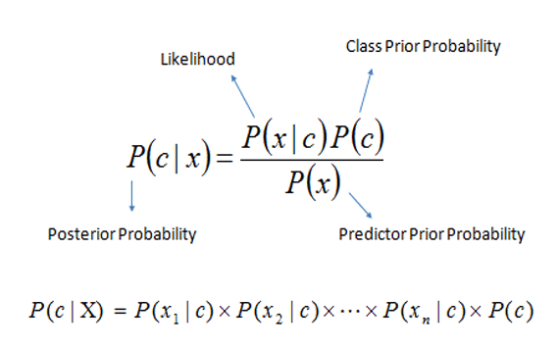


Figure 2: Naive Bayes - ML algorithm for classification (Sunil Ray, 2017)

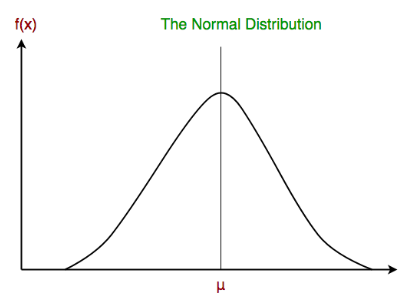


Figure 3: Gaussian Naive Bayes classifier (Geeksforgeeks, 2017)

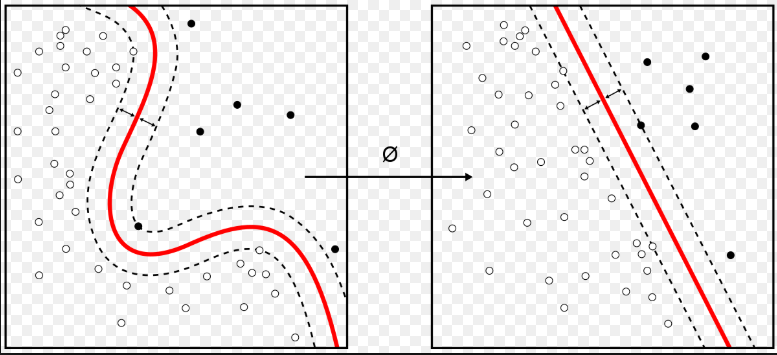


Figure 4: Kernel machines are used to compute non-linearly separable functions into a higher dimension linearly separable function. (Wikipedia, 2019)

* + - * 1. **Natural Language Processing (NLP):** Natural language processing (NLP) is a subfield of [computer science](https://en.wikipedia.org/wiki/Computer_science), [information engineering](https://en.wikipedia.org/wiki/Information_engineering_(field)), and [artificial intelligence](https://en.wikipedia.org/wiki/Artificial_intelligence) concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of [natural language](https://en.wikipedia.org/wiki/Natural_language) data.

Challenges in natural language processing frequently involve [speech recognition](https://en.wikipedia.org/wiki/Speech_recognition), [natural language understanding](https://en.wikipedia.org/wiki/Natural_language_understanding), and [natural language generation](https://en.wikipedia.org/wiki/Natural_language_generation).

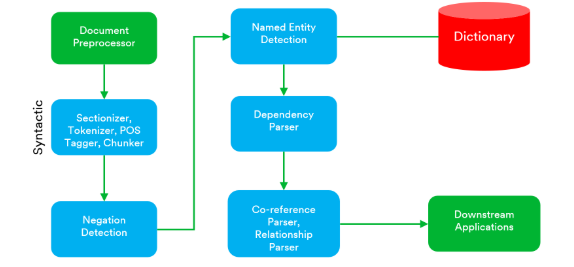


Figure 5: Typical Natural Language Processing Data Flow

* + - * 1. **Support Vector Machines (SVMs):** In this algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate (Sunil Ray, 2017). “Then, we perform classification by finding the hyperplane that differentiates the two classes very well.”

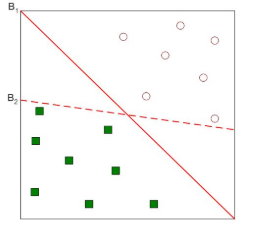


Figure 6: Observation for Individual SVM’s Coordinates. (Sunil Ray, 2017)

**3. Methodology**

**3.1 Project Tools:**

* **Jupyter Notebook**

According to Jupyter Notebook, it is an open-source web application developed by Fernando Perez. The notebook allows you to create and share the code written and even lets you create visualizations, equation, and narrative text. These are significant uses such as data visualization, simulation, data cleaning and transformation, machine learning, statistical modeling, and much more can be done using Jupyter Notebook.

