Combining Lexicon based and Machine Learning based Methods for Twitter Sentiment Analysis

Siddhi Adhiya[1], Samruddhi Shirbhate[2], Anurag Kadu[3], Aditya Kalambe[4], Tejal Mohod[5], Atharva Belge[6]

[1,2]Department of Information Technology, Sipna College of Engineering and Technology, Amravati, Maharashtra, India

[3]Department of Information Technology, Government College of Engineering Amravati, Maharashtra, India

[4,5,6] Department of Computer Science, Sipna College of Engineering and Technology, Amravati, Maharashtra, India

---

Abstract - *Twitter is one of the most used social media. People often use Twitter to express their thoughts and feelings, users give their opinions about several topics on this social platform. There are almost 330 million people over the globe who are active on twitter and share their opinion widely over this application. There are lots of tweets which express human sentiments on twitter. Opinions tweeted by people generally depict the sentiments and this makes it easy for industries and business companies to analyze and study the opinion on their product. Thus Sentiment Analysis works wonders for business companies, industries and organizations to acquire different opinions from the users. This study focuses on the techniques of combining lexicon based and machine learning based methods for twitter sentiment analysis. This includes techniques like NLP (Natural Language Processing) which can understand texts and words the same way human beings can.*

Key Words: Lexicon, Random Forest, TextBlob, Vader Sentiment Analyzer, Support Vector Machine.

1. INTRODUCTION

Nowadays, Twitter is one of the most actively used social media platforms. A large number of people are willing to post their thoughts and reactions on Twitter. Tweets are the best way to analyze people's real opinions regarding a specific product, organization, service, movie, individual, political events, topics, and their attributes regarding various products and sectors. With the help of sentiment analysis. We can collect feedback on any product from the users and can improve the product if needed. Sentiment analysis especially works on two terms positive or negative, we can collect opinions from the users positively or negatively, Example: If any industry launches their phone then there will be some negative and positive reactions by the users. This data will be collected accordingly and will further be analyzed and the result will be provided to the company. Through this analysis, companies can modify their product and provide better products in future. Hence companies can gain a good profit and satisfy their customer's needs. Twitter is a kind of micro-blogging social networking site and billions of users use it to give their opinion or reactions on a specific topic. A Review of the product can be estimated through sentiment analysis.

Before purchasing a product, people often search for reviews of the product online. This can help the users to decide whether the product is worth spending money on. These reviews usually contain expressions that carry all types of opinions, such as "great" (positive valence) or "terrible" (negative valence), leaving readers with a positive or negative impression. This approach of using opinion words (the lexicon) to determine opinion orientations is called the lexicon-based approach to sentiment analysis. Lexicon sentiment analysis is based on calculations of polarity scores given to positive and negative words in a document.

2. DATASET DESCRIPTION

To effectively perform sentiment analysis on the tweets in the datasets, it is necessary to first clean the data. This involves removing any irrelevant symbols or characters that may have been included in the tweets and could potentially interfere with the accuracy of the analysis. This cleaning process can also involve converting all text to lowercase to ensure consistency and eliminating any stop words that may not add much value to the analysis. Once the data has been cleaned, it is ready to be used in various machine learning algorithms. The goal of sentiment analysis is to classify the tweets as positive, negative, or neutral. To achieve this, we need to train a machine learning model using the cleaned training dataset (train.csv) and evaluate its accuracy using the test dataset (test.csv). sentiment analysis, including Naive Bayes, Support Vector Machines (SVM), and Long Short-Term Memory (LSTM) networks. The accuracy of the model can be measured by calculating metrics such as precision, recall, and F1 score. These metrics will give us an idea of how well the model is able to accurately classify the

tweets in the test dataset. Once we have obtained a well-performing model, we can use it to classify the sentiment of any new tweets that come in. Sentiment analysis has numerous applications, including customer feedback analysis, brand reputation management, and social media monitoring. In this Research, we have combined Lexicon Sentiment Analysis and Machine Learning Techniques to analyze the reviews. To estimate the overall opinion of the user about a particular product or service and that can be implemented in different sectors to improve their product or service quality.

