**PROJECT REPORT**

**on**

**Twitter Sentimental Analysis**

**(CSE – 4th Sem)**

**2021**



**Under guidance of Submitted by:**

Mr. Ashwini Kumar Ansh Kumar Garg

University Roll. No.:2015014

Class roll no.:10

Contact no.:9997332302

Gmail: anshgarg110@gamil.com

B-techCSE- ML 4TH Sem

Session: 2020-2021

**Department of computer science and engineering with specialization in ml and ai**

**CERTIFICATE**

### Certified that Ansh Kumar Garg (Roll No. 2015014) has developed mini project on “Twitter Sentimental Analysis” for the CSE IV Semester Mini Project in Graphic Era Deemed to be University, Dehradun. The project carried out by Students is their own work as best of my knowledge.

Date: 08-05-2021

(Mr. Ashwini Kumar)

**Project Resource Person**

(CSE Department)

GEU Dehradun.

**ACKNOWLEDGMENT**

I have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to extend my sincere thanks to all of them.

I am highly indebted to Graphic Era, for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

I would like to express my gratitude towards my parents & our resource person Mr. Ashwini Kumar for their kind co-operation and encouragement which help me in completion of this project.

I would like to express my special gratitude and thanks to industry persons for giving me such attention and time.

My thanks and appreciations also go to my colleague in developing the project and people who have willingly helped me out with their abilities

**Ansh Kumar Garg**

**Roll No.- 2015014**

**CSE-ML & AI**

**Session: 2020-2021**

**GEU, Dehradun**

**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.

**OBJECTIVE:**

To implement an algorithm for classification of text into positive or negative.

## **INTRODUCTION:**

Social network is a rich platform to learn about people’s opinion and sentiment regarding different topics as they can communicate and share their opinion actively on social medias including Facebook and Twitter. There are different opinion-oriented information gathering systems which aim to extract people’s opinion regarding different topics. We will use Twitter to perform sentiment analysis of the written text.

The project would heavily rely on techniques of “Natural Language Processing” in extracting significant patterns and features from the large data set of tweets and on “Machine Learning” techniques for accurately classifying individual unlabelled data samples (tweets) according to whichever pattern model best describes them.

**Technical Approach:**

We will be using python language in implementation of model and Jupyter Notebook which supports machine learning and data science projects.

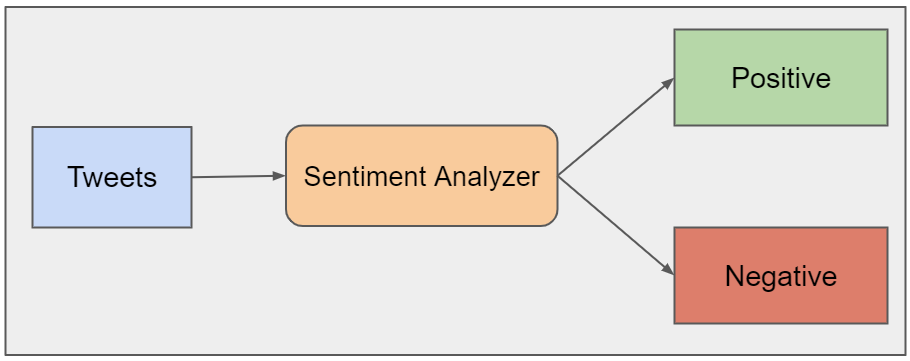
**Important Terms used :**

**Sentimental Analysis :-** [Sentiment analysis](https://monkeylearn.com/sentiment-analysis/) is the automated process of identifying and classifying subjective information in text data. This might be an opinion, a judgment, or a feeling about a particular topic or product feature. The most common type of sentiment analysis is ‘polarity detection’ and involves classifying statements as positive or negative. A polarity sentiment analysis model, for example, automatically tags this tweet as positive. Sentiment analysis uses [Natural Language Processing (NLP)](https://monkeylearn.com/natural-language-processing/) to make sense of human language, and [machine learning](https://monkeylearn.com/machine-learning/) to automatically deliver accurate results. Connect sentiment analysis tools directly to your social platforms, so you can monitor your tweets as and when they come in, 24/7, and get up-to-the-minute insights from your social mentions.