Jupyter is a server-client web browser application where we can run the code and edit it. The notebook doesn't need any internet access and using remote server we can access it anywhere. Here we are using python language but when we are implementing in Jupiter notebook, its extension changes to ipynb.

* **Visual studio**

Visual studio is software where we can create applications and run the code, and it is developed by Microsoft. To create a front-end user interface where we can enter the input and predict the outcome. Visual Studio is a very fastest way to run the code and test often. To run the flask framework which incorporates both HTML has front-end and python has back-end.

* **Python**

Python is an interpreted high-level programming language which is also used for implementing the machine learning algorithms. Python 3.7 tool is available on the website and its free tool which acts as an interpreter between the Jupyter and the code. It is used in Jupyter to analyze the data using graphs and plots.

**3.2 Techniques:**

* **Pandas**: Using python, we can use open source library pandas which provides the tools required for the data analysis. The main purpose of this library is to prepare the data for machine learning algorithms.
* **NumPy:** Mathematical and scientific operations can be performed by using numpy even multi-dimensional can be handled by it. In this project, NumPy is used for performing some mathematical operation to the pandas library.
* **Scikit-learn:**  According to the scikit learn website, it is open-source Python machine learning library which provides numerous classifications, clustering algorithms, and regression. The main purpose of this library is to prepare the model and train the data using several machine learning algorithms.

**3.3 Data Requirements:**

* **Training Source of Data Collection**: In order to analyze twitter data, first, we need a dataset of tweets. We are looking into the dataset: **ProcessedPositive.csv, ProcessedNegative.csv** and **ProcessesNeutral.csv** consisting of 10000 tweets from Kaggle. It is composed of one columns that is tweet text. We are adding the sentiment as positive, negative and neutral accordingly to suit the tweet texts from the different files

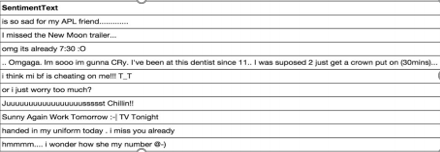


Figure 7: Sample dataset from Kaggle

* **Test Source of Data Collection:** For the test dataset, we are also looking at tweepy to extract twitter data. Tweepy is open-sourced, hosted on GitHub and enables Python to communicate with Twitter platform and use its API. We have installed Tweepy and using it to extract the dataset from twitter based on user specific data. Using the Consumer Key, Consumer Secret, Access Token and Access Token Secret, we are using the Twitter app to connect to the twitter database and extract dataset.

**3.4 System Requirements:**

* **Hardware Specifications:**
* **Processor**: 64-bit, core i5
* **RAM**: 8 GB
* **Operating System:** Windows 10
* **Hard Disk:** 160 GB
* **Software Specifications:**
* **Python 3.7.3**
* **Flask Framework**
* **Python Libraries**
* **Tweeter API - Tweepy**
* **Jupyter Notebook**
* **Visual studio**
* **Programming Languages Used:**
* **Python**
* **JavaScript**
* **HTML**
* **CSS**

