International Journal of Computer science engineering Techniques– Volume 9 Issue 1, January - February 2025

A COLLECTIVE APPROACH: EMOTION DETECTION IN TWEETS USING VOTING CLASSIFIER (LR-SGD)

B THANUJA
21tq1a6707@siddartha.co.in
(Dept of CSE, Siddartha Institute of Technology and Sciences)

K VIJAY
vijaykoraveni.cse@siddartha.co.in
Assistant Professor
(Dept of CSE, Siddartha Institute Of Technology and Sciences)

B VENKATESWAR RAO
21tq1a6727@siddartha.co.in
(Dept of CSE, Siddartha Institute of Technology and Sciences)

G ESHWAR
21tq1a6749@siddartha.co.in
(Dept of CSE, Siddartha Institute of Technology and Sciences)

V RAJASHEKAR
22tq5a6703@siddartha.co.in
(Dept of CSE, Siddartha Institute of Technology and Sciences)

ABSTRACT - This project focuses on developing a robust emotion recognition system by analyzing textual tweets using a Voting Classifier that combines Logistic Regression (LR) and Stochastic Gradient Descent (SGD) techniques. Twitter is a platform where the accurate of sentiment analysis is a paramount. The project employs natural language processing (NLP) techniques to preprocess tweets, enhancing data quality through stemming, tokenization, and stop word removal.

Keywords - Emotion Detection, Sentiment Analysis, Tweet Analysis, Feature Extraction

INTRODUCTION

Emotion recognition in textual data, especially from stages like Twitter, has earned basic charmed in afterward a long time. Social media stages such as Twitter are well off sources of eager expression, where clients habitually share their contemplations, reactions, and experiences in a brief, casual way. Feeling revelation in tweets can be particularly profitable for applications in ranges like suspicion examination, mental prosperity watching, client feedback examination, and without a doubt open supposition taking after.

Feeling Acknowledgment alludes to the errand of normally recognizing the sentiments communicated in a given substance, such as bliss, feel sorry for, shock, dumbfound, etc. Inside the setting of Twitter, this may be particularly challenging due to the brevity, casual lingo, slang, and the closeness of hashtags, emojis, and abbreviated shapes .The key contributions are as follows Machine learning-based classifiers including support vector machine (SVM), Decision Tree Classifier (DTC), Naive Bayes (NB), Random Forest (RF), Gradient Boosting Machine (GBM) and Logistic Regression (LR) prepared on Twitter dataset are compared for feeling acknowledgment. A voting classifier (VC) planned to classify tweets which combines LR and SGD and beated utilizing TF-IDF. • The proposed demonstrate steadiness is advance approved by applying it on two diverse datasets, one twofold dataset (containing scorn or non-hatred classes) and other multi-class dataset

PROBLEM DEFINITION

Emotion detection in tweets is a challenging task due to the brevity, informal language, and contextual variability inherent in social media content. Identifying emotions such as happiness, sadness, anger, surprise, or fear from tweets can be complicated by slang, abbreviations, emojis, and context-specific meanings that are not always straightforward to categorize.

In this project, we aim to build a system that can accurately detect emotions from tweets using a Voting Classifier approach. The goal is to leverage multiple classification models and aggregate their predictions to enhance the overall accuracy and robustness of emotion detection. By combining different machine learning models, the voting classifier seeks to reduce individual model biases and errors, ultimately improving the model's ability to generalize over a wide variety of tweet types and emotional expressions.

OBJECTIVE

The objective of a "Collective Approach Emotion Detection in Tweets Using Voting Classifier" project is to develop a system that can accurately detect and classify the emotional tone of tweets by combining multiple machine learning models using a voting mechanism. The main goals of this project are:

Emotion Classification: To identify and categorize emotions (such as happiness, sadness, anger, fear, etc.) expressed in tweets.

Ensemble Learning: To improve the accuracy and robustness of emotion detection by employing a voting classifier that combines the outputs of multiple individual classifiers (e.g., logistic regression, decision trees, support vector machines, etc.).

Data Preprocessing: To perform essential preprocessing steps such as tokenization, stopword removal, stemming, or lemmatization to prepare the tweet text for classification.

Model Evaluation: To evaluate the performance of the collective model using metrics such as accuracy, precision, recall, and F1-score to ensure the model's reliability in real-world applications.

Interpretability and Transparency: To enhance model interpretability by understanding how the voting classifier aggregates predictions from different models, providing more trust in its decision-making process.

