Sentiment Analysis and Bot Detection Using Logistic Regression

Jitendra
Assistant Professor
Galgotias University
Jitendra@galgotiasuniversity.edu.in

Shivangi Singh
Assistant Professor
Galgotias University
shivangisingh@galgotiasuniversity.edu.in

Deepanshu Singh
Galgotias University
Noida, India
deepanshu.21scse1010840@galgotiasuniversity.edu.in

Nikhil Mishra
Galgotias University
Noida, India
nikhil.21scse1010918@galgotiasuniversity.edu.in

Abstract Sentiment analysis is a way of identifying the emotion behind the statements provided and bot detection is a way of identifying if the statement is provided by the bot or the human. There can be too many feedback data or reviews that can be much time consuming whereas there is another problem that is if the feedback or the review is provided by the human or the bot. Through this project we are going to solve the mentioned problem in efficient way using machine learning technique such as support vector machine (SVM) that sorts text into categories like positive or negative by finding the best way to separate these feelings based on the words used and also identifies whether an interaction is from a real person or an automated bot by learning patterns in the behavior or text and then classifying it accordingly.

Keywords— Sentiment Analysis, Support Vector Machines, Bot Detection, Social Media, TF-IDF, Machine Learning

I. INTRODUCTION

People use social media to talk about a wide range of products and policies to social issues. Companies can use this platform to get information that helps them customize their products, respond to public concerns, and measure people how they feel about their brand. The biggest problem with social media data is that it is not aligned. This has been worse by the fact that B ots is toe and automatic accounts. These types of accounts can stick to the results of sentiment analysis and people seem to think something different than them. We need good ways to analyze the spirit and the boat so we can focus on the content generated by the actual user. Many people use logistic regression (LR). This paper details a project that gives LR the benefit of classification when integrating the boat detection mechanism to filter the synthetic material. The objective is that with the methods of choice, such as TF-IDF, how effective logistic regression is, to explore how effective the logistic regression is, to accurately analyze the spirit and ensure the integrity of the data.

II. LITERATURE REVIEW

A. Timeline of Sentiment Analysis Using SVM

In recent years, SVM has gained popularity as an automatic learning model for emotions analysis. Because the model can handle high -dimension data, it is ideal for the analysis of text-

based emotions, which often uses dispersed data sets with many characteristics. Ahmad, Aftab and Ali [1] tested SVM in data related to Apple products and autonomous cars to show their versatility. Their respective F-measures were 57.2% and 69.9%. Fikri and Sarno [2] demonstrated in a different comparative study that SVM, when combined with TF-IDF, outperformed rule-based techniques using SentiWordNet, achieving accuracy of 89%. SVM and k-nearest neighbors (KNN) were investigated by Huq, Ali, and Rahman [3] for Twitter sentiment analysis. They discovered that, although KNN went well in some situations, the precision of SVM increased with high - dimension data, which was consistent with its advantages in the management of scarce characteristics. Although they also point out that SVM performance may differ depending on the dimensionality and characteristics of the data set, these studies feel the bases for SVM as a reliable model for the classification of emotions.

B. Enhanced SVM Models with Advanced Features

To overcome some limitations of traditional SVM models, researchers have developed advanced models with advanced facility engineering. For example, Hen et al. [4] This MODEL Dell achieved 87.2% accuracy, with the complexity of language such as synonym and policy more effectively than the basic SVM models. Similarly, Mulen and Colier [5] created a hybrid SVM model by integrating the semantic orientation (SO) features, which enhances the performance of SVM on a complex linguistic noise text like sarcasm.

C. Application-Specific SVM Studies

The adaptability of SVM expands into different application-specific references. Tyagi and Sharma [6] SVM use to classify smartphone reviews, achieving 89.98% accuracy with preprocessing methods such as TF-IDF and POS tagging. Patil et al. [7] In the Healthcare Domain, Rahardi et al. [8] Used SVM to analyze the public spirit on the Kovid -19 vaccination, obtained a 92% accuracy rate with the RBF kernel.

D. Hybrid Approaches and Feature Selection Techniques

Hybrid approaches, which combine SVM with other optimization algorithms, have proven promising to further improve the precision of the classification of emotions. For example, Sharma and Sabharwal [9] combine the optimization of particle swarm (PSO) with a search for cuckoo to optimize the selection of characteristics. This model significantly exceeded convolutional neural networks (CNN) with a Twitter data set, achieving a 91.91% accuracy and an accuracy of 98.34%, illustrating the advantage of hybrid models in subset subsections of refining characteristics.