3. Natural Language Processing (NLP)

Natural Language Processing (NLP) is one of the branches of Artificial Intelligence. It forms a link for interactions with Humans through natural languages. This kind of Intelligent System requires computational and linguistic technology to build it, and the system processes natural language like humans. The development cycle in NLP is shown in Fig. 1. The development cycle of NLP begins with collecting the dataset. The Text collected in the dataset is known corpus. Further the data is analyzed to process the text. The process then carries on till the initial processing stage. This stage focuses on the cleaning and selecting appropriate text. Then it will get attributes from unstructured texts through the computational techniques such as Machine Learning. Now the dataset is in the structured format and it is ready for the Sentiment Analysis Test.

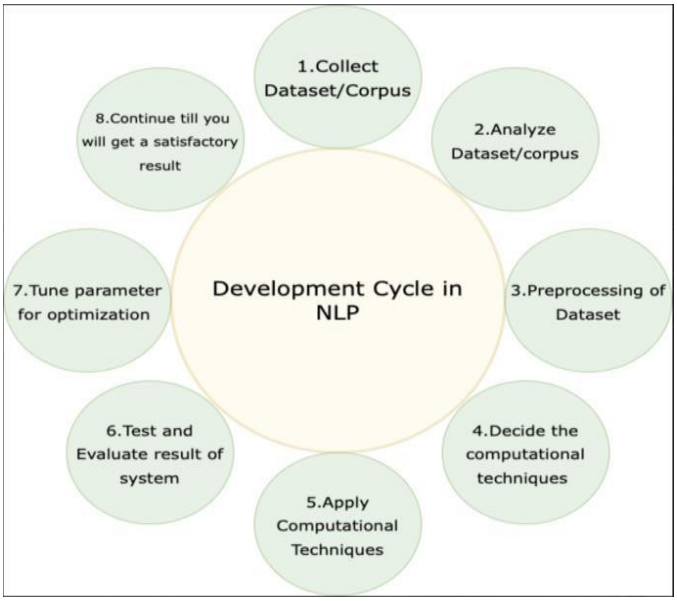

Fig. 1. Development Cycle of NLP

4. Sentiment Analysis

The word Sentiment Identification is an Important job in many Applications of sentiment Analysis and opinion mining, such as mining tweets, discovery of opinion holders, and classification of tweets. Sentiment can be further categorized into positive, negative and neutral words. Sentiment Analysis comes under the study of NLP and represents people's beliefs and point of view on particular topics or an event.

