International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 05 | May 2021 www.irjet.net p-ISSN: 2395-0072

Twitter Sentiment Analysis

Ramachandiran [1], Arunnkumar [2], Balachander [3]

[1] *Associate Professor, Department of Information Technology, Sri Manakula Vinayagar EngineeringCollege, Puducherry, India*
[2,3] *Student, Department of Information Technology, Sri Manakula Vinayagar EngineeringCollege, Puducherry, India*

Abstract - *Twitter is one of the most used applications by the people to express their opinion and show there sentiments towards different occasions. In the era of accelerating social media users, Twitter has a large number of regular users who post their views in the form of tweets. This paper proposes a method for collecting feelings from tweets as well as a method for categorising tweets as positive, negative, or neutral. Any business that is listed or tagged in a tweet will benefit from this strategy in a number of ways. Since most tweets are in an unstructured format, they must first be translated into a structured format. Tweets are resolved in this paper using a pre-processing phase, and tweets are accessed using libraries that use the Twitter API. The datasets must be trained using algorithms in such a way that they are capable of checking tweets and extracting the requisite sentiments from the feed. The goal of this research is to give an model to this fascinating problem and to present a framework which will perform sentiment analysis associating XGBOOST and Natural Language Processing Classification Techniques*

Key Words: Sentiment Analysis, Twitter, XGBOOST, opinion mining.

1.INTRODUCTION

This research, which analyses tweet sentiments, falls under the categories of "Pattern Classification" and "Data Mining." Both of these concepts are closely linked and interconnected, and they can be formally identified as the method of identifying "useful" trends in large amounts of data. The project will largely depend on "Natural Language Processing" techniques for extracting significant patterns and features from a big corpus of tweets, as well as "Machine Learning" techniques for precisely identifying different unlabeled data samples (tweets) with whatever pattern model best describes them.

There are two types of features that can be used for modelling patterns and classification: formal language-based features and informal blogging-based features. Prior intention polarity of individual terms and phrases, as well as parts of speech tagging of the sentence, are examples of language-based features. Prior sentiment polarity defines the inherent propensity of certain terms and phrases to convey similar and related feelings in general. The word "excellent," for example, has a positive social connotation, while "evil" has a strong negative connotation. When a word with a positive connotation appears in a sentence, the whole sentence is likely to express a positive emotion. On the other hand, Parts of Speech tagging is a syntactical solution to the issue. It means determining which part of speech each individual word in a sentence belongs to, such as noun, pronoun, adverb, adjective, verb, interjection, and so on.

Patterns can be derived by studying the frequency distribution of these parts of speech in a

specific class of labelled tweets (either separately or in combination with another parts of speech). Twitter-based features are more casual and relate to how people express themselves on online media networks and compact their feelings into the limited 140-character space provided by Twitter. Twitter hashtags, retweets, term capitalization, word lengthening, question marks, URL presence in tweets, exclamation marks, internet emoticons, and internet shorthand/slangs are only a few examples

1.1 Problem Statement

The project's aim is to do a sentiment study on a certain product or service. Sentiment will be categorised as positive, neutral, or negative; there will be no in-between categories. For example, if an emotion is positive or highly positive, both will be classified as "positive."A study on how the product or service is viewed by the target audience would be the outcome of this inquiry. This project will create a variety of tools and computer programmes, but it is important to remember that these are not the expected outputs. It is worth repeating that the project's output is the study of the target audience. That being said a system will have to be designed in order to perform this analysis.

This method will collect data from Twitter, cleanse it, and then categorise it. The classified data would then be subjected to review. The most important criterion for this study is that it provides the client with a reasonably high degree of precision at a low cost. The aim of this project is to construct a practical classifier that can reliably and instantly characterise an unknown tweet stream's sentiment.

2. LITERATURE SURVEY

Many theories have been offered to explain why humans feel the way they do. Despite the fact that the literature includes a wide range of ideas, this study will concentrate on five main themes that appear across the literature analyzed in Table 1.

Table -1: Literature Survey

Algorithm Used	Dataset	Description
LIWC text analysis software	104,003 tweets	[1] Twitter is being used as a platform for political debate, and it's being investigated if online postings on Twitter accurately reflect offline political attitude. Our findings indicate that Twitter is widely utilised for political discussion. We discovered that the quantity of communications referencing a political party accurately matches the election outcome.
NB enhanced SVM	12002 tweets	[2] In both trials, the unified model integrating syntactic context of words and emotion information of sentences performed best.
Naive Bayes	6,408 tweets	[3] The algorithm employs a basic rule that looks for polarity terms in the tweets/texts being analysed. When a polarity lexicon and multiword are given to the classifier.
NLP & SVM	2 billion tweets	[4] The SVM was evaluated using measures like the Area Under the Curve (AUC) and the Receiver Operating Characteristic (ROC) curve. The ROC curve was created using the

		mean values of the 1000 repetitions. The average prediction accuracy over 1000 iterations was 0.74, and the average AUC value was 0.82.
Hadoop Classification	1000 positive and 1000 negative tweets	[5] They ran their algorithm against the Cornell dataset, which yielded an average accuracy of 80.85%. The software was able to categorise several subsets of the Amazon movie review dataset with equal accuracy without modifying the Hadoop code. The size of the dataset in their experiment ranges from 1,000 to one 100,000 reviews in each class to evaluate the adaptability of the Naive Bayes classifier.
Unigram Naive Bayes	10000 tweets	[6]To gather Twitter data, use the Twitter API. Their training data is divided into three groups (camera, movie , mobile). The data is divided into three categories: favourable, negative, and non-opinions. Opinion-based tweets were censored. The Naive Bayes simplified independence assumption was used with the Unigram Naive Bayes model. They also used the Mutual Information and Chi Square feature extraction methods to remove unnecessary features. Finally, the tweet's orientation is anticipated. i.e. if it is good or bad.
KNN	18000 tweets	[7] This paper presented a method for classifying sentiment types in tweets utilising punctuation, single words, n-grams, and patterns as distinct feature types, which were then merged into a single feature vector for sentiment classification. They constructed a feature vector for each sample in the training and test sets and used the K-Nearest Neighbor method to assign emotion labels.
SVM	50000 tweets	[8] This paper suggested a method for sentiment analysis for Twitter data utilising remote supervision, with tweets containing emoticons serving as noisy labels as training data. Nave Bayes, MaxEnt, and Support Vector Machines are used to create models (SVM). Unigrams, bigrams, and POS were part of their feature space. SVM outperformed other models, and unigrams were more useful as features, they found.

