ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14965910

Sentiment Analysis of Tweets: An Emoticon-Focused Method

[1]U Karthik; [2]Srinidhi K S; [3]Mithun Gowda H

[1,2,3] SJB Institute of Technology
Bangalore, India

Publication Date: 2025/03/08

Abstract: A recent development in natural language processing (NLP) is sentiment analysis of tweets leveraging emoticons, which uses emoticons' expressive potential to determine the sentiment contained in textual data. Emoticons, which are tiny visual representations of emotions, offer a natural approach to improve comprehension of the feelings expressed in conversations, social media posts, and other unofficial text formats. This approach consists of a number of crucial processes, beginning with data preprocessing, which cleans and normalizes texts. Next, emoticon extraction is used to find and classify emoticons into predetermined sentiment classifications, such positive and negative. By assembling related emoticons and the written information that goes with them into clusters, the k-means clustering algorithm plays a crucial part in this study by making it easier to spot common sentiment patterns. By dividing the dataset into k clusters according to feature similarity, the unsupervised learning algorithm K-means minimizes the variance within each cluster. The analysis can effectively manage massive amounts of data by using k-means clustering, which offers insights into the prevailing sentiment trends and how they change over time.

Additionally, by guaranteeing that contextual subtleties are recorded, the combination of clustering and natural language processing (NLP) approaches improves sentiment analysis and sentiment classification accuracy. The generated clusters help with activities like market analysis, customer feedback evaluation, and social media monitoring by providing a detailed view of the sentiment landscape. Essentially, a strong foundation for extracting and analysing feelings is provided by emoticon-based sentiment analysis utilizing NLP with k-means clustering, which promotes improved decision-making and a deeper comprehension of the audience.

Keywords: *Machine learning, Sentiment analysis, Tweet Classification, Natural Language Processing (NLP), Machine Learning, Text Preprocessing, Tokenization, Noise Reduction, Emoticon Extraction, Sentiment Score Assignment, BiLSTM, DistilBERT, Textual Sentiment, Emoticon-Based Sentiment.*

How to Cite: U Karthik; Srinidhi K S; Mithun Gowda H (2025) Sentiment Analysis of Tweets: An Emoticon-Focused Method. *International Journal of Innovative Science and Research Technology*, 10(2), 1689-1694.
https://doi.org/10.5281/zenodo.14965910

I. INTRODUCTION

Due to its numerous practical uses, sentiment analysis is significant. Businesses seek to learn new things to improve their customer service. They wish to draw in new clients while keeping their current clientele. Without using surveys or questionnaires, sentiment analysis enables businesses to do market research to assess consumer feedback.

The change in public perception of their candidates is something that election parties like to research. Any conventional algorithm, meanwhile, can miss genuine feelings that are concealed in tweet language. The difficulty of identifying irony, sarcasm, or humor in the text can result in tweets being misclassified. They might signify many things depending on the situation. Grammatical mistakes and misspelled words might increase the data set's noise.

It is possible for tweets that have both positive and negative language to be mistakenly categorized as neutral mood. Therefore, it is not enough to identify true emotions from tweets based solely on their language because it is crucial to comprehend the author's true intentions. Emotional iconography can significantly enhance sentiment analysis, sometimes referred to as opinion mining. Emoticons were introduced to social media platforms to depict an author's face features and to add emotion to text communications. Simple emoticons may communicate a lot of complicated ideas.

Instead of just looking at the words in a tweet to figure out if it's positive or negative, we can also pay attention to the emoticons. Think of it like this: traditional sentiment analysis tries to understand the tone of a message by dissecting the text for emotions such as sadness or joy. But people often use emoticons to add another layer of meaning, signalling whether something is good or bad, or even to be sarcastic. This

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14965910

research is all about digging into how people use emoticons in their tweets to express how they feel. Most studies focus on the text itself, but we're trying to broaden our understanding by including these visual cues to better grasp the feelings behind tweets. By understanding how emoticons are used, we hope to get a richer, more accurate sense of the emotions people share on Twitter.

