iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** III **Month of publication:** March 2024

DOI: https://doi.org/10.22214/ijraset.2024.59237

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

Exploring Stress Detection in Tweets: A Comparative Business Analysis of Classification Algorithms

Akhil Joseph[2], Ms. Naveena Tresa Joseph[1]
[1]*IT Department, AJCE*
[2]*Assistant Professor, IT Department, AJCE*

Abstract: ***In today's digitally-driven landscape, social media platforms such as Twitter, Facebook, and WordPress have emerged as critical conduits for public discourse, resulting in an overwhelming influx of unprocessed data, particularly on Twitter. Coping with the sheer volume of this data presents a significant challenge. To address this, sentiment analysis, a technique that categorizes sentiments into positive, negative, or neutral, offers a promising solution. This study delves into three main approaches for sentiment analysis: Machine learning-based methods, Sentiment lexicon-based approaches, and Hybrid methods. Its primary research objectives revolve around the identification of suitable algorithms and metrics for evaluating the performance of Machine Learning Classifiers. Additionally, the study aims to compare these metrics with respect to the dataset size, gauging their impact on the most appropriate sentiment analysis algorithm.The research methodology applied is experimental, entailing a rigorous assessment of algorithms using carefully selected metrics. The results of this investigation spotlight Naïve Bayes, Random Forest, XGBoost, and CNN-LSTM as the leading machine learning algorithms under consideration. These algorithms are assessed based on key performance metrics such as precision, accuracy, F1 score, and recall. Notably, the CNN-LSTM model emerges as the optimal choice for sentiment analysis of Twitter data within the specified dataset size, achieving a remarkable accuracy rate of 88%.In summary, this research successfully pinpoints the most suitable algorithm for sentiment analysis of Twitter data, especially in the context of dataset size. The CNN- LSTM model showcases its effectiveness, serving as a robust tool for sentiment analysis and delivering an impressive accuracy rate. This study significantly enhances our comprehension of public sentiment on Twitter, providing valuable insights into the ever-evolving realm of digital discourse and the analysis of vast unprocessed data***

Keywords: Sentimental Analysis workflow, Naïve Bayes Classifier predictions, Naïve Bayes Classifier Metrics, Random Forest Classifier predictions

I. INTRODUCTION TO DOMAIN

In the realm of digital communication, tweet-level stress detection has emerged as a compelling field, aiming to uncover and comprehend stress levels conveyed through tweets. With social media, particularly Twitter, serving as a prominent platform for individuals to express their emotions and experiences, this domain leverages natural language processing and mental health analysis to identify stress indicators.

Stress, a significant factor in mental health, necessitates continuous monitoring and management. Tweet-level stress detection systems offer real-time insights by analyzing language, sentiment, and contextual cues in tweets. They have potential applications in stress management, mental health monitoring, and crisis intervention. These systems rely on intricate models and algorithms to categorize tweets based on their stress levels. As social media continues to play a vital role in our lives, the development of accurate and ethical tweet-level stress detection systems represents an exciting intersection of technology and mental health, promising to enhance our understanding of stressors and offer timely support.

II. DEFINITIONS

A. Sentimental Analysis workflow

Sentiment analysis, also known as opinion mining, is the process of computationally determining the sentiment or opinion expressed in a piece of text. The sentiment analysis workflow outlines the steps involved in this process, which typically include data preprocessing, feature extraction, model training, prediction, and evaluation

B. Naïve Bayes Classifier prediction

Naïve Bayes is a simple probabilistic classifier based on Bayes' theorem with the "naive" assumption of independence between features. It's commonly used for classification tasks, including sentiment analysis. Despite its simplicity, Naïve Bayes often performs well and is computationally efficient.

C. Naïve Bayes Classifier Metrics

Metrics are used to evaluate the performance of a classifier, such as Naïve Bayes, in sentiment analysis. Common metrics include accuracy, precision, recall, F1-score, and confusion matrix. These metrics provide insights into the classifier's ability to correctly classify sentiments and identify areas for improvement.

D. Random Forest Classifier Predictions

Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes for classification tasks or the mean prediction for regression tasks. It's known for its high accuracy, robustness, and ability to handle large datasets with high dimensionality.

III. RELEVANCE OF THE TOPIC

Labeled data plays a pivotal role in training models and evaluating their accuracy by providing explicit sentiment labels for each text. On the other hand, unlabeled data serves multifaceted purposes such as feature extraction, data augmentation, and domain adaptation. By extracting valuable features and generating synthetic examples, unlabeled data enriches model training and enhances its robustness. Moreover, unlabeled data facilitates semi-supervised and unsupervised learning approaches, enabling models to learn from large, diverse datasets without exhaustive labeling efforts. This synergy between labeled and unlabeled data fosters more accurate, adaptable, and scalable sentiment analysis models across different domains and languages.

IV. IMPLEMENTATION DETAILS

The process of implementing classification based on both labeled and unlabeled data involves several key steps. Firstly, data preparation entails obtaining a labeled dataset containing text samples and their corresponding class labels, as well as creating a separate dataset with unlabeled text samples for semi-supervised learning. Preprocessing the text data by removing stopwords, punctuation, and stemming or lemmatizing words is essential for standardization. Next, feature extraction transforms the text data into numerical vectors, commonly achieved through techniques like TF-IDF or word embeddings such as Word2Vec or GloVe. For algorithms like CNN-LSTM, text is converted into sequences of word embeddings. The labeled data is then used for supervised learning, where the dataset is split into training and validation sets. Each algorithm, including Naive Bayes, Random Forest, XGBoost, and CNN-LSTM, is trained on the labeled training data using their respective methods. Model evaluation follows, assessing performance using metrics like accuracy, precision, recall, F1 score, and confusion matrices on the validation set. Moving to semi-supervised learning with unlabeled data, the trained models are applied to predict labels for the unlabeled dataset. Techniques such as Naive Bayes, Random Forest, XGBoost, and CNN-LSTM are employed accordingly to assign labels or make predictions. Optionally, active learning techniques may be utilized to label some unlabeled data based on model uncertainty to enhance accuracy. Finally, the chosen classification model, typically determined by its performance on the validation set, can be deployed for further use on new, unseen data. This comprehensive approach provides a structured framework for classifying labeled and unlabeled data using a variety of algorithms, with the selection of the most suitable algorithm dependent on the data's characteristics and the specific classification task at hand. Experimentation and parameter tuning are crucial for achieving optimal classification results

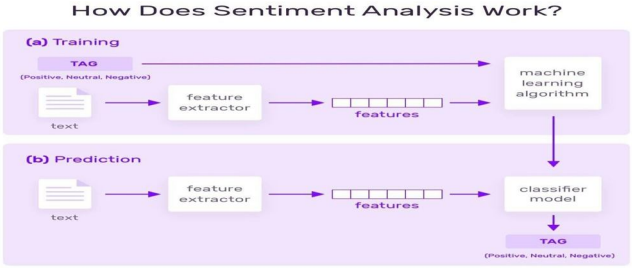


Fig 3.1 Sentimental analysis workflow

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

The above diagram shows in detail about the way sentimental analysis works by first starting by training and then following with prediction. The various Machine learning algorithms such as Naïve Bayes, Random Forest, XGBoost and CNN-LSTM are used after feature extraction for further classification of labelled or unlabeled data.After the data is classified, it is then trained for being a better model. The model after being trained is then evaluated to understand the model performance. Once the model performance has achieved good precision and accuracy values, the unlabeled tweet is then evaluated using the trained model and the tweets are labelled accordingly to the way the model was trained. After the tweets being the labelled through the trained model, a brief conclusion is drawn from the data. Then we conclude the emotions, sentiment and opinions of people for the particular dataset.

A. Block Diagram

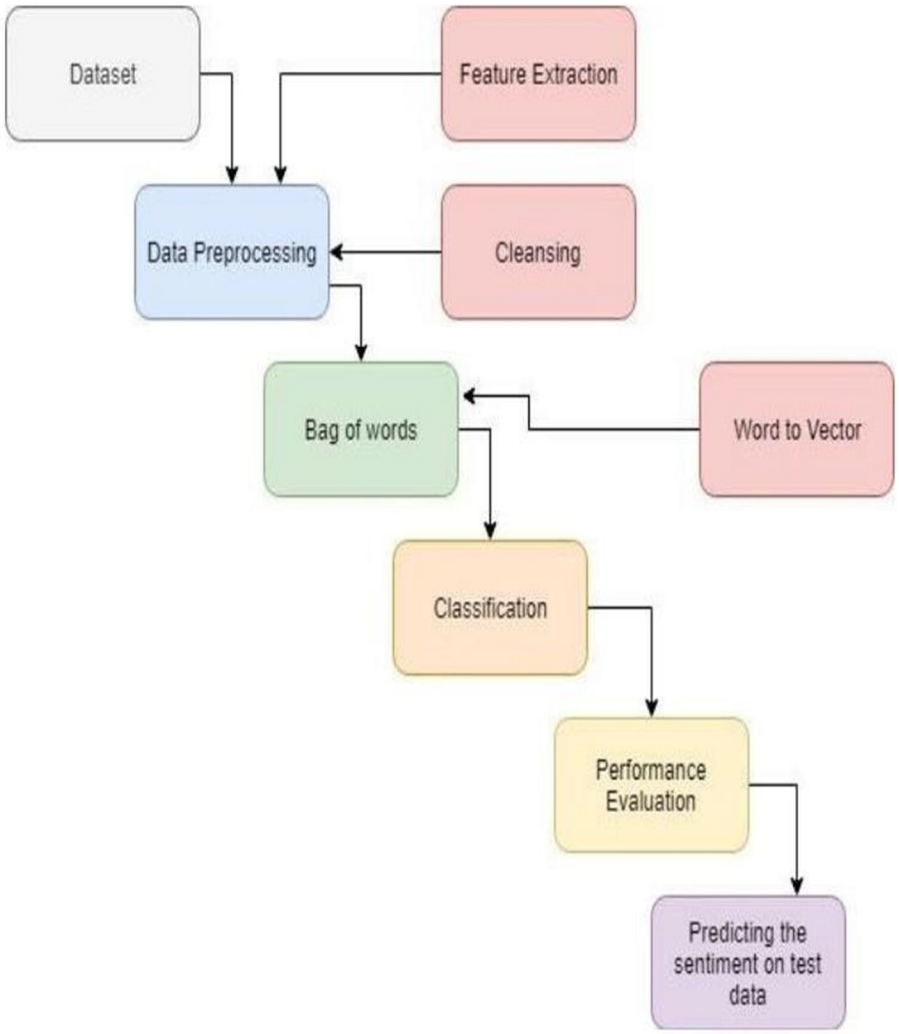


Fig 3.2 Block Diagram

V. PERFORMANCE ANALYSIS

A. Comparative Study

The comparative study involves the comparisons of various algorithm used for classification of labelled and unlabeled data. The algorithms considered for sentimental analysis for classification of data tweