International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 12 | Dec - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Emotion Recognition by Textual Tweets Classification Using Voting Classifier

[1]CH.Saikrishna, [2]D.Sathwik reddy, [3]G.Ajith, [4]DR.khushbu doulani

1,2,3 UG Scholars, 4 Associate Professor

1,2,3,4 Department of Computer Science & Engineering,

1,2,3,4 Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India

ABSTRACT:

The proliferation of social media platforms has revolutionized sentiment analysis, transforming it into a pivotal research area. Twitter, known for its brevity and real-time engagement, serves as a significant platform for analyzing public opinions on diverse topics. Sentiment analysis leverages machine learning to automatically identify emotions, providing valuable insights into societal trends, government policies, and market dynamics. This study explores the performance of seven machine learning classifiers for emotion detection in tweets, culminating in a novel voting classifier (LR-SGD) that utilizes TF-IDF features. The proposed model achieved a notable accuracy of 79% and an F1 score of 81%.

1 INTRODUCTION:

Emotion recognition, a subset of artificial intelligence encompassing pattern recognition and natural language processing, plays a critical role across industries. Platforms like Twitter generate vast datasets that demand efficient tools for extracting meaningful insights. For instance, during the COVID-19 pandemic, misinformation on social media highlighted the necessity of advanced sentiment analysis to filter noise and misinformation.

Twitter’s concise format makes it an ideal candidate for analyzing user sentiment, but the platform’s unstructured data requires sophisticated techniques for text classification. Businesses leverage Twitter to understand consumer behavior, promote products, and receive feedback. This study focuses on employing machine learning models to automate the sentiment classification process, allowing organizations to make data-driven decisions more effectively.

1.1 OBJECTIVE:

The research uses a Twitter dataset obtained from Kaggle, which undergoes preprocessing to remove noise. The data is divided into training (70%) and testing (30%) sets. Feature engineering techniques are applied, followed by the training and evaluation of multiple machine learning classifiers.

1.2 SCOPE OF THE WORK:

The study investigates emotion recognition through the classification of tweets using TF and TF-IDF representations. Key contributions include:

- Comparison of machine learning classifiers such as Support Vector Machine (SVM), Decision Tree Classifier (DTC), Naive Bayes (NB), Random Forest (RF), Gradient Boosting Machine (GBM), and Logistic Regression (LR).

- Introduction of a voting classifier (LR-SGD) that demonstrated superior performance.

1.3 PROBLEM STATEMENT:

The research examines various supervised machine learning models for tweet classification. The methodology employs classifiers such as SVM, NB, RF, DTC, GBM, LR, and a voting classifier (LR-SGD). Key highlights of the models include:

- Random Forest (RF): Utilizes decision trees to enhance classification accuracy through majority voting.

- Support Vector Machine (SVM): Identifies optimal hyperplanes for effective class separation.

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 12 | Dec - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

- Naive Bayes (NB): A probabilistic classifier based on Bayes' Theorem, suitable for text classification.

- Decision Tree (DTC): Constructs a tree structure to make decisions, often utilizing Gini Index or Information Gain for attribute selection.

1.4 LIMITATIONS OF EXISTING SYSTEM:

Previous studies applied Naive Bayes and SVM for sentiment analysis but faced challenges due to noisy and imbalanced data. For instance, misinformation during the COVID-19 infodemic underscored the limitations of traditional approaches. Additionally:

- Random Forest can produce inaccurate results with unrepresentative bootstrap samples.

- Gradient Boosting is sensitive to noisy data, leading to potential overfitting.

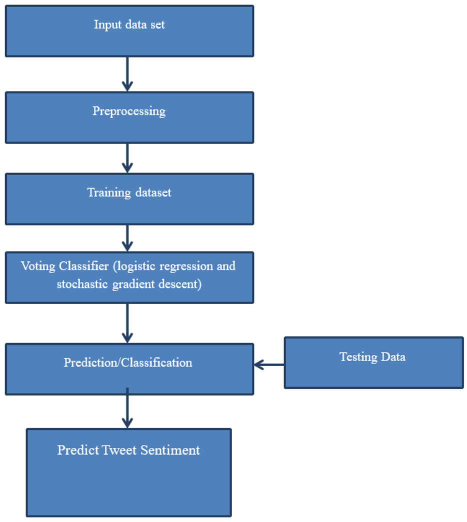
1.4.1 Existing System Disadvantages:

Previous studies applied Naive Bayes and SVM for sentiment analysis but faced challenges due to noisy and imbalanced data. For instance, misinformation during the COVID-19 infodemic underscored the limitations of traditional approaches. Additionally:

- Random Forest can produce inaccurate results with unrepresentative bootstrap samples.

- Gradient Boosting is sensitive to noisy data, leading to potential overfitting.

