A Report on

SENTIMENTAL ANALYSIS IN TWITER USING PYTHON

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

(INTERNET OF THINGS)

Submitted by

TEJA VARMA-20BTRCD05

JAINI KARTHIKEYA-20BTRCD011

SRI HARSHA-20BTRCD019

NANDA KISHORE REDDY-20BTRCD035

HEMAA CHANDHAN-20BTRCD039

KAUSTUBH NAIK-20BTRCD051

Under the guidance of

R. VIJAY KUMAR SIR

Faculty of Engineering & Technology

Jain (Deemed-To-Be University)

Department of Computer Science & Engineering

Jain Global Campus, KanakapuraTaluk - 562112

Ramanagara District, Karnataka, India

2021-20

Table of contents:

1.Abstract:

The purpose of this project is to create a functional classifier that can accurately and automatically classify the sentiment of an unknown tweet stream. Sentiment analysis, also known as opinion mining, is the computational study of people's written opinions, sentiments, attitudes, and emotions. Sentiment analysis is a method that employs Natural Language Processing (NLP) to extract, convert, and interpret opinions from text and categorise them as positive, negative, or natural sentiment.

Keyword:

1. Sentimental Analysis

2. Machine Learning

3. Natural Language Processing

4. Python

2.Introduction:

1. The Internet has been extremely beneficial in allowing people in today's world to express themselves globally.Individuals all over the world rely on this client, which produces a lot of content.
2. However, it is not humanly possible for a single person to sit and read every single review available.It would simply be a waste of time.
3. As a result, this process can be automated to make it more efficient.
4. Machine Learning (ML) is crucial in this case.The ML process of Sentiment Analysis (SA) assists the system in understanding the sentiment of a specific statement made.
5. The system is built with several ML algorithms that can understand the nature of sentiment or a set of sentiments.
6. The paper proposes a system for extracting data from Twitter and analysing it.

The paper proposes a system for extracting Twitter data and performing analysis on it. Machine Learning (ML) assists the system in determining the sentiment of a given statement. The system is built with several machine learning algorithms that can understand the nature of sentiment or a set of sentiments.

3.Literature survey/review:

S.NO	TITLE OF THE PAPER	TECHNIQUE
1)	Supervised Machine Learning approach for Urdu sentiment analysis in multiple domain.	Support Vector Machine Algorithm
2)	E-Commerce product review sentiment classification.	Naive Baye's
3)	Data Extraction and Sentiment analysis using stack of Deep Learning algorithms.	Bayesian Logistic Regression
4)	Multi task Learning model based on recurrent convolutional neural network.	Maximum Entropy

4.Methodology:

The system plans to perform sentiment analysis on tweets from the Twitter dataset. Various algorithms were used and tested against the available dataset, and the best algorithm was chosen. gives an idea of how the sentiment analysis will be performed The dataset will be pre-processed using the techniques listed below after it has been cleaned and divided (isolated) into preparing (training) and testing datasets. To reduce the size of the dataset, features will be extracted. The next step is to develop a model that will be used by the classifier to categorise tweets as positive or negative. Again, the classifier will be given real-time tweets to test the real-time data. The proposed system is not interactive

4.1. Data Cleaning

The information gathered was not in the proper format. Information cleaning is the process of ensuring that information is correct, predictable, and usable. Typically, data sets must be cleansed because they contain a large amount of noisy or unwanted data known as outliers The presence of such outliers may result in incorrect results. Data cleaning ensures that such data is removed and improved, resulting in a much more reliable and stable dataset.

Data cleaning can be done in given ways:

• Monitors error:. The entry point or source of errors should be tracked and monitored constantly. This will help in correcting the corrupted data

• Process standardization: The point of entry should be standardized. By standardizing the data process, the risk of duplication reduces.

• Accuracy validation: Data should be validated once the existing database is cleaned. Studying and using various data tools that can help in cleaning the datasets is very important

• Avoid data duplication: The identification of duplicate data is a very mandatory process. Several AI tools help in identifying duplicates in large corpora of data.

The above-mentioned steps are a few of the many ways to clean datasets. Making use of these methods will end up giving good, usable, and reliable datasets.