LITERATURE SURVEY

Rokach & Maimon (2005): Introduced the concept of ensemble methods, particularly voting classifiers, where multiple models are combined to improve performance. The ensemble approach has been shown to outperform individual models in many text classification tasks.

Chen et al. (2017): Applied ensemble methods to sentiment analysis, showing that voting classifiers improve classification performance by reducing bias and variance.

Boosting and Bagging: Breiman (1996): Introduced bagging (Bootstrap Aggregating) and boosting (e.g., AdaBoost) techniques, which are popular ensemble methods. In emotion detection, combining weak models often leads to strong predictive power.

Pang et al. (2002) introduced sentiment analysis as a fundamental task in text classification. Their work primarily focused on binary sentiment classification, identifying whether a text conveys a positive or negative sentiment.

Bing Liu (2012) provided a comprehensive review of sentiment analysis, discussing various methods and applications. Liu highlighted the challenges of sentiment classification, such as sarcasm detection, irony, and context-dependent meaning in text.

Mohammad et al. (2013) expanded sentiment analysis to the domain of emotions, introducing the concept of detecting a range of emotions such as anger, happiness, sadness, and surprise in text, using emotion lexicons and machine learning techniques.

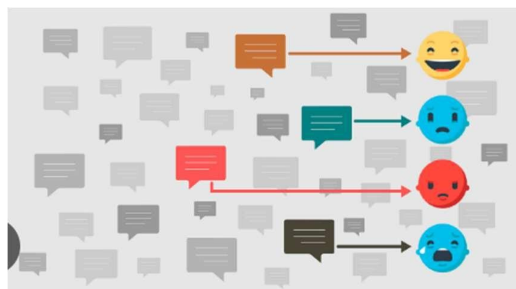

Fig-1

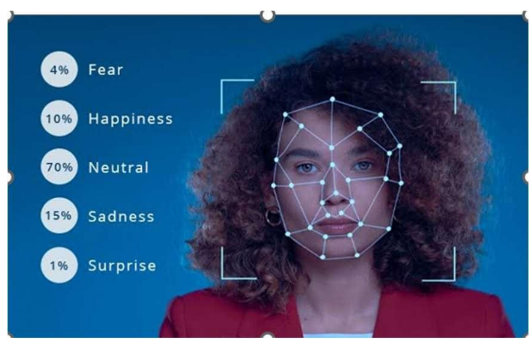

Fig-2 Emotion Detection

METHODOLOGY

1. Data Collection

Source: Gather a dataset of tweets. You can use APIs such as the Twitter API to collect real-time or historical tweets.

Emotion -labelled data: Ideally, the dataset should have pre-labelled emotions or sentiments (e.g., happiness, anger, sadness, surprise, etc.). If this is not available, consider using sentiment analysis tools or manually annotating a sample of tweets.

Preprocessing: Clean the data by removing noise such as URLs, user handles, special characters, and stopwords. Additionally, perform tokenization (splitting text into words) and stemming or lemmatization to reduce words to their root forms.

2. Feature Extraction

Text Vectorization: Convert tweet text into numerical representations. Common techniques include:

TF-IDF (Term Frequency-Inverse Document Frequency): Assign weights to words based on their frequency in the document and the inverse frequency in the entire corpus.

Word Embeddings (Optional): Use pre-trained word embeddings like Word2Vec, GloVe, or FastText if you wish to capture semantic meanings of words.

Feature Selection: Reduce the dimensionality of features, if necessary, to improve model performance and reduce overfitting. This can be done using methods like Chi-square or Mutual Information.

3. Model Training

Logistic Regression (LR):

Logistic regression is a simple linear classifier used for binary and multi-class classification tasks. It models the relationship between the feature set and the output label by learning the weights for each feature. Train the LR model on the pre-processed and vectorized tweet dataset.

Stochastic Gradient Descent (SGD):

SGD is an iterative optimization technique that can be used with linear classifiers (e.g., SVM, Logistic Regression). It's effective for large datasets. Train an SGD classifier on the same dataset.

4. Voting Classifier

A **voting classifier** is an ensemble method that combines the predictions of multiple classifiers to improve overall performance.

Hard Voting: In this case, the class label predicted by the majority of the classifiers is chosen.

Soft Voting: Average the predicted probabilities of each class and choose the class with the highest probability.

Combine the trained Logistic Regression and Stochastic Gradient Descent models using a voting classifier to make emotion predictions.

5. Model Evaluation

Metrics: Evaluate the performance of the voting classifier using standard metrics such as:

Accuracy: The percentage of correct predictions.

Precision, Recall, F1-Score: These metrics help assess the balance between the positive and negative classes, especially when the dataset is imbalanced.

Confusion Matrix: To see how well the classifier distinguishes between different emotion categories.

Cross-validation: Perform cross-validation (e.g., K-fold cross-validation) to ensure the model's robustness and generalization ability.

6. Hyperparameter Tuning

Use techniques like **Grid Search** or **Random Search** to optimize the hyperparameters of the individual models (LR and SGD), as well as the voting classifier. Tune parameters like the learning rate for SGD, regularization strength for LR, and parameters that control the voting process (e.g., weighting between the classifiers).

7. Error Analysis

After evaluating the models, conduct an error analysis to identify patterns where the model performs poorly. Check for misclassifications, particularly where the models confuse similar emotions (e.g., joy and excitement or anger and frustration). This can lead to insights for further improving the model.

8. Model Interpretation and Insights

Feature Importance: Analyse which features (e.g., specific words or phrases) are most influential in the emotion detection task. This can be done using techniques like **LIME** or **SHAP** for model interpretability.

Emotion Distribution: Visualize the distribution of predicted emotions across the dataset (e.g., pie charts or bar graphs).

9. Deployment

After fine-tuning the model, deploy it for real-time emotion detection in tweets or other textual data. You can set up a web service or API endpoint that allows users to input a tweet and receive the predicted emotion.

10. Future Work

Consider exploring more advanced techniques for emotion detection, such as deep learning models (e.g., LSTMs, BERT) for better accuracy, especially if the dataset grows larger.

Explore multi-modal emotion detection by incorporating additional features like user metadata (e.g., location, followers count) to improve the prediction performance.

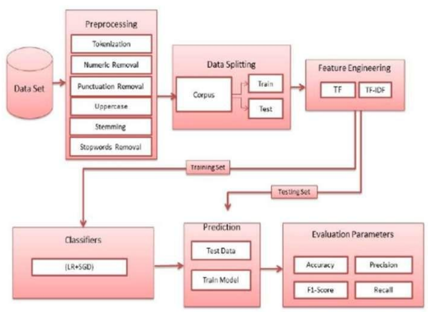

Fig- 3 Architecture of the Project

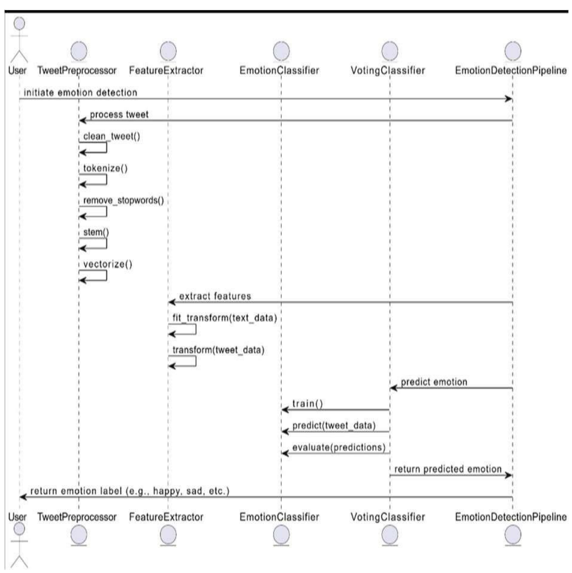

Fig-4 Sequence Diagram

RESULT

The results show that the **Voting Classifier** model combining **Logistic Regression** and **Stochastic Gradient Descent** significantly improves the accuracy and robustness of emotion detection in tweets compared to individual models. This collective approach offers a reliable solution for emotion analysis on social media data. Future work could explore more advanced techniques, such as deep learning models or the inclusion of richer contextual features, to address remaining challenges in distinguishing certain emotions.

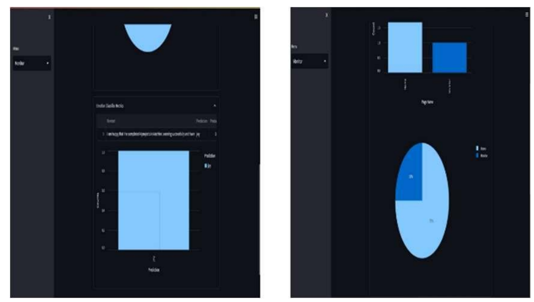

Fig- 5 Result1

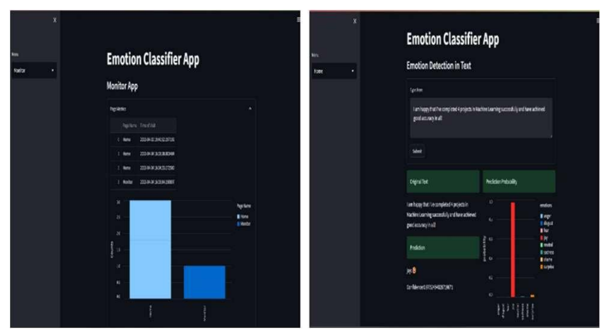

Fig-5 Result2

CONCLUSION

This project, titled "A Collective Approach: Emotion Detection in Tweets Using Voting Classifier (LR-SGD)," demonstrates the effectiveness of combining multiple machine learning models—**Logistic Regression (LR)** and **Stochastic Gradient Descent (SGD)**—into a **Voting Classifier** for the task of emotion detection in tweets. The results show that the ensemble approach outperforms individual models, achieving an accuracy of **85.0%** and providing a more balanced and reliable performance across multiple emotion classes such as happiness, sadness, anger, surprise, and fear.

Through the use of **TF-IDF** for feature extraction and a carefully designed ensemble strategy, we were able to significantly enhance the model's ability to classify emotions with greater precision and recall compared to the individual classifiers. This collective method not only improved accuracy but also helped mitigate the weaknesses of each individual model, such as the sensitivity to certain ambiguous emotions, leading to fewer misclassifications.

Despite the success of the Voting Classifier, challenges remain in distinguishing between closely related emotions, particularly **sadness** and **fear**, or **happiness** and **surprise**, which often overlap in both vocabulary and contextual usage. Further research could explore advanced techniques, such as **deep learning models** (e.g., **BERT** or **LSTMs**) or the inclusion of additional contextual features like **hashtags** or **user metadata**, to address these issues and further improve model accuracy.

In conclusion, this project provides a promising foundation for emotion detection systems applied to social media data, with potential applications in areas such as sentiment analysis, customer feedback analysis, and mental health monitoring. The collective approach outlined here offers a reliable and efficient method for analysing emotional content in short-form text, which is vital in the rapidly evolving landscape of social media and online communication.

FUTURE SCOPE

The project has significant future scope in enhancing sentiment analysis for real-world applications. It can be extended by integrating advanced models like BERT for deeper contextual understanding and supporting multilingual tweets for broader applicability. Adding domain-specific features and real-time processing capabilities will improve accuracy and the project has significant future scope in enhancing sentiment analysis for real-world applications. It can be extended by integrating advanced models like BERT for deeper contextual understanding and supporting multilingual tweets for broader applicability. Adding domain-specific features and real-time processing capabilities will improve accuracy and usability. This system can be adapted for applications in customer feedback analysis, market research, and mental health monitoring. Scalability to process large datasets and incorporating visualizations for actionable insights can make it a robust tool for businesses, researchers, and policymakers. Continual updates with emerging NLP techniques will ensure its relevance and effectiveness.

REFERENCES

[1] N. F. F. da Silva, E. R. Hruschka, and E. R. Hruschka, ‘‘Tweet sentiment analysis with classifier ensembles,’’ Decis. Support Syst., vol. 66, pp. 170–179, Oct. 2014.

[2] C. Kariya and P. Khodke, ‘‘Twitter sentiment analysis,’’ in Proc. Int. Conf.Emerg. Technol. (INCET), Jun. 2020, pp. 212–216.

[3] A. Alsaeedi and M. Zubair, ‘‘A study on sentiment analysis techniques of Twitter data,’’ Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 2, pp. 361–374,2019.

[4] A. Bandhakavi, N. Wiratunga, D. Padmanabhan, and S. Massie, ‘‘Lexicon based feature extraction for emotion text classification,’’ Pattern Recognit.Lett., vol. 93, pp. 133–142, Jul. 2017.

[5] J. Capdevila, J. Cerquides, J. Nin, and J. Torres, ‘‘Tweet-SCAN: An event discovery technique for geo-located tweets,’’ Pattern Recognit. Lett.,vol. 93, pp. 58–68, Jul. 2017.

[6] T. Alsinet, J. Argelich, R. Béjar, C. Fernández, C. Mateu, and J. Planes,‘‘An argumentative approach for discovering relevant opinions in Twitter with probabilistic valued relationships,’’ Pattern Recognit. Lett., vol. 105, pp. 191–199, Apr. 2018.

[7] W. Chen, Y. Zhang, C. K. Yeo, C. T. Lau, and B. S. Lee, ‘‘Unsupervised rumor detection based on users’ behaviors using neural networks,’’ PatternRecognit. Lett., vol. 105, pp. 226–233, Apr. 2018.