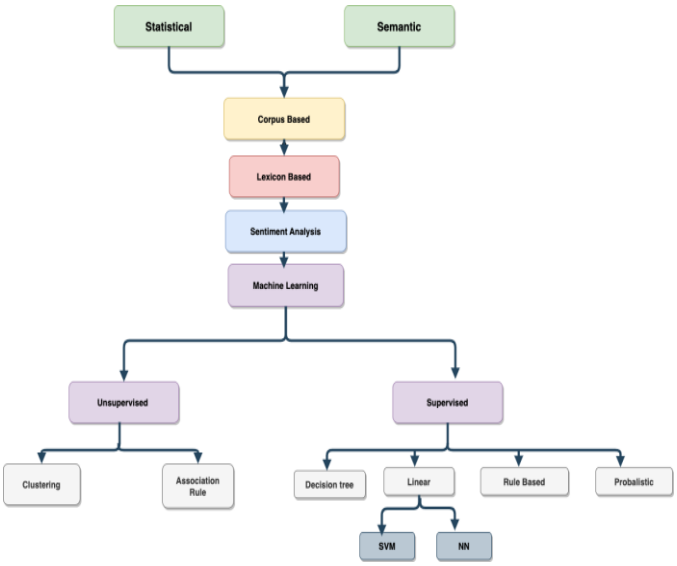

Fig 2. Sentiment Analysis Methods

4.1 Lexicon Based Sentiment Analysis

This approach uses sentiment Lexicon to interpret the polarity (positive, negative or neutral) of text content. This approach can be easily implemented then algorithm-based approach. The only drawback here is the involvement of humans which can cause errors. Lexicon which has been prepared using a dictionary based on this dictionary only contains adjectives that are collected and then have labels. Following steps are involved in Lexicon-based analysis tweets retrieval, then those tweets are pre-processed by deleting repeated tweets, removing URLs, special characters, stop words, numbers, punctuation marks, stemming and tokenization. The more prominent the data set the more noteworthy the text will be. Identification lexicon-based approach can be divided into two categories dictionary-based corpus based. Dictionary based on word net and corpus based on corpus data which can further be divided into statistical and Semantic approaches.

4.1.1 TextBlob

Text blob is a python library that is used for text analysis, mining, processing, and modules for the python community. It supports the latest as well as version 2.6 of Python. It utilizes NLTK corpora. These commands will install textblob and download NLTK corpora.

Installation:

step1: $ pip install -U textblob

step2: $ python -m textblob.download_corpora

First the data is input, then reviews are splitted into sentences. A common way of determining polarity for an entire dataset is to count the number of positive and negative sentences/reviews and decide whether the response is positive and negative based on the total number of reviews.

4.1.2 VADER

Vader is a popular tool used for sentiment analysis, which is the process of determining the sentiment expressed in a given text. It is based on a lexicon and rule-based approach, which means that it uses a dictionary of words and their corresponding sentiment scores to analyze the sentiment of a given text. The sentiment scores range from negative to positive, and the score of each word is used to determine the overall sentiment of the text. Vader takes into account the context of words, so it is able to understand the sentiment expressed even in complex sentences. For example, if the text says "I did not love the movie", Vader will be able to determine that the sentiment expressed is negative, despite the word *"love"* typically conveying a positive sentiment. Vader also takes into account the emphasis of capitalization and punctuation, such as exclamation marks or question marks, to determine the sentiment expressed. For example, if the text says "I ENJOY the movie!" Vader will be able to determine that the sentiment expressed is positive, based on the use of all caps and an exclamation mark. Overall, Vader is a fast, efficient, and accurate tool for sentiment analysis, making it a popular choice for sentiment analysis in various fields, such as social media analysis, customer feedback analysis, and opinion mining.

Installation:

C:\Users\Admin>pip install vaderSentiment

Advantages of VADER:

1. It does not suffer from speed performance trade off.

2. Fast enough to use online.

3. It also allows emojis for the classification of sentiments in the sentences.

4. Works accurately on social media type text.

5. It constructed from generalizable human-curated gold standard lexicon, valence based.

6. Training data is not essential for VADER.

5. Machine Learning Based Approaches

In the Machine Learning approach, the Data is collected from the Twitter Module. This process of collecting data is performed one time as we don't need this process after the data is collected. After this the Data is sent to the Preprocessing Module after undergoing through certain algorithms the analysis of sentiment is performed and the required result is formed.