3.PROPOSED SYSTEM

Our objective is to use Twitter data to do sentiment analysis. We'll combine a number of machine learning classifiers to create a classifier. We'll keep going till our classifier is ready and well-trained.

Step 1: Using Python's Textblob module, we'll start by streaming tweets into our build classifier.

Step 2: We next pre-process these tweets to make them suitable for mining and feature extraction.

Step 3: After pre-processing, we give this information to our trained classifier, which categorises them as positive or negative depending on the outcomes of the training.

Step -4 Since Twitter is our data source, we'll use it to analyse it. The tweets from Twitter will be streamed into our database. We'll utilise the Twitter application for this

.4.ARCHITECTURE

This article aims to provide a high-level overview of the data mining, text classification, and machine learning algorithms used in this project. The fundamental organization of a system, as expressed in its elements, their interactions with one another and with the environment, and the design

and evolution principles. These representations begin with a high-level, general definition of a functional organisation and gradually become more detailed and concrete. The figure 4.1 displays a block diagram of the method.

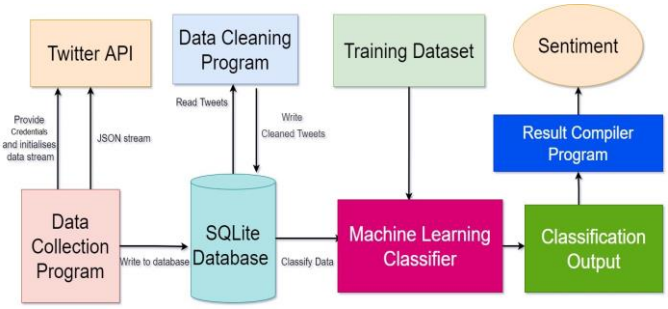

Figure 4.1

5. IMPLEMENTATION

The implementation phase includes the construction of comprehensive data model listed below:

1.Data Gathering

2.Data Preprocessing

3.Tokenisation

4.FeatureExtraction

5.Data Training

5.1.DATA GATHERING

Twitter provides a set of streaming APIs that provide developers low-latency access to Twitter data flows. For data collection, the public streams API was utilised; it was determined that this was the ideal means of obtaining information for data mining purposes because it gave access to a worldwide stream of twitter data that could be filtered as needed.

5.2.DATA PREPROCESSING

Preprocessing a Twitter dataset entails eliminating all sorts of unnecessary data, such as emoticons, special characters, and blank spaces. It may also entail making format changes, deleting duplicate tweets, or tweets with less than three characters.

5.3 TOKENIZATION

It is the process of separating a continuous stream of text into words, symbols, and other significant pieces known as "tokens." Whitespace and/or punctuation characters can be used to separate tokens. It's done this way so that tokens may be seen as distinct components of a tweet. Emoticons and abbreviations (e.g., OMG, WTF, BRB) are recognised as separate tokens during the tokenization process.

5.4 FEATURE EXTRACTION

Feature extraction is a dimensionality reduction method that reduces a large collection of raw data into smaller groupings for processing. The high number of variables in these large data sets necessitates a lot of computer resources to process.

5.5 DATA TRAINING

If the training procedure is followed correctly, the machine learning algorithm should be able to generalise the training data and map new data it has never seen before. Training data must include a class label; this may be accomplished by manually assigning a class to each tweet, but this is a time-consuming procedure, and because Twitter has tight regulations about the sharing of its data, finding trustworthy hand-annotated twitter datasets has proven challenging.

6. CONCLUSIONS

Our solution provides the best technique to predict the sentiment analysis with an efficiency of 95%. Machine-Learning Algorithm predicts and identifies the exact feedback/review of the user on a specific product.In future, this model focuses to enhance the accuracy of prediction. The various approaches to sentiment analysis, primarily Machine Learning and

Cognitive approaches, are discussed in depth in this study. It gives a comprehensive overview of the numerous applications and challenges that Sentiment Analysis can present, making it a difficult task.

REFERENCES

[1] Efthymios Kouloumpis and Johanna Moore,IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2018

[2] S. Batra and D. Rao, "Entity Based Sentiment Analysis on Twitter", Stanford University,2019

[3] Saif M.Mohammad and Xiaodan zhu ,Sentiment Analysis on of social media texts:,2018

[4]Ekaterina kochmar,University of Cambridge,at the Cambridge coding Academy Data Science.2017

[5] Manju Venugopalan and Deepa Gupta ,Exploring Sentiment Analysis on Twitter Data, IEEE 2019

[6] Brett Duncan and Yanqing Zhang, Neural Networks for Sentiment Analysis on Twitter.2017

[7] Afroze Ibrahim Baqapuri, Twitter Sentiment Analysis: The Good the Bad and the OMG!, Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.2020

[8] Suman, D.R, & Wenjun, Z., "Social Multimedia Signals: A Signal Processing Approach to Social Network Phenonmena",