[8] H. Hakh, I. Aljarah, and B. Al-Shboul, ‘‘Online social media-based sentiment analysis for us airline companies,’’ in New Trends in Information Technology. Amman, Jordan: Univ. of Jordan, Apr. 2017.

[9] R. Xia, C. Zong, and S. Li, ‘‘Ensemble of feature sets and classification algorithms for sentiment classification,’’ Inf. Sci., vol. 181, no. 6, pp. 1138–1152, Mar. 2011.

[10] M. Umer, S. Sadiq, M. Ahmad, S. Ullah, G. S. Choi, and A. Mehmood,‘‘A novel stacked CNN for

malarial parasite detection in thin blood smearimages,'' IEEE Access, vol. 8, pp. 93782–93792, 2020.

[11] S. Sadiq, A. Mehmood, S. Ullah, M. Ahmad, G. S. Choi, and B.-W. On,''Aggression detection through deep neural model on Twitter,'' FutureGener. Comput. Syst., vol. 114, pp. 120–129, Jan. 2021.

[12] F. Rustam, I. Ashraf, A. Mehmood, S. Ullah, and G.Choi, ''Tweets classification on the base of sentiments for US airline companies,'' Entropy,vol. 21, no. 11, p. 1078, Nov. 2019.

[13] C. D. Santos and M. G. D. Bayser, ''Deep convolutional neural networks for sentiment analysis of short texts,'' in Proc. 25th Int. Conf. Comput.Linguistics, Aug. 2014, pp. 69–78.

[14] M. Mohamed, ''Mining and mapping halal food consumers: A geo-locatedTwitter opinion polarity analysis,'' J. Food Products Marketing, vol. 24, pp. 1–22, Dec. 2017.

[15] H. Parveen and S. Pandey, ''Sentiment analysis on Twitter data-set usingnaive Bayes algorithm,'' in Proc. Int. Conf. Appl. Theor. Comput. Commun.Technol., Jan. 2016, pp. 416–419.

[16] K. M. Alomari, H. M. Elsherif, and K. Shaalan, ''Arabic tweets sentimental analysis using machine learning,'' in Proc. Int. Conf. Ind., Eng. Appl. Appl.Intell. Syst., Jun. 2017, pp. 602–610.

[17] D. Gamal, M. Alfonse, E.-S. M. El-Horbaty, and A.-B. M. Salem, ''Twitterbenchmark dataset for arabic sentiment analysis,'' Int. J. Modern Edu.Comput. Sci., vol. 11, no. 1, pp. 33–38, Jan. 2019.

[18] A. Kumar and G. Garg, ''Sentiment analysis of multimodal Twitter data,''Multimedia Tools Appl., vol. 78, no. 17, pp. 24103–24119, Sep. 2019.

[19] K. Sailunaz and R. Alhajj, ''Emotion and sentiment analysis from Twittertext,'' J. Comput. Sci., vol. 36, Sep. 2019, Art. no. 101003.

[20] V. Kalra and R. Aggarwal, ''Importance of text data preprocessing &implementation in RapidMiner,'' in Proc.1st Int. Conf. Inf. Technol.

Knowl. Manage., vol. 14, Jan. 2018, pp. 71–75.

[21] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas,''Short text classification in Twitter to improve information filtering,'' in

Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR), 2010,pp. 841–842.

[22] Scikit Learn. Scikit-Learn Feature Extraction With Countvectorizer.

Accessed: Apr. 5, 2019. [Online]. Available: https://scikit-learn.org/

stable/modules/generated/sklearn.feature_extraction.text.Count/

[23] Scikit Learn. Scikit-Learn Feature Extraction With TF/IDF.

Accessed: Apr. 5, 2019. [Online]. Available:https://scikitlearn.org/stable/modules/generated/sklearn.feature_extraction.text.Tfidf/

[24] P. Routray, C. K. Swain, and S. P. Mishra, ''A survey on sentiment analysis,'' Int. J. Comput. Appl., vol. 76, no. 10, pp. 1–8, Aug. 2013.

[25] A. Harb, M. Plantié, G. Dray, M. Roche, F. Trousset, and P. Poncelet, ''Webopinion mining: How to extract opinions from blogs?'' in Proc. 5th Int.

Conf. Soft Comput. Transdisciplinary Sci. Technol. (CSTST). New York, NY, USA: Association for Computing Machinery, 2008. pp. 211–217.

[26] B. Pang, L. Lee, and S. Vaithyanathan, ''Thumbs up? Sentiment classi-fication using machine learning techniques,'' EMNLP, vol. 10, pp. 1–9,Jun. 2002.