**3.5 System Architecture:**

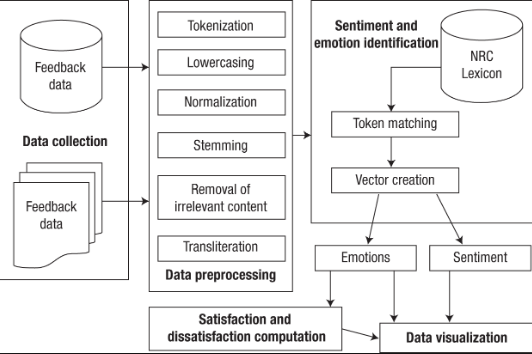


Figure 8: System Architecture of Data Preprocessing and Data Visualization

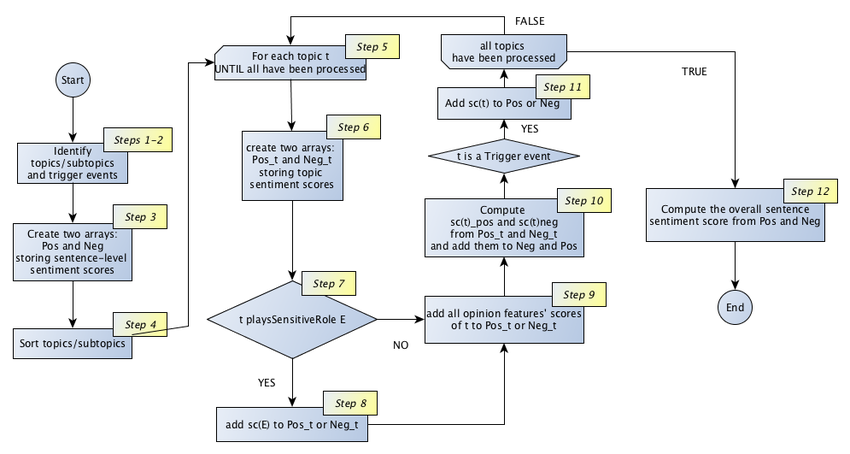
****

Figure 9: System Architecture of the Classifiers

**3.6 Data Analysis and Models Used:**

We have used three different models to train and test our Twitter data, NLP- Natural Language Processing, Naïve Bayes Classifier and SVM- Support Vector Machine. We also have divided our data set into a ratio of 80 and 20 for training and testing respectively. For the data analyzation the steps we followed were data preprocessing, part-of-speech tagging, stemming and lemmatization, handling of negation, and tokenization.

## 

Figure 10: Schematic representation of methodology for the sentiment calculation. ([Kolchyna](https://arxiv.org/search/cs?searchtype=author&query=Kolchyna%2C+O),  [Souza](https://arxiv.org/search/cs?searchtype=author&query=Souza%2C+T+T+P), [Treleaven](https://arxiv.org/search/cs?searchtype=author&query=Treleaven%2C+P), [Aste](https://arxiv.org/search/cs?searchtype=author&query=Aste%2C+T). 2015)

**4. Result and Discussion:**

* 1. **Data Analysis**: In order to perform the sentimental Analysis of The Twitterdata the user has to give input of the “Topic” which he wants to analyze. According to the user input, we will access the Twitter API and collect 200 latest test datasets from Twitter.

The data is divided into two phases which are training data and testing data. Training data consist of has been downloaded from Kaggle. It consists of three individual files:

**ProcessedPositive.csv**(tweet text of all positive tweets)

**ProcessedNegative.csv**(tweet text of all negative tweets)

**ProcessedNeutral.csv**(tweet test of all neutral tweets)

Total 10000 tweets have been used to train the model. The training data is used for fitting the models of machine learning algorithms and testing data is further used for prediction for the fitted model. The predicted values are then compared with original values of the testing data. The best algorithm is chosen on the basis of accuracy.

By implementing flask framework which main purpose is to implement python code as back-end and front-end as HTML. It takes input values from the HTML page which is the topic of analyzation.

**4.2 Data Preprocessing:**

There might some emoticons, bad words in the Twitter dataset. All other attributes which are required for the fitting the model are fully labelled and found completely clean.

Furthermore, the efficiency of the algorithms for sentimental analysis completely depends on the quality of the data that is fed to the algorithm. Therefore, formatting and cleaning of the data before feeding it to the algorithms to get the desired result is the most crucial step. Oversampling of the data is made with shuffling for randomization and omitting dominance of the examination order of each collected data. The several steps for the preprocessing are mentioned below:

