A Survey on Twitter Sentiment Analysis

Ramachandran [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 proposed to explain what sentiment of humans. Although the literature covers a wide variety of such theories, this review will focus on five major themes which emerge repeatedly throughout the literature reviewed in Table 1.

Table -1: Literature Survey

Algorithm Used	Dataset	Description
LIWC text analysis software	104,003 tweets	[1]Twitter is used as a forum for political deliberation and whether online messages on Twitter validly mirror offline political sentiment. Our results show that Twitter is indeed used extensively for political deliberation. We find that the mere number of messages mentioning a party reflects the election result.
NB enhanced SVM	12002 tweets	[2]The unified model combining syntactic context of words and sentiment information of sentences yields the best performance in both experiments
Naive Bayes	6,408 tweets	[3] The system makes use of fundamental rule that search's for polarity words within the analysed tweets/texts. When the classifier is provided with a polarity lexicon and multiword.
NLP & SVM	2 billion tweets	[4]Evaluation of the SVM was done using parameters such as the Area under the Curve (AUC) value, and the Receiver operating characteristic (ROC) curve. The ROC curve using the mean values of the 1000 iterations was drawn. The prediction accuracy on average over the 1000 iterations was evaluated to 0.74, and the mean AUC value is 0.82
Hadoop Classification	1000 positive and 1000 negative tweets	[5]They tested their code on Cornell dataset and resulted in an 80.85% average accuracy. Without changing the Hadoop code, the program was able to classify different subsets of Amazon movie review dataset with comparable accuracy. To test the scalability of Naive Bayes classifier, the size of the dataset in their experiment varies from one thousand to one million reviews in each class.
Unigram Naive Bayes	10000 tweets	[6] Twitter API to collect twitter data. Their training data falls in three different categories (camera, movie , mobile). The data is labeled as positive, negative and non-opinions. Tweets containing opinions were filtered. Unigram Naive Bayes model was implemented and the Naive Bayes simplifying independence assumption was employed. They also eliminated useless features by using the Mutual Information and Chi square feature extraction method. Finally , the orientation of an tweet is predicted. i.e. positive or negative.
KNN	18000 tweets	[7]This proposed a approach to utilize Twitter user- defined hastags in tweets as a classification of sentiment type using punctuation, single words, n-grams and patterns as different feature types, which are then combined into a single feature vector for sentiment classification. They made use of K-Nearest Neighbor strategy to assign sentiment labels by constructing a feature vector for each example in the training and test set.
SVM	50000 tweets	[8]This proposed a solution for sentiment analysis for twitter data by using distant supervision, in which their training data consisted of tweets with emoticons which served as noisy labels.They build models using Naïve Bayes, MaxEnt and Support Vector Machines (SVM). Their feature space consisted of unigrams, bigrams and POS. They concluded that SVM outperformed other models and that unigram were more effective as features

3.PROPOSED SYSTEM

Our aim is to perform sentiment analysis using data from Twitter. We're going to make a classifier out of a variety of machine learning classifiers. We'll proceed with the steps until our classifier is ready and educated.

Step-1 First we are going to stream tweets in our build classifier with the help of Textblob library in python

Step-2 Then we pre-process these tweets, so that they can be fit for mining and feature extraction.

Step-3 After pre-processing we pass this data in our trained classifier, which then classify them into positive or negative class based on trained results.

Step -4 Since, Twitter is our source of data for analysis. We are going to stream the tweets from twitter in our database. For this we are going to use Twitter Application

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.

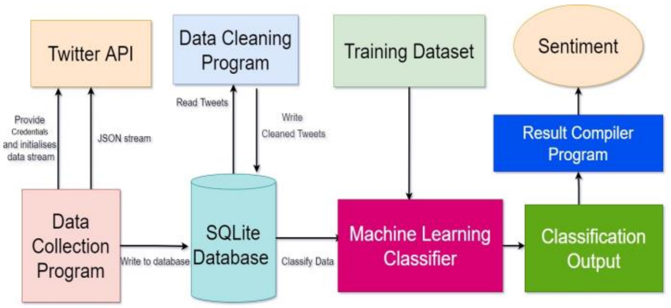

Figure 4.1

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 enables developers with a collection of streaming APIs that provide low-latency access to Twitter data flows. The public streams API was used for data collection; it was discovered that this was the best method of gathering information for data mining purposes because it provided access to a global stream of twitter data that could be filtered as needed.

5.2.DATA PREPROCESSING

Preprocessing a Twitter dataset involves a series of tasks like removing all types of irrelevant information like emojis, special characters, and extra blank spaces. It can also involve making format improvements, delete duplicate tweets, or tweets that are shorter than three characters

5.3 TOKENIZATION

It is the process of breaking a stream of text up into words symbols and other meaningful elements called “tokens”. Tokens can be separated by whitespace characters and/or punctuation characters. It is done so that we can look at tokens as individual components that make up a tweet. Emoticons and abbreviations (e.g., OMG, WTF, BRB) are identified as part of the tokenization process and treated as individual tokens

5.4 FEATURE EXTRACTION

Feature extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing. A characteristic of these large data sets is a large number of variables that require a lot of computing resources to process.

5.5 DATA TRAINING

The training process is implemented correctly the machine learning algorithm should be able to generalize the training data so that it can correctly map new data that it has never seen before. Training data must contain a class label, this can be achieved through manually assigning each tweet with a class but this is a tedious process and as twitter enforces strict rules on the distribution of its data it has proved difficult to source reliable hand annotated twitter datasets.

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”,

[9] Sang-Hyun Cho and Hang-Bong Kang, “Text Sentiment Classification for SNS-based Marketing Using Domain Sentiment Dictionary”,