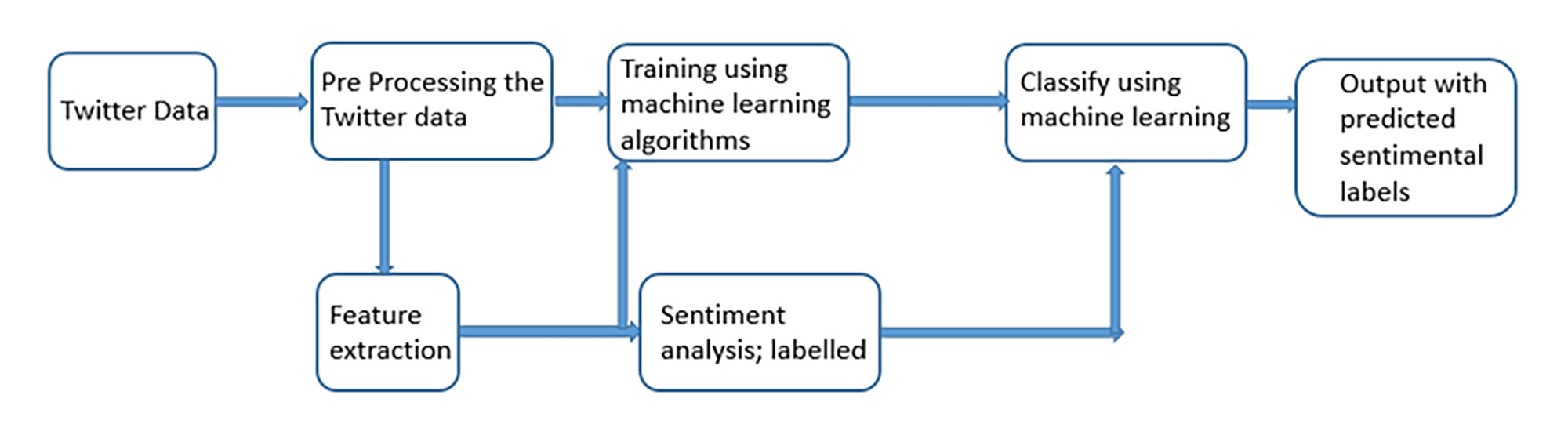
**NLP :-** Natural Language Processing (NLP) is a unique subset of Machine Learning which cares about the real life unstructured data. Although computers cannot identify and process the string inputs, the libraries like NLTK, TextBlob and many others found a way to process string mathematically.

**Diagrammatic Representation of the project:**

* **Identifying positive and negative words:**



* **Process on which the project works :**



**Data collection :**

Taking a training sample of tweets and labels, where label '1' denotes the tweet is negative and label '0' denotes the tweet is not negative, your objective is to predict the labels on the test dataset.

For training the models, we provide a labelled dataset of 31,962 tweets. The dataset is provided in the form of a csv file with each line storing a tweet id, its label and the tweet.

Shape of two data sets used are :

Train: (31962, 3)

Test: (17197,2)

**Libraries used :** Pandas, re, NLTK, Numpy, warnings

**Data Preprocessing :**

* Creating a function and doing preprocessing tasks .
* Converting data to lower case characters
* Then removing all the twitter handles
* Then removal of # tags
* Then removing the re-tweets
* Then finally importing the stop words and removing them from the data .

**Algorithm used**

**Logistic regression** :- Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

Accuracy of Logistic classifier on training set {:.2f} 0.9593425370907809

Accuracy of Logistic classifier on test set {:.2f} 0.9567127746135069

**Naïve bayes :-**

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions

Accuracy of NB classifier on training set: {:.2f} 0.9482763297078847

Accuracy of NB classifier on test set: {:.2f} 0.9449145646867372

**Learning outcome:**

After completion of this project I had learned about the different libraries of python and there functioning. Learned about data pre-processing techniques logistic regression, naïve bayes which are of the important algorithms of the machine learning.

**Motivation :**

I have chosen to work with twitter since we feel it is a better approximation of public sentiment as opposed to conventional internet articles and web blogs. The reason is that the amount of relevant data is much larger for twitter, as compared to traditional blogging sites. Moreover the response on twitter is more prompt and also more general.

**CONCLUSION AND FUTURE RECOMMENDATIONS :**

### The task of sentiment analysis, is still in the developing stage and far from complete. So proposed a couple of ideas which we feel are worth exploring in the future and may result in further improved performance. Right now I had worked with only the very simplest models; we can improve those models by adding extra information like closeness of the word with a negation word.

### Using an different data visualization process.

### Using different machine learning algorithm and comparing them.