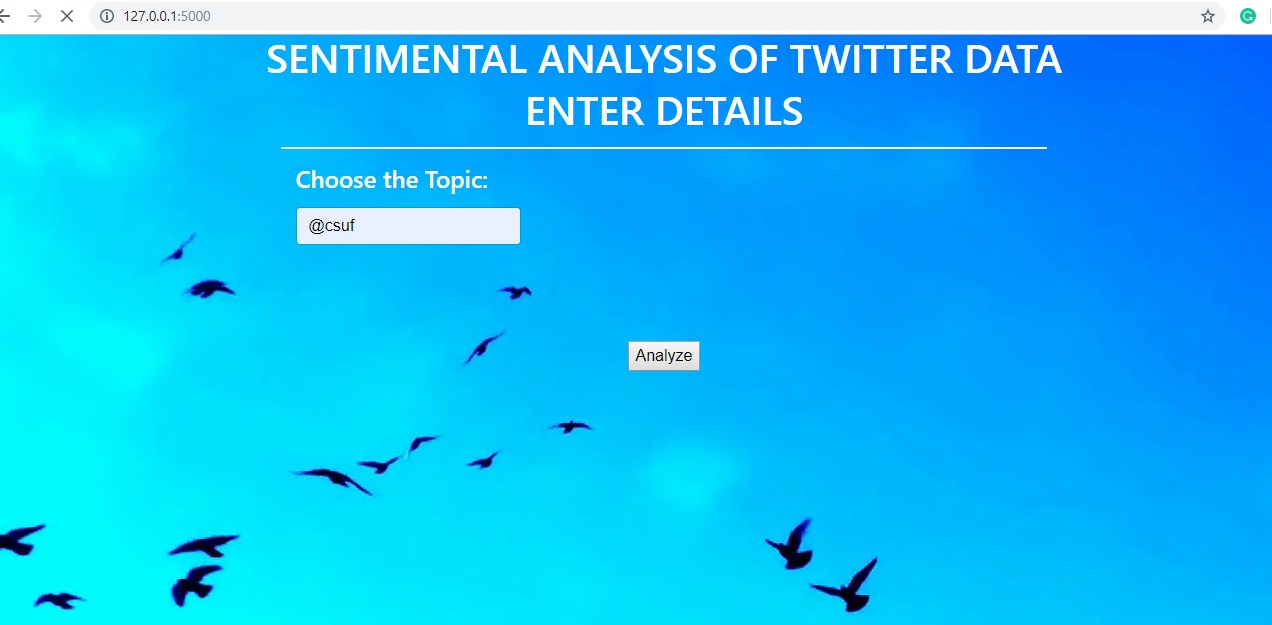
## Part-of-Speech Tagging (POS): This process helps in tagging each word of text automatically to which part of speech it belongs to such as verbs, nouns, adjectives, adverb, pronoun etc. The major goal of this technique is to extract the pattern in text format based on the analysis of frequency distribution of the above-mentioned parts-of-speech.

## Stemming and lemmatization: In the next step of stemming we have to replace the words with their roots or stems. For example, all the words such as “playing”, “played”, “play” should be mapped to one word “play”. Furthermore, while applying this technique we should also take care so that bias does not increase.

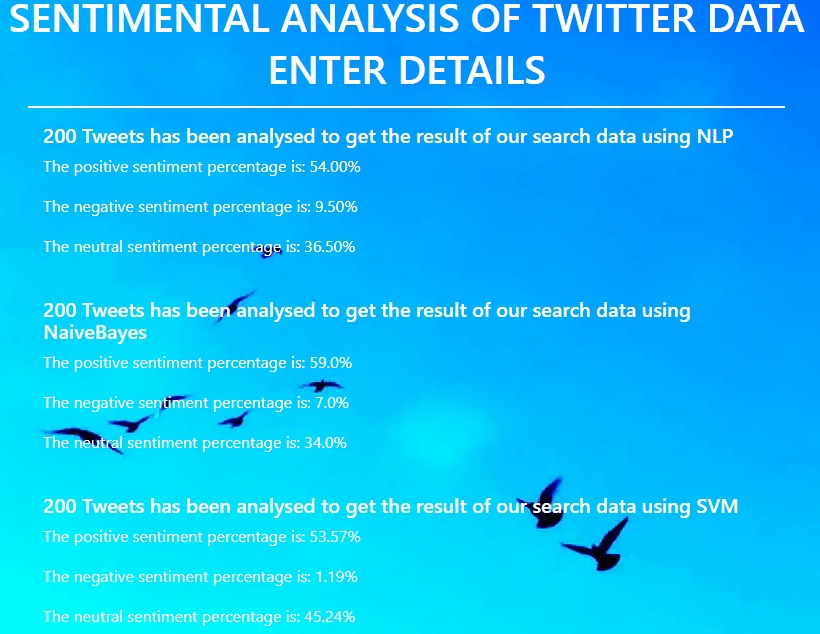
## Handling of Negation: In this process, the sentiment of the text is converted either from positive to negative or from negative to positive by using several words such as “do not”, “not”, “no” etc. The process to achieve this is to revert the polarity of the words that are found between the first punctuation mark and the negation.