1.5 SYSTEM ARCHITECTURE:



1.6 PROPOSED SYSTEM

The proposed voting classifier (LR-SGD) integrates logistic regression and stochastic gradient descent to achieve optimal performance. Tested across binary and multi-class datasets, the system consistently demonstrated robustness with 79% accuracy and an 81% F1 score.

1.6.1 PROPOSED SYSTEM ADVANTAGES:

1) Easy to predict.
2) It monitors the accuracy.

2 PROJECT DESCRIPTION:

2.1 GENERAL:

In this chapter, various supervised machine learning approaches are used. This section provides a general description of these approaches.

2.2 METHODOLOGIES:

- Data Collection
- Dataset
- Data Preparation
- Model Selection
- Analyze and Prediction
- Accuracy on test set

- Saving the Trained Model

DATASET:

The dataset consists of 10314 individual data. There are 3 columns in the dataset, which are described below

Importing the Necessary Libraries:

The rise of user-generated content on social media has made opinion mining increasingly challenging. Twitter, as a microblogging platform, is widely used to gather opinions on products, trends, and political topics. Sentiment analysis is a technique used to assess the attitudes, emotions, and opinions of individuals on various subjects, and it can be applied to tweets to understand public sentiment on news, policies, social movements, and personalities. Machine Learning models enable opinion mining without the need for manual reading of individual posts.

RETRIEVING EMOTIONS:

The process begins by collecting images and their corresponding labels. The images will then be resized to a consistent dimension of (180, 180) to standardize them for recognition. After resizing, the images will be converted into numpy arrays for further processing and input into the model.

Splitting the Dataset:

The dataset will be divided into training and testing sets, with 80% of the data allocated for training and 20% for testing.

Building the Model:

Training the Model and Plotting Performance Graphs:

The model will be compiled and trained using the fit function, with a batch size of 10. After training, graphs for accuracy and loss will be generated to evaluate the model's performance. The model achieved an average validation accuracy of 93.00% and a training accuracy of 91.00%.

Accuracy on Test Set:

The model achieved an accuracy of 91.00% on the test set.

Saving the Trained Model:

After the model has been successfully trained and tested, it will be saved for deployment in a production environment. This can be done by saving the model as a .h5 or .pkl file using a library like pickle. Ensure that pickle is installed in the environment. Once installed, the module can be imported, and the model can be saved as a .h5 file for later use.

2.3 TECHNIQUES OR ALGORITHM USED:

Learning Rate Schedules and Adaptive Learning Rate Methods for Deep Learning.

Adaptive learning rate methods and predefined schedules such as time-based decay and exponential decay are used to improve the training of convolutional neural networks (CNNs). These methods adjust learning rates dynamically, optimizing model performance during training.

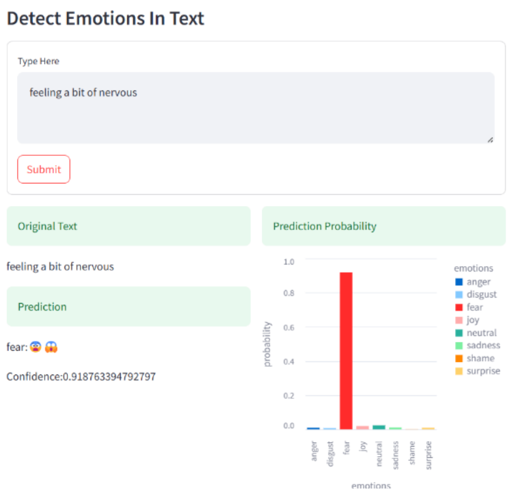
Learning Rate Schedules:

Learning rate schedules are designed to adjust the learning rate during training, typically by gradually decreasing it according to a set plan. Common types of learning rate schedules include time-based decay, step decay, and exponential decay. To demonstrate this, I train a CNN on CIFAR-10 using the stochastic gradient descent (SGD) optimization algorithm, experimenting with different learning rate schedules to compare their effectiveness.

3 RESULT:

The text that has been entered which will wwe there could be enttred and all they could the each an every word that has been recognazied and every thing will be shown for who much content will be as percentage wise gor grouch.

4. SNAPSHOTS OF THE PROJECT:



5.FUTURE ENHANCEMENTS:

The proposed model has been validated on two additional datasets with strong results. Future work will focus on comparing more feature engineering techniques and exploring additional ensemble model combinations to enhance performance. Additionally, new methods will be explored to address sarcastic comments.

6.CONCLUSION:

This study highlights the effectiveness of combining LR and SGD for emotion recognition. By employing TF-IDF features and testing various machine learning classifiers, the proposed system delivers robust results. The findings underscore the potential of ensemble models in enhancing sentiment analysis accuracy and applicability.

.7.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," *Pattern Recognit. 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 smear images," *IEEE Access*, vol. 8, pp. 9378-93792, 2020.