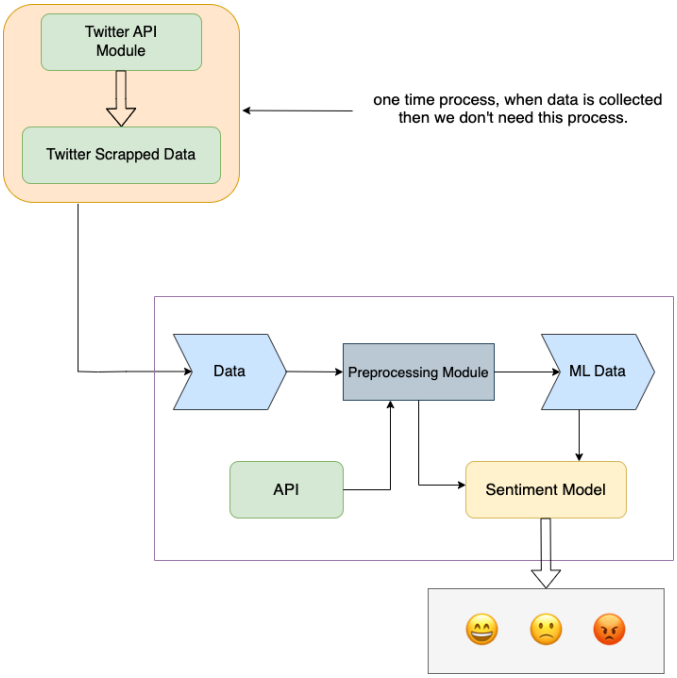

Fig 3. Process for sentiment classification

The machine learning algorithm analyzed the input text of the customer reviews to determine the sentiment expressed in the review. This was achieved by using natural language processing techniques to identify and extract important features from the reviews. The algorithm then used these features to train a machine learning model to accurately predict the sentiment expressed in the reviews. The model was trained using a supervised learning approach, where it was given labeled data (positive or negative sentiment) to learn from. The algorithm then used this trained model to make predictions on the sentiment of new reviews. The accuracy of the predictions was evaluated using various evaluation metrics such as precision, recall, and F1 score.

The Machine Learning Approach was able to provide valuable insights into the customer reviews of Amazon's Alexa products. It allowed for the analysis of the overall sentiment towards the products and helped determine the strengths and weaknesses of the products. By analyzing the input text of the customer reviews, the machine

learning algorithm was able to identify important features that had a significant impact on the sentiment expressed in the reviews. This information was useful in developing future marketing strategies and in making improvements to the products to better meet customer needs and expectations.

5.1 Random Forest Algorithm

Random Forest, which were formally proposed in 2001 by Leo Breiman and Adèle Cutler, are part of the automatic learning techniques. This algorithm consists of the concepts of "bagging". Random Forest is a classification based on combining tree structures with training on the sample data they have. It is a supervised Machine Learning Algorithm that depends on ensembling the learning and data. Random Forest works in two phases. The first is to create a random forest by combining N decision trees and the second is to make predictions for each tree created in the first phase.

5.2 Support Vector Machine

SVM is a supervised machine learning algorithm that is used for both classification and regression. SVM chooses extreme points/vectors that help in creating the hyperlane. The points closest to the hyperlane and which affects the position of the hyperlane are termed as support vectors. The data is collected and fitted into the training set for the classification and then the result is calculated.

5.3 Decision Tree Classification

Decision Tree Classification is a type of supervised machine learning algorithm that is used for classification problems. It works by constructing a tree-like model of decisions and their possible consequences. Each internal node of the tree represents a test on an input feature, and each leaf node represents a class label. The tree is constructed by recursively splitting the training data into subsets based on the feature test that results in the largest reduction in impurity, with the goal of resulting in pure subsets (i.e., subsets where all instances belong to the same class). When a new instance is to be classified, it traverses the tree by testing the feature values of the instance against the feature test at each internal node, and finally reaches a leaf node that determines the class label of the instance. Decision tree classifiers can handle both continuous and categorical features, and are simple and easy to interpret, but may be prone to overfitting if the tree is not pruned.

5.4 Logistics Regression

Logistic Regression is a type of supervised machine learning algorithm that is used for classification problems. It is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is modeled using a logistic function, which is a S-shaped curve that maps any real-valued number to a value between 0 and 1, and is used to represent the probability of a binary outcome. In Logistic Regression, the input features are combined linearly using coefficients (also known as weights) to predict the log-odds of the binary outcome. The coefficients are learned from the training data using maximum likelihood estimation. The resulting model can then be used to make predictions on new data instances. Logistic Regression is simple, efficient, and easy to interpret, but may not perform well when the relationship between the features and the outcome is non linear or when there is high complexity in the data.