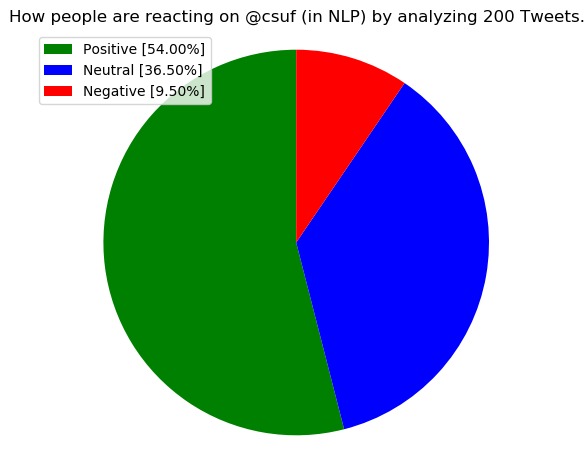
## Tokenization: For tokenization, we will create a bag full of words from the text collected. In this technique, we combine the accompanying word into different n-grams or phrases, which can be either trigram, bigram or unigram. A trigram is a combination of three neighboring words, bigram is the collection of two and so on.

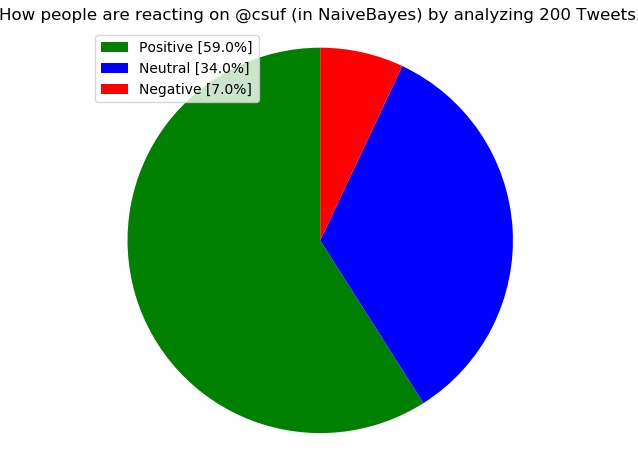
The below figure is the user interface of our HTML page and it includes the input parameter that is the topic to choose by the user which will be analyzed by our system. Also, the next picture is where all the data from 3 different algorithms NLP, Naïve Bayes and SVM will be consolidated together and compared.

Figure 11: Sentimental Analysis of Twitter Data

**4.3 Model Comparisons:**

All the three classifiers has been used for the sentimental analysis of the twitter data that has been entered by the user in the GUI. We have also plotted the results using the pie chart to get better visualization of the data. Figure 12: Consolidated data from all the Classifiers: NLP, Naïve Bayes and SVM