4.2. Data Pre-processing

Data pre-processing comes after data cleaning. It is a significant step forward in machine learning. It is the process of transforming or encoding data so that it can be understood by a machine. In other words, the algorithms can easily interpret the dataset's features. The feature is a measurable property of the observed entity. Height, age, and gender, for example, are all characteristics of a person. A twitter stream will extract every related tweet from Twitter in an unstructured format. Before applying any classifier to these unstructured tweets, they must be pre-processed. Tokenization and cleaning will be performed on the tweets ahead of time. Initially, all HTML content from tweets is removed.To process, all tweets must be in the same case; thus, it will change to lower case, and each word is split based on space. Following that, collect all stop words and structure them as a single set before removing them. Finally, return a string of significant words [11]. Thus, before extracting the features, a pre-processing step is performed to filter out slang words and misspellings[12].

4.3. Classifiers to Be Used

Naive Bayes: The naive Bayes is a supervised machine learning algorithm that returns probability values as the output. Naive Bayes classifier is very

useful in solving high-dimensional problems. It assumes the probabilities of the different events that are completely independent.

Logistic regression: Logistic regression predicts a binary outcome, i.e., (Y/N) or (1/0) or (True/False). It also works as a special case of linear regression. It produces an S-shaped curve better known as a sigmoid. Ittakes real values between 0 and 1.Basically, logistic regression has a binary target variable. There can be categories of target variables that can be predicted by it. The logistic classifier uses a crossvalidation estimor.

Support vector machine: It is a non-probabilistic model that utilizes a portrayal of text models as focuses in a multidimensional space. These examples are mapped with the goal that the instances of the diverse categories (sentiments) have a place with particular areas of that space. Later, the new messages are mapped onto that equivalent space and are predicted to have a place with a classification dependent on which category they fall into. In the SVM algorithm, the fundamental goal is to boost the edge between information points and the hyperplane. The loss function that helps with this is called a hinge loss

5.Conclusions and Future Work: -

The work in this paper is done to classify a relatively huge corpus of twitter data into two groups of sentiments, positive and negative, respectively. Higher accuracy is achieved by using sentiment features instead of conventional text classification. This feature can be used by various establishments, business organizations, entrepreneurship, etc., to evaluate their products and get a deeper insight into what people say about their products and services. Future work includes working not only in the English language but in other regional languages too. Also, it will include analysis of complex emotions like sarcasm and generate a hybrid classifier to get the best accuray.

6.References:-

[1] A. Feizollah, S. Ainin, N. B. Anuar, N. A. B. Abdullah and M. Hazim, "Halal Products onTwitter: Data Extraction and Sentiment Analysis Using Stack of Deep Learning Algorithms,"IEEE Access, vol. 7, pp. 83354-83362, 2019.

[2] Neelam Mukhtar, Mohammad AbidKhan, and NadiaChiragh, "Lexicon-based approach out performs Supervised Machine Learning approach for Urdu Sentiment Analysis in multiple domains", Telematics and Informatics, vol. 35, no. 8, pp. 2173-2183, December 2018.

[3] AbdallahYousif, ZhendongNiu, JamesChambua, and Zahid YounasKhan, "Multi-task learningmodel based on recurrent convolutional neural networks for citation sentiment and purposeclassification", Neurocomputing, vol. 335, pp. 195-205, 28 March 2019.

[4] FengXu, ZhenchunPan, and RuiXia, "E-commerce product review sentiment classificationbased on a naïve Bayes continuous learning framework", Information Processing &Management, Available online 13 February 2020.

[5] Hyun-jungPark, MinchaeSong, and Kyung-ShikShin, "Deep learning models and datasets foraspect term sentiment classification: Implementing holistic recurrent attention on target-dependent memories", Knowledge-Based Systems, vol. 187, January 2020.

[6] S. E. Saad and J. Yang, "Twitter Sentiment Analysis Based on Ordinal Regression," IEEEAccess, vol. 7, pp. 163677-163685, 2019.

[7] ParamitaRay, and AmlanChakrabarti, "A Mixed approach of Deep Learning method and Rule-Based method to improve Aspect Level Sentiment Analysis", Applied Computing andInformatics, Available online 4 March 2019.

[8] ZiyuanZhao, HuiyingZhu, ZehaoXue, ZhaoLiu, JingTian, Matthew Chin HengChua, andMaofuLiu, "An image-text consistency driven multimodal sentiment analysis approach forsocial media", Information Processing & Management, vol. 56, no. 6, November 2019