6. Accuracy

Random forest classification is an ensemble machine learning algorithm that combines multiple decision trees to make a prediction. It creates several decision trees and aggregates their predictions, reducing the variance and improving the accuracy of the model. This makes it well suited for large datasets and complex problems. In the analysis of Twitter tweets, random forest classification showed the highest accuracy of 0.9992, which means that the algorithm was able to correctly predict the outcome of the tweets in 99.92% of the cases. This high accuracy suggests that random forest classification was able to effectively capture the patterns and relationships in the data.

The other algorithms, logistic regression, support vector machine (SVM), and decision tree classification, were also used in the analysis. However, they did not produce the same level of accuracy as random forest classification, indicating that this algorithm may be better suited for this particular task.

The accuracy of different machine-learning algorithms on Twitter tweets is shown below:

S. No	Different Machine Learning Model Accuracy Report	Accuracy Report
1	Random Forest classification	0.9992
2	Logistic Regression	0.9515
3	Support Vector Machine	0.9143
4	Decision Tree Classification	0.9492

7. Result

In the lexicon-based approach, the sentiment of a tweet is determined based on the presence of positive or negative words in a predefined lexicon or dictionary. The TextBlob algorithm uses this approach by counting the number of positive and negative words in a tweet and determining the overall sentiment based on the count. The accuracy of this approach was 0.81, which suggests that it was able to correctly classify the sentiment of the tweets in the test dataset 81% of the time. In the machine learning approach, the algorithm is trained on the training dataset to identify patterns and relationships between the tweet text and its sentiment. Four different models were used in this approach: Decision Tree Classification, Random Forest Classification, Logistics Regression, and Support Vector Machine (RBF). Of these models, Decision Tree Classification was found to be the most accurate, with an accuracy of 0.9992. This high accuracy value indicates that the Decision Tree Classification algorithm was able to correctly classify the sentiment of the tweets in the test dataset 99.92% of the time.

It is important to note that while the Random Forest Classification Algorithm had a higher accuracy than the lexicon-based approach, both methods have their advantages and disadvantages. The lexicon-based approach is simpler and easier to implement, but its accuracy may be limited by the size and quality of the lexicon used. On the other hand, the machine learning approach can achieve higher accuracy, but requires a larger amount of data for training and can be more complex to implement. The choice between these two approaches will ultimately depend on the specific requirements of the task at hand

8. CONCLUSIONS

The purpose of the paper was to study sentiment analysis of tweets using two different approaches, namely the lexicon-based approach and the machine learning approach. The lexicon-based approach relied on pre-existing lexicons to determine the sentiment of a tweet, while the machine learning approach used algorithms to learn and make predictions based on the data. In the lexicon-based approach, two algorithms were tested, Vader and TextBlob. TextBlob was found to have the highest accuracy of 0.81, while Vader was found to have a lower accuracy. In the machine learning approach, four algorithms were tested: Random Forest Classification, Logistic Regression, Support Vector Machine, and Decision Tree Classification. Out of these, Random Forest Classification was found to have the highest accuracy of 0.9992, while Support Vector Machine had the lowest accuracy of 0.9143. Based on these results, we conclude that for Twitter sentiment analysis, the Random Forest Classification algorithm is the best option for the machine learning approach and TextBlob is the best option for the lexicon-based approach.

REFERENCES

[1]github.com›mohammed 97 ashraf›Sentiment-Analysis Using-Unsupervised-Lexical-Models

[2]github.com›laurenedeans›DataMining_Mining_For_Cryptocurrency

[3]Implementation of Naive Bayes Algorithm on Sentiment Analysis Application

[4] ieec.neduet.edu.pk›2018›Papers_2018›15.pdf

[5] "A comprehensive study on Lexicon based approach on Sentiment Analysis", Article: March 2019, Nandhini Kumaresh, Naulegari Janardhan, Central University of Tamil Nadu.

[6] Accuracy achieved using various classifiers for different books https://www.researchgate.net/figure/Accuracy-achieved using-various-classifiers-for-different books_tbl2_323503127