E. SVM in Product and Event Sentiment Tracking

Burequat and Mourad [10] took advantage of SVM to analyze consumer reactions to iPhone launches, achieving an 89.21%accuracy. Its preprocessing pipe included tokenization, elimination of stop and weighting words of TF-IDF, highlighting the SVM potential for applications in the monitoring of consumer opinion, where the classification of timely and precise emotions is critical.

F. Timeline of Sentiment Analysis Using Logistic Regression

Logistic Regression is a method of classification where you differentiate the information or data in two classes and when there is need for multiple classes then multinomial logistic regression is used [11]. For sentiment classification where you need to tell the sentiment in only two classes that is positive or negative, we can use logistic regression.

G. Logistic Regression with TF-IDF

TF-IDF (Term Frequency – Inverse Document Frequency) that is an advance feature extraction method can be implanted with logistic regression to vectorize the data that is converting the textual data into quantifiable form [12]. Data can be classified easily if its in quantifiable form.

H. Sentiment Analysis Using Naïve Bayes

The most commonly used classifier for sentiment analysis is naïve bayes as it is not as complex as other models like SVM. The Naive Bayes classifier is a simple and fast way to predict which category something belongs to, based on probabilities. It works by assuming that all the features it looks at are independent of each other. This makes it easy to use and gives good results in many cases. However, in real life, features often depend on each other, which can make this assumption unrealistic and lead to errors [13][14]. After implying naïve bayes it was noticeable that this method works well for negative comments but it was not much accurate with positive comments and it faced more problems with comments with ironic comments or other complex content [15].

G. Analysis on social networks using automatic learning: trends, challenges and performance evaluation

As observed in recent times, many people publish surveys and reviews about products, movies, games or events on social media platforms. It is essential that organizations determine the feelings of such reviews to understand public opinion about the product [16]. Most people tend to give neutral feelings, being the least common negative feelings. This indicates that the general state of mental health of people around the world remains neutral to positive [17]. Several studies have used the DOC2VEC model for the analysis of feelings together with the machines of support vectors; However, the best results have been obtained using logistics regression [18]. When working with large data sets, it is prudent to use algorithms that work in linear time [19]. Experimental results show that logistic regression achieves better precision, up to 94%, compared to other automatic learning algorithms, such as naive bayes, decision trees and more near neighbors [20]. While logistics regression exceeds naive bayes with precision, Naïve Bayes provides a more efficient data processing rate. In addition, the classification of feelings works better when using data sets marked with binary instead with three labels, since this facilitates the categorization of more effective feelings in market application reviews [21]. The number of data sets is far from being sufficient; Second, it is necessary to use Big Data frames, such as Hadoop and Spark due to a large amount of data in electronic commerce; Third, the data obtained are not complete enough, so more multidimensional data for experimentation is needed. [22] The length and the combination Perplexity helps improve performanceMerce slightly, compared to the use of language reflected in specific, frequency and emotional terms Expression, which captures more propaganda information. [23]

III. METHODOLOGY

A. **Data Collection and Preprocessing**

We gathered data from Twitter using its API, making sure we followed all the rules about how Twitter data can be used. This helped protect people's privacy and kept us in line with Twitter’s guidelines. To get a good variety of tweets, we focused on collecting messages related to specific keywords, hashtags, or user mentions.

After gathering the tweets, we cleaned up the data to make it easier to analyze and remove any unnecessary or confusing information. Here’s what we did to clean the text:

- **Removed Links**: To ensure that they wouldn't interfere with our analysis, we removed all of the website links (URLs).
- **Removed Mentions**: To make the text focus on the message itself instead of the person the tweet referred to, we removed references to other users (such as "@username").
- **Removed Special Characters and Emojis**: We got rid of symbols, special characters, and emojis, so we could focus only on the words and their meanings.
- **Normalized the Text**: To ensure that words like "Happy" and "happy" would be treated equally, we made all of the text lowercase. Additionally, we eliminated words that don't really contribute much to the analysis, such as "the," "is," and "in."

The goal of all this cleaning was to remove distractions and make the dataset cleaner, so it would be easier to analyze in the next steps.

B. Feature Engineering and Model Selection

The TF -DF was chosen as the method of selecting the primary installation to convert the lesson data in a numerical way. This technique effectively captures the importance of words in the context, which is suitable for emotions analysis. Logistics regression was used as the main classification model with its capacity for high -dimension data management. We used linear and RBF kernels to determine the optimal model performance required to convert text into a format, which the computer could understand. We used a method called TF -DF (Term Frequence-Invers Document Frequency), which helps identify the most important words in tweets. This ensures that general words do not dominate analysis and highlight words that are most important to understand the emotion. After converting the text into numbers, we selected a logistic region model to analyze the emotion (positive or negative emotions). Logistic regression works well with data like us because it can handle many characteristics (words) and is good in separating a variety of information. We tried two different settings for the model.

- **Linear Kernel**: A simpler method that assumes the data is easy to separate.
- **RBF Kernel**: A more flexible method that works well when the data is more complex.

We tested both to see which one worked best for our task.

C. Bot detection

Our Bot detection approach involved patterns recognition techniques, focusing on distinguishing automated behavior, such as excessive publication frequency and identical messages. Bot detection algorithms marked potentially synthetic accounts, which were then excluded from feelings analysis, improving data reliability**.**

D. Model training

Model training with multiple epochs so that the model can learn the pattern better and more efficient. We made a total of 300 times we improve the efficiency of the model**.**

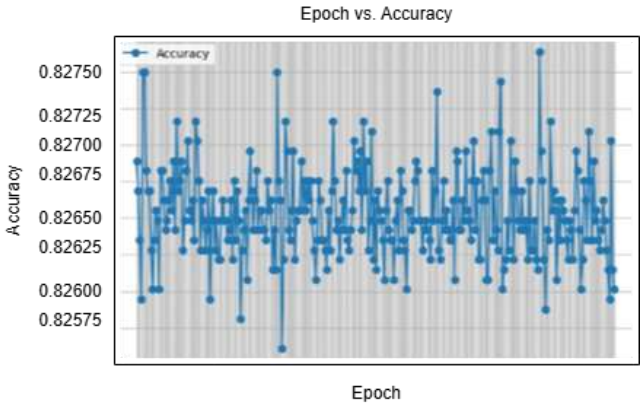

Figure.1. Training accuracy graph

E. Model Evaluation

We tested three models to find the sentiment of the textual data SVM, Naïve bayes and Logistic Regression and calculated the accuracy of each to choose the better one.

These formulas can be used to find out metrics of the models.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad \text{Eq.1}$$

$$Recall = \frac{TP}{(TP+FN)} \quad \text{Eq.2}$$

$$Precision = \frac{TP}{(TP+FP)} \quad \text{Eq.3}$$

$$\text{F1 Score} = 2 \times \left[\frac{(\text{Recall} \times \text{Precision})}{(\text{Recall} + \text{Precision})}\right] \quad \text{Eq.4}$$

In this regard, TP. (True Positive) refers to patterns where the model correctly identifies a comment as a positive emotion. FN (False negative) occurs when the model predicts a positive

emotion, but the real spirit is negative. On the other hand, FP (false positive) means that the model mistakenly classifies a negative comment as a positive. Finally, TN (True negative) represents cases where the model identifies the negative emotion accurately. These four categories help evaluate how well the sentiment analysis model is performing. A high number of true positives and true negatives usually indicates strong accuracy. Meanwhile, frequent false positives or false negatives suggest the model may be misinterpreting emotional tone. By analyzing these metrics, we can fine-tune the model to improve its predictions.