II. LITERATURE SURVEY

In recent years, the focus of research in sentiment analysis has been on developing specific models for tasks such as emotion detection, sentiment classification, opinion mining, and context-aware sentiment prediction. Though this project integrates all these models into a single interface, which had not been developed earlier, it offers a more comprehensive approach to sentiment analysis. Some of the existing works in this field have been described below:

Alfreihat et al., [1] developed an Emoji Sentiment Lexicon (Emo-SL) for Arabic sentiment analysis on social media, utilizing a dataset of 58,000 Arabic tweets from source: the Arabic Sentiment Twitter Corpus . The study integrated emoji characteristics with text for classification using various ML models, including Linear Support Vector Machine (SVM),Support Vector Machine (SVM), Multinomial Naive Bayes, Bernoulli Naive Bayes, stochastic Gradient Descent (SGD) Classifier, Decision Tree Classifier, Random Forest classifier, and K-Neighbours classifier. Limitations include the shortage of large-scale public Arabic corpora and sentiment lexicons, challenges in handling informal dialectical Arabic, and the need for more advanced techniques to capture linguistic nuances. Future work aims to address these limitations by incorporating context-aware ML models, expanding the lexicon to include a wider range of emojis, exploring deep learning techniques for automatic feature extraction, and improving handling of noise, sarcasm detection, and dialect variations.

Sharma et al., [2] developed a model for binary sentiment classification of Hindi movie reviews ,using a manually annotated dataset of 10K reviews and additional IIT-P datasets. The study applied Term Frequency - Inverse Document Frequency (TF-IDF) with word-level N- gram features and ensemble-based classifiers, culminating in a Stacked Ensemble-Based Architecture (SEBA) that outperformed individual models. SEBA achieved an F1-score of 0.807 on the HLMR dataset. The study addressed challenges specific to Hindi, including lack of fixed word order, spelling variations, morphological richness, and use of "Hinglish". Limitations included scarcity of research, language complexities, dataset challenges, and label imbalance. Future work includes extending the model to other domains and improving feature extraction techniques.

Bengesi et al., [3] designed a system that involved two stages: (1) collecting over 500,000 multilingual tweets related to monkeypox and performing sentiment analysis using VADER and TextBlob to annotate them into positive, negative, and neutral sentiments; (2) designing, developing, and evaluating 56 classification models using stemming, lemmatization, CountVectorizer, TF-IDF, and machine learning algorithms like KNN, SVM, Random Forest, Logistic Regression, MLP, Naïve Bayes, and XGBoost. Performance was evaluated based on accuracy, F1 Score, Precision, and Recall.The study identifies gaps in previous research, including the limited time frame of data collection (focus on initial outbreak cases), lack of detailed preprocessing descriptions, and a focus on English-language tweets.The study utilized a dataset of over 500,000 multilingual tweets related to the monkeypox outbreak collected from Twitter1. The dataset comprised tweets in 103 languages.

Godard et al., [4] developed and validated the Multidimensional Lexicon of Emojis (MLE) to assess emotional content of emojis across specific and non-specific emotional domains. The dataset used included over 3 million Twitter posts collected at three time points: 1,014,363 tweets from November 7-20, 2019; 1,122,438 tweets from September 30 - November 2, 2020; and 1,021,715 tweets from February 20 - March 4, 2021. Emotion ratings were provided by 2,230 human raters for Study 2. Limitations included not accessing tweeter's actual experienced or intended emotions, assuming emojis matched the emotional content of the words, and data only from Twitter. Future works should validate the MLE against sender and recipient emotional experiences, investigate how emojis shift emotional tone, integrate human and automated ratings, consider culture-dependent use of emojis, and develop methods for rating non-face emojis.

He et al., [5] proposed a fusion sentiment analysis method for e-commerce product experience analysis, combining a sentiment dictionary with machine learning algorithms. They used review data from Amazon for Tao Te Ching books, with 5,480 total reviews (4,678 positive, 802negative). The method showed improved performance over other approaches on metrics like precision, recall and F-score. Limitations included only analyzing sentiment as positive/negative without finer-grained categories. The authors noted that the future work should conduct more comprehensive multi-dimensional sentiment analysis considering diverse affective aspects like surprise, anger, etc. beyond just positive and negative.

III. METHODOLOGY

In this paper, we aim to implement a comprehensive platform that integrates both textual and emoticon-based sentiment detection. The system leverages advanced machine learning models to classify sentiment, detect sarcasm, and cluster similar sentiment patterns. By utilizing large-scale datasets, the platform ensures accurate sentiment classification and trend analysis.

The platform focuses solely on sentiment classification by analyzing textual content and emoticons to determine sentiment polarity as positive, negative, or neutral. By employing machine learning techniques such as K-means clustering, and Naïve Bayes, the system enhances

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14965910

classification accuracy and identifies patterns in sentiment trends.

A. Datasets

In this project, we have used a dataset containing tweets for sentiment analysis. The dataset includes both textual content and associated emoticons, which help in refining sentiment classification. A total of 13094 tweets were selected, where emoji has about 3000 tweets each.

Preprocessing involved collecting raw tweet data, cleaning unnecessary elements, and standardizing the content for sentiment classification. Unnecessary columns were removed, and missing values were handled to maintain data integrity. Tokenization and normalization were applied to process text efficiently, while emoticons were extracted and mapped to sentiment categories. Additionally, noise such as URLs, special characters, and redundant data was eliminated to ensure a structured and optimized dataset for accurate sentiment analysis.

After preprocessing, the cleaned and structured data was prepared for feature extraction and sentiment scoring. Text data was converted into numerical representations using word embeddings and TF-IDF, while emoticons were assigned predefined sentiment values. These extracted features were then used to train sentiment classification models, ensuring that both textual content and emoticons contributed effectively to sentiment prediction.

B. Model Selection

First, we must select suitable machine learning techniques for sentiment classification. Use the dataset to train all of these models, including Naïve Bayes, K-Nearest Neighbors (KNN), Bidirectional LSTM (BiLSTM), DistilBERT, K-Means Clustering. Using ensemble techniques like model stacking, which combines multiple models and compares their performance, these predictions from various models are integrated for increased accuracy.

➢ *Sentiment Score Assignment*

Before sentiment classification, a sentiment score is assigned to each word in the tweet. These scores are derived from predefined sentiment lexicons and machine learning models. The overall sentiment score of the tweet is then calculated as the sum or weighted combination of individual word scores.

The final sentiment score determines the tweet's polarity:
Positive Sentiment: Score > 0
Neutral Sentiment: Score = 0
Negative Sentiment: Score < 0

ALGORITHMS USED:

- *LSTM (Long Short-Term Memory):*

LSTM is a deep learning-based sequential model used for sentiment regression. It captures long-range dependencies in text and dynamically assigns sentiment scores to words based on contextual meaning. The model learns from labeled sentiment datasets and improves its predictions over time. The performance of the LSTM model is evaluated using metrics such as Mean Squared Error and R-squared (R^2).

- *DistilBERT-based Sentiment Regression:*

DistilBERT is a transformer-based model optimized for efficient sentiment regression. It assigns sentiment scores to words based on contextual embeddings, capturing nuanced sentiment variations. The model is fine-tuned on sentiment datasets to improve accuracy and robustness. The performance of DistilBERT is evaluated using MSE and R-squared metrics.

➢ *Sentiment Classification*

Utilizing input features such as word embeddings, sentiment lexicon scores, emoticon sentiment values, punctuation usage, and contextual dependencies, we train suitable classification models, including Naïve Bayes, K-Nearest Neighbors (KNN), Bidirectional LSTM (BiLSTM), and DistilBERT.

ALGORITHMS USED:

- *Naïve Bayes Classifier*

Naïve Bayes is a probabilistic classifier based on Bayes' Theorem, commonly used for text classification. It assumes that words in a sentence are independent of each other and calculates the probability of a tweet belonging to a specific sentiment class. When used on our dataset, it achieved an accuracy of 87.5% for sentiment classification.

- *K-Nearest Neighbors (KNN)*

KNN is a distance-based classification algorithm that assigns sentiment by comparing the new input tweet with the k most similar tweets from the training dataset. It determines sentiment polarity based on majority voting among neighbors. When applied to our dataset, it achieved 89.73% accuracy in sentiment classification.

- *Bidirectional LSTM (BiLSTM)*

BiLSTM is a deep learning model that captures long-range dependencies in text by processing it in both forward and backward directions. This helps in understanding the sentiment context more accurately. When trained on our dataset, it provided 88.76% accuracy, effectively detecting nuanced sentiments.

- *DistilBERT*

DistilBERT is a transformer-based model optimized for efficient sentiment classification. It processes the entire text at once and uses contextual embeddings to understand word meanings in relation to surrounding words. When fine-tuned on our dataset, it achieved 90.2% accuracy, making it the best-performing model in our sentiment classification system.

C. Model Training

The input parameters are processed as features, and the selected sentiment classification model is trained using labeled datasets. Performance metrics such as accuracy, precision, recall, and F1-score are used for evaluation. The model's generalization capability is validated using holdout

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14965910

validation or cross-validation, ensuring its effectiveness across unseen data.

- *Naïve Bayes:*

A probabilistic classification model that estimates sentiment polarity based on word frequency and prior probabilities. It is simple yet effective for text classification but assumes word independence, which may limit accuracy in complex sentences

- *K-Nearest Neighbors (KNN):*

A distance-based classification technique that assigns sentiment labels by comparing a new tweet with the most similar examples in the dataset. It works well with small datasets but can be computationally expensive for large-scale sentiment analysis.

- *Bidirectional LSTM (BiLSTM):*

A deep learning-based model that captures context by processing text in both forward and backward directions. This allows it to better understand sentiment nuances, making it suitable for analyzing complex sentence structures.

- *DistilBER`T:*

A transformer-based model designed for fast and efficient sentiment classification. It processes entire text inputs simultaneously, using contextual embeddings to determine sentiment with high accuracy while being computationally optimized compared to larger transformer models.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

D. Proposed Architecture

The architecture integrates machine learning and deep learning models for preprocessing tweets, assigning sentiment scores, classifying text and emoticons separately, and combining both for overall sentiment classification. The system efficiently processes large-scale pre-existing tweet datasets, ensuring accurate and context-aware sentiment classification.

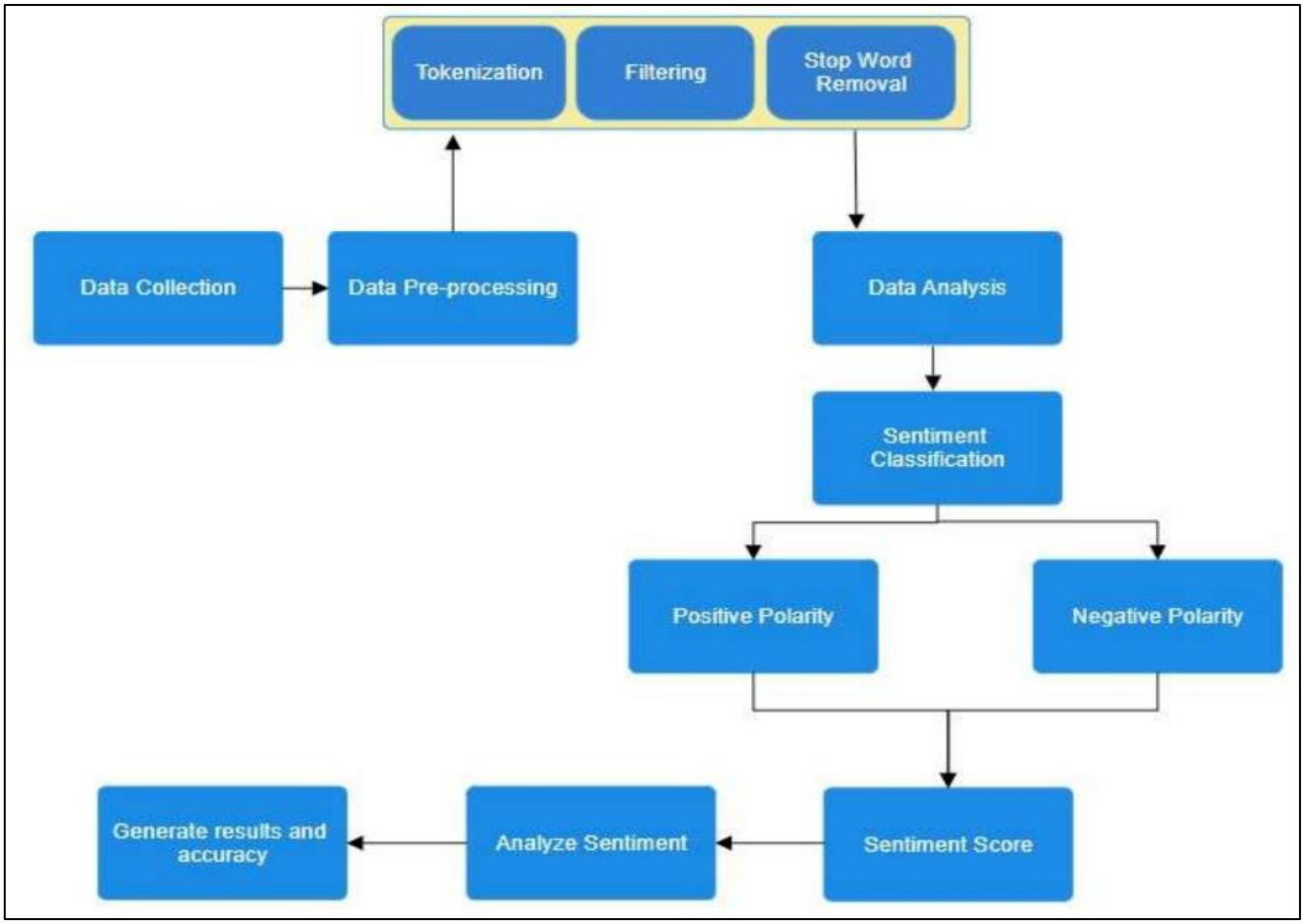

Fig.1. Proposed Architecture

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14965910

Since tweets contain both text and emoticons, it is crucial to analyze them separately because they contribute to sentiment in different ways. While text provides contextual meaning, emoticons serve as direct sentiment indicators, sometimes reinforcing or altering the sentiment of the text. By processing them separately, we can capture cases where emoticons contradict or intensify the sentiment expressed in words.

A variety of data preprocessing, feature extraction, sentiment scoring, and classification techniques are integrated into this modular system. The architecture follows a structured pipeline, ensuring that the text content and emoticons are processed separately before combining them for the final sentiment prediction.

E. Evaluation

The model was tested using metrices such as:

➢ *Balanced Performance Index (BPI):*

This metric ensures that both textual sentiment classification and emoticon sentiment classification contribute equally to the final sentiment prediction. It prevents the model from favoring one modality over the other.

$$BPI = \frac{\text{Text Accuracy} + \text{Emoticon Accuracy}}{2}$$

➢ *Classification Success Rate (CSR):*

Measures the proportion of correctly classified sentiment labels across all predictions. It provides an overall assessment of the model's performance.

$$CSR = \frac{\text{Correctly Classified Tweets}}{\text{Total Tweets}} \times 100$$

➢ *Error Rate:*

This metric calculates the percentage of incorrectly classified tweets, highlighting potential misclassifications and areas for improvement.

$$ErrorRate = \frac{\text{Mis Classified Tweets}}{\text{Total Tweets}} \times 100$$

➢ *Sentiment Score Deviation (SSD):*

Measures the difference between the predicted sentiment score and the actual human-annotated sentiment score. A lower SSD value indicates that the model is effectively predicting sentiment with minimal deviation.

$$SSD = \frac{|\text{Predicted Sentiment Score} - \text{Actual Score}|}{\text{Total Samples}}$$

IV. RESULTS

The sentiment classification model was evaluated based on its ability to accurately analyze tweets by considering both text and emoticons. The effectiveness of the system was assessed using relevant performance metrics, ensuring that the integration of text-based and emoticon-based sentiment contributed to a well-rounded classification. The following section presents the evaluation of sentiment classification accuracy, highlighting key findings.

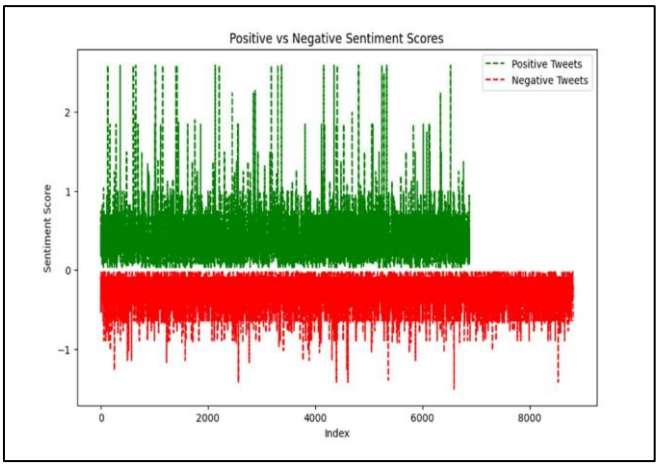

Fig 2. Scatter Plot of Sentiment Scores

Figure 2 illustrates the distribution of positive and negative tweets, providing insight into the variance in sentiment scores.

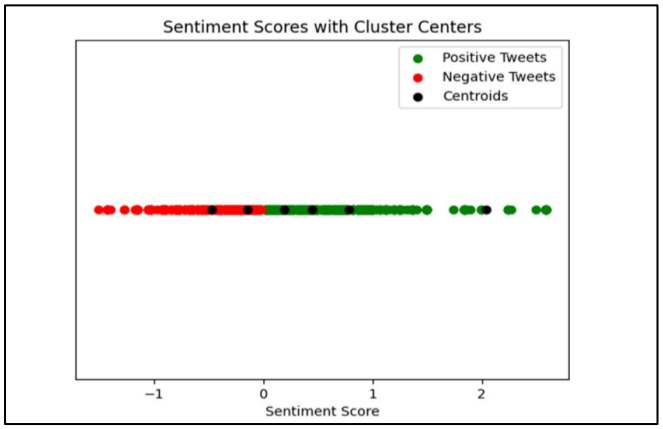

Fig 3. Data Distribution with Final Cluster Centroids

Fig 3 depicts the presence of multiple centroids indicates that the tweets can be grouped into different clusters based on their sentiment scores, providing insights into the overall sentiment distribution within the dataset.

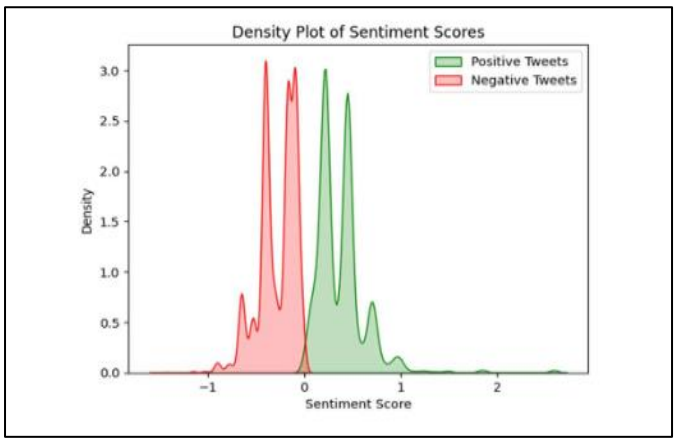

Fig 4. Data Distribution with Final Cluster Centroids

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14965910

The peaks in the density plot indicate the most common sentiment scores within each category, suggesting that most tweets tend t have moderate sentiments rather than extreme ones. The overlap around the sentiment score of 0 suggests that there are some tweets with neutral sentiments that could be categorized as either slightly positive or slightly negative.

V. FUTURE SCOPE

The results of the analysis show that some clusters of emoticons have achieved higher accuracy rates compared to others, indicating that the model was good at some sentiments and poor at others. This means that further refining is required to improve the overall performance. Future development of this project could emphasize enlarging the dataset to have more diverse and balanced ranges of emoticons to enable adequate representation of the different sentiments. Data quality and feature extraction could be further improved by making use of advanced preprocessing techniques like sophisticated tokenization and stemming. Implementing even more complex models, including deep learning techniques, would further boost the accuracy in sentiment classification. Hybrid approaches using a combination of traditional machine learning with neural networks might also help. Fine-tuning hyperparameters and leveraging transfer learning may finally be used for the optimization of efficiency and robustness of the model.

VI. CONCLUSION

The project on emoticon-based sentiment analysis of random tweets demonstrates the potential of machine learning in classifying sentiments through emoticons. Using a dataset with sentiment scores for positive and negative emoticons, we effectively preprocessed, clustered, and analyzed the data.

Efficient handling of data loading, normalization, and shuffling ensured a robust dataset for model training. The KMeans algorithm grouped sentiment scores into meaningful clusters, capturing emoticons' emotional nuances. The accuracy of sentiment analysis was calculated for each cluster, with the average accuracy providing a comprehensive performance measure.

Bar charts illustrated accuracy distribution across clusters, offering insights into strengths and areas for improvement. The project achieved a satisfactory average accuracy, showcasing the model's capability to classify sentiments based on emoticons.

While the model performs well, future work could expand the dataset, enhance preprocessing techniques, and explore deeper neural networks for improved accuracy. This project highlights the feasibility of sentiment analysis using emoticons in tweets, forming a foundation for advanced sentiment classification in social media analysis.

REFERENCES

[1]. M. Alfreihat, O. S. Almousa, Y. Tashtoush, A. AlSobeh, K. Mansour and H. Migdady, "Emo-SL Framework: Emoji Sentiment Lexicon Using Text-Based Features and Machine Learning for Sentiment Analysis," in IEEE Access, vol. 12, pp. 81793-81812,doi:10.1109/ACCESS.2024.3382836 ,2024.

[2]. A. Sharma and U. Ghose, "Toward Machine Learning Based Binary Sentiment Classification of Movie Reviews for Resource Restraint Language (RRL)—Hindi," in IEEE Access, vol. 11,pp. 58546-58564, doi: 10.1109/ACCESS.2023.3283461, 2023.

[3]. S. Bengesi, T. Oladunni, R. Olusegun and H. Audu, "A Machine Learning-Sentiment Analysis on Monkeypox Outbreak: An Extensive Dataset to Show the Polarity of Public Opinion From Twitter Tweets," in IEEE Access, vol. 11, pp. 11811-11826,doi:10.1109/ACCESS.2023.3242290, 2023.

[4]. R. Godard and S. Holtzman, "The Multidimensional Lexicon of Emojis: A New Tool to Assess the Emotional Content of Emojis," Frontiers in Psychology, vol. 13. doi:10.3389/fpsyg.2022.921388, 2022.

[5]. H. He, G. Zhou and S. Zhao, "Exploring E-Commerce Product Experience Based on Fusion Sentiment Analysis Method," in *IEEE Access*, vol. 10, pp. 110248-110260, 2022, doi: 10.1109/ACCESS.2022.3214752.

[6]. Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. Sentiment of emojis. PLoS ONE 10, 12 ,e0144296. http://dx.doi.org/10.1371/journal.pone.0144296.

[7]. M. J. Li, M. K. Ng, Y. m. Cheung, and J. Z. Huang. Agglomerative Fuzzy K-Means Clustering Algorithm with Selection of Number of Clusters. IEEE Transactions on Knowledge and Data Engineering.https://doi.org/10.1109/TKDE.2008.88

[8]. Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J.Bethard,and David McClosky. The Stanford Core NLP Natural Language Processing Toolkit. In Association for Computational Linguistics (ACL) System Demonstrations. 55–60.

[9]. K. Rani Narejo, H. Zan, D. Oralbekova, K. Parkash Dharmani, M. Orken and K. Mukhsina, "Enhancing Emoji-Based Sentiment Classification in Urdu Tweets: Fusion Strategies With Multilingual BERT and Emoji Embeddings," in *IEEE Access*, vol. 12, pp. 126587-126600, 2024, doi: 10.1109/ACCESS.2024.3446897.

[10]. K. R. Narejo *et al*., "EEBERT: An Emoji-Enhanced BERT Fine-Tuning on Amazon Product Reviews for Text Sentiment Classification," in *IEEE Access*, vol. 12, pp. 131954-131967, 2024, doi: 10.1109/ACCESS.2024.3456039