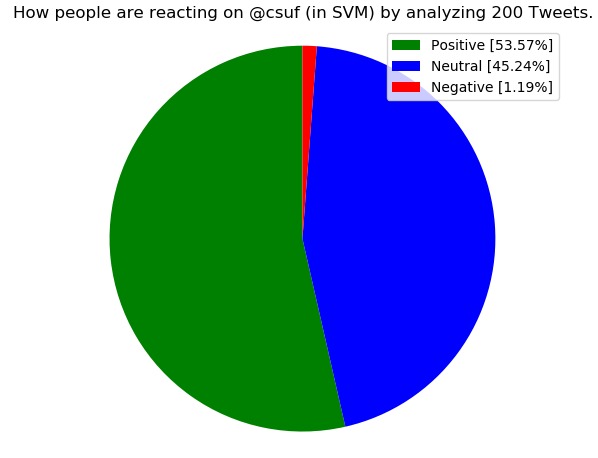


Figure 13: Model Comparison for NLP, Naïve Bayes and SVM

1. **Conclusion, Implication and Recommendation**
   1. **Conclusion:**

In a nutshell, in this era of the internet, it is extremely required to understand the sentiments of your customer and you target people in the market. And nothing can be better than the social media to track down various information about various people, product or any controversial topic. We can very easily by utilizing various machine learning algorithms such as Naïve Bayes, Decision Trees, Support Vector Machines (SVMs) determine people’s sentiments from collecting data from social media. These data can then be used for running your business strategically and generating more revenue out of it. This paper also discusses several benefits, advantages, and disadvantages of the used algorithms in this research paper.

* 1. **Implications:**
* Flask Framework is the easiest way to integrate HTML based web application with Python based backend.
* Python programming language is used for its ease of understanding and robustness.
* Sentimental Analysis can be done based on various attributes such as the topic and the number of data you want to analyze.
* Jupyter helps us to understand the effectiveness of analyzed data.
* Jupyter Notebook is the most effective open source tool for visualization of data using Graphs and charts.
* Understood significance of multiple Python libraries like NumPy and Scikit-Learn.
  1. **Future Work/ Recommendations:**
* Although the system right now if making analysis based on the historical data patterns, if future system can be enhanced to train model based on real-time data which would help in providing more accurate result.
* In future system can be implemented in such a way that it trains itself daily with more accurate data available on that day, so it becomes more accurate over time.
* Current system analyzes data based on the input provided by the user, the future system could be enhanced in such a way that it notifies user about the sentiments which there are currently browsing.
* In current dataset and the models they are not capable of understanding any sarcastic comments such as “Oh, yeah, Fast Food Restaurant. I just LOVE the 30 minute wait for my food.”

We humans understand sarcasm. We understand the sentiment of this comment is clearly negative. Yet a machine would flag it as positive, possibly even very positive because of the all-caps LOVE.

* The one way of increasing the accuracy of the system is by adding more features to the system which would help in making analyzation of sentiment more efficient and accurate.

**References**

Min Yuh Day, 2016, “*Sentimental Analysis in Social Media*” Retrieved from <https://slideplayer.com/slide/12541406/>

M. F. Çeliktuğ, 2018. "Twitter Sentiment Analysis, 3-Way Classification: Positive, Negative or Neutral?," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, pp. 2098-2103.

Shekhar Amit. (2018). “What Is Feature Engineering for Machine Learning?”. Retrieved  from <https://medium.com/mindorks/what-is-feature-engineering-for-machine-learning-d8ba3158d97a>

[Olga Kolchyna](https://arxiv.org/search/cs?searchtype=author&query=Kolchyna%2C+O), [Tharsis T. P. Souza](https://arxiv.org/search/cs?searchtype=author&query=Souza%2C+T+T+P), [Philip Treleaven](https://arxiv.org/search/cs?searchtype=author&query=Treleaven%2C+P), [Tomaso Aste](https://arxiv.org/search/cs?searchtype=author&query=Aste%2C+T). (2015). *Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination.*

Sunil Ray, 2017, “*6 easy steps to learn Naïve Bayes Algorithm*” Retrieved from <https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/>

Geeksforgeeks, 2017, “*Naïve Bayes Classifier*” Retrieved From <https://www.geeksforgeeks.org/naive-bayes-classifiers/>

[Emma Grimaldi](https://towardsdatascience.com/@emmagrimaldi). (2018). “*Decision Tree: an algorithm that works like the human brain*.” Retrieved from <https://towardsdatascience.com/decision-tree-an-algorithm-that-works-like-the-human-brain-8bc0652f1fc6>

Sunil Ray. (2017). *“Understanding Support Vector Machine algorithm from examples.”* Retrieved from <https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/>

Wikipedia, 2019. “*Naïve Bayes Classifier*” Retrieved from <https://en.wikipedia.org/wiki/Naive_Bayes_classifier>

[Shiho Hashimoto](https://blog.insightsatlas.com/author/shiho-hashimoto). (2017). “7 Benefits of Sentiment Analysis You Can’t Overlook.” Retrieved from <https://blog.insightsatlas.com/7-benefits-of-sentiment-analysis-you-cant-overlook>