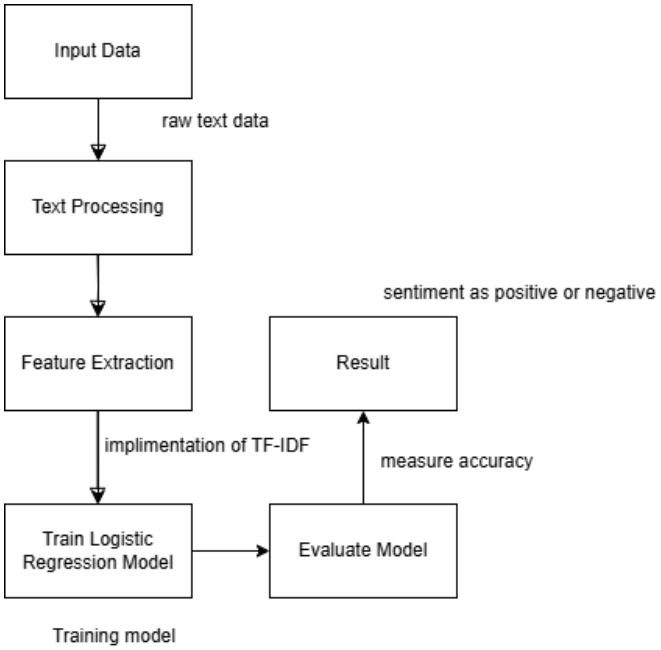

Figure. 2. Process flow of the system

F. Visualization

Visualization played an important role in providing emotion trends and comfortable understanding of major subjects in data. Many types of visual representations were used:

- Pie Chart: The entire dataset is used to show the distribution of emotions (positive, negative and neutral) throughout the dataset. This made it easier to figure out the public's general mood.
- Bar graph: Bar graph was used to display emotion distribution in various subgroups or categories, such as emotion by field, language or time period.
- Word Clouds: These were the most frequently used words in dataset to represent visually, with more prominent words indicate strong emotional associations. Word clouds are particularly helpful in identifying the keywords or trends within emotion data

These visualizations provided a comprehensive view of the data, enhancing its interpretability. They allowed stakeholders to quickly grasp the main sentiment trends and key topics, offering insights into public opinion and discourse patterns.

IV. RESULTS AND DISCUSSION

A. Sentiment Analysis Results

Using the cleaned Twitter data set, the logistics regression achieved an average accuracy of 82% for the classification of feelings. The results indicated that the model was particularly effective in distinguishing between positive, negative and neutral feelings, highlighting SVM robustness in the management of social networks data.

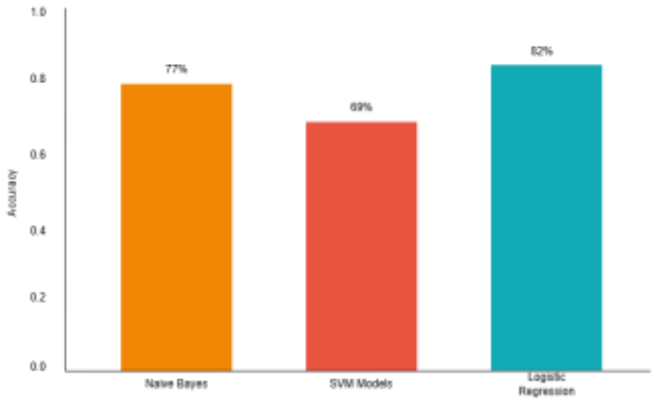

Figure 3. Comparison chart of models

B. Bot Detection Efficacy

The bot detection algorithm flagged approximately 5% of the dataset as inauthentic content, which was subsequently

excluded from analysis. This exclusion improved the overall accuracy of sentiment classification by ensuring that only genuine user sentiments were considered.

C. Comparative Analysis with Other Models

Table 1. Performance comparison of different models in percentage

Machine Learning Model	Precision	Recall	F1Score	Accuracy
SVM	72.13	75.86	73.95	69.00
Naïve Bayes	78.14	77.97	80.00	77.00
Logistic Regression	79.00	79.00	79.00	82.60

Compared to SVM logistics and naive bayes, logistics regression exhibited higher performance in terms of precision and processing efficiency, particularly in high -dimension and scarce data. The use of TF-IDF further improved SVM's ability to capture relevant characteristics, reinforcing SVM's suitability for sentiments analysis tasks on social media platforms.

V. CONCLUSION AND FUTURE WORK

This study demonstrates the effectiveness of logistics regression in the analysis of feelings. The ability of logistics regression to administer scarce and high-dimension data, together with the extraction of TF-IDF characteristics, positions it as a valuable tool for the analysis of feelings on social networks. However, future work could explore advanced NLP techniques, such as deep learning, to capture more complex feelings, including sarcasm and mixed emotions. In addition, the expansion of bot detection to identify more sophisticated automated behaviors would further improve data reliability. The multiplatform feelings analysis could also provide a broader understanding of the trends of public feelings, offering ideas on several social network platforms.

You still have to work on bot detection, bot detection can be very useful for detecting bots that are only there for spam comments if we eliminate the bots, then the planned sensation will be more significant.

REFERENCES

[1] Ahmad, M., Aftab, S., & Ali, I., "Sentiment analysis of tweets using SVM," *Int. J. Comput. Appl.*, vol. 177, no. 5, pp. 25-29, 2017.

[2] Fikri, M., & Sarno, R., "A comparative study of sentiment analysis using SVM and SentiWordNet," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 13, no. 3, pp. 902-909, 2019.

[3] Huq, M. R., Ali, A., & Rahman, A., "Sentiment analysis on Twitter data using KNN and SVM," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 6, 2017.

[4] Han, K. X., Chien, C. M., Chiu, S. M., & Cheng, C. T., "Application of support vector machine (SVM) in the sentiment analysis of Twitter dataset," *Appl. Sci.*, vol. 10, no. 3, p. 1125, 2020.

[5] Mullen, T., & Collier, N., "Sentiment analysis using support vector machines with diverse information sources," *Proc. of the 2004 Conf. on Empirical Methods in Nat. Lang. Process.*, 2004.

[6] Tyagi, E., & Sharma, A. K., "Sentiment analysis of product reviews using support vector machine learning algorithm," *Indian J. Sci. Technol.*, vol. 10, no. 35, 2017.

[7] Patil, G., et al., "Sentiment analysis using support vector machine," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 2, no. 1, pp. 2607-2612, 2014.

[8] Rahardi, M., et al., "Sentiment analysis of COVID-19 vaccination using support vector machine in Indonesia," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 6, 2022.

[9] Sharma, D., & Sabharwal, M., "Sentiment analysis for social media using SVM classifier of machine learning," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 9, pp. 39-47, 2019.

[10] Bourequat, W., & Mourad, H., "Sentiment analysis approach for analyzing iPhone release using support vector machine," *Int. J. Adv. Data Inf. Syst.*, vol. 2, no. 1, pp. 36-44, 2021

[11] Mayur Wankhade1, A Chandra Sekhara Rao1, Suresh Dara2, Baijnath Kaushik3" A Sentiment Analysis of Food Review using Logistic Regression" 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

[12] Sai Prasad, Dr. Vadivu G "Comparative Study of Logistic Regression and LSTM for Sentiment Classification Across Diverse Textual Dataset" November 2023

[13] Rothfels, J., Tibshirani, J. (2010). Unsupervised sentiment classification of En-glish movie reviews using automatic selection of positive and negative sentiment items. CS224N-Final Project

[14] Thakkar, H., Patel, D. (2015). Approaches for sentiment analysis on twitter: A state-of-art study. arXiv preprint arXiv:1512.01043

[15] Vikas Malik, Amit Kumar "Sentiment Analysis of Twitter Data Using Naive Bayes Algorithm" April 2018

[16] Abhilasha Tyagi1, Naresh Sharma2 1Dept. of Computer Science and Engineering, SRM University (India)-2012 042Assistant Professor, Dept. of CSE, SRM University (India)-201204, July 2018

[17] Imamah; Fika Hastarita Rachman, Twitter Sentiment Analysis of Covid-19 Using Term Weighting TF-IDF And Logistic Regresion,2020

[18] Tirta Hema Jaya Hidayat*, Yova Ruldeviyani, Achmad Rizki Aditama, Gusti Raditia Madya, Ade Wija Nugraha, Muhammad Wijaya Adisaputra, 2022

[19] Supriya raheja Amity University, Noida, India Anjani Asthana, Amity University, Noida, India

[20] Bahtiar, S. A., Dewa, C., & Luthfi, A. (2023). Comparison of Naïve Bayes and Logistic Regression in Sentiment Analysis on Marketplace Reviews Using Rating-Based Labeling. Journal of Information Systems and Informatics, 5(3), 915-927.

[21] Shuwei Xiao and Weiqin Tong 2021 J. Phys.: Conf. Ser. 1757 012089, 2020

[22] Li, Jinfen, Zhihao Ye, and Lu Xiao. "Detection of propaganda using logistic regression." *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*. 2019