**MINI PROJECT – I**

**(2019-20)**

# Sentimental Analysis Using Python

# 

**SYNOPSIS**



**Institute of Engineering & Technology**

**Team Members**

PUSHKAR KRISHNA OJHA

(171500246)

## Supervised By

**Mr.AMIR KHAN**

**Technical Trainner**

**Department of Computer Engineering & Applications**

**About the Project:**

With the huge amount of increase in the web technologies, the no of people expressing their views and the opinion via web are increasing. This information is useful for everyone like businesses, governments and individuals . with 500+ million tweets per day , twitter is becoming a major source of information. Twitter is a microblogging site, which is popularly known for its short messages known as tweets. It has a limit of 140 characters. Twitter has a user base of 240+ million active users and hence it is a useful source of information. The users often discuss their personal views on various subjects and also on current affairs via tweets. Out of all popular social medias like Facebook , Twitter, Google+, and Myspace we choose Twitter because of the reasons like

* Twitter contains vast number of text posts and it grows day by day. The collected corpus can be arbitrarily large.
* Twitter’s audience varies from regular users to celebrities, Politicians , company representatives, and even country’s president. Therefore it is possible to collect text posts of users from different social and interests groups.
* Tweets are small in length and thus less ambiguous and are unbiased in nature.

            Using social media, models are built for classifying “tweets” into positive , negative, and neutral classes .The models are build for two classification tasks : a 3-way classification of already separated phrases in a tweet into positive, negative , and neutral classes and another 3 way classifications of entire message into positive , negative and neutral classes.

            This paper is experimented with baseline model and feature based model. An incremental analysis is done to the features. It is also experimented with a combination of models: combining baseline and feature based model. The baseline model is done to the phrase based classification task which achieves an accuracy of 62.24% and is 29% more than the chance baseline. The feature based model uses features and achieves an accuracy of 77.86%.

            These combinations achieves an accuracy of 77.90% which outperforms the baseline by 16%. For message based classification task the baseline model comes out with 51% of accuracy which is 18% more than the chance baseline.  The feature based model uses the features with the accuracy of 57.43% . The combination achieves 58.00% of accuracy which outperforms the baseline by 7%.

Twitter is a popular social networking website where users posts and interact with messages known as “tweets”. This serves as a mean for individuals to express their thoughts or feelings about different subjects. Various different parties such as consumers and marketers have done sentiment analysis on such tweets to gather insights into products or to conduct market analysis. Furthermore, with the recent advancements in machine learning algorithms,the accuracy of our sentiment analysis predictions is able to improve

            In this report, we will attempt to conduct sentiment analysis on “tweets” . We attempt to classify the polarity of the tweet where it is either positive or negative. If the tweet has both positive and negative elements, the more dominant sentiment should be picked as the final label. The data provided comes with emoticons, usernames and hashtags which are required to be processed and converted into a standard form. It also need to extract useful features from the text such unigrams and bigrams which is a form of representation of the “tweet”.

**Motivation:**

In the past decade, new forms of communication, such as microblogging and text messaging have emerged and become ubiquitous. While there is no limit to the range of information conveyed by tweets and texts, often these short messages are used to share opinions and sentiments that people have about what is going on in the world

around them.

Tweets and texts are short: a sentence or a headline rather than a document. The language used is very informal,

with creative spelling and punctuation, misspellings, slang, new words, URLs, and genre-specific terminology and abbreviations, such as, RT for "re-tweet" and # hashtags, which are a type of tagging for Twitter messages.

**Future Prospects:**

In the future, one can collect large sentiment related data from multiple social networking sites which may provide better results by overcoming the limitation of the project. In future, one caneven collect data which include videos and images for analyzing the effectiveness of sentimental analysis.

**Requirements:**

1. **Hardware:**

* CPU:- 2x64 bit 2.8 GHz GT/s CPUs
* RAM:- 32 GB
* Sufficient storage
* Internet access

1. **Software:**

* Python

1. **Module Used:**

* String
* Nltk
* Matplotlib
* Pandas
* Numpy
* Collection
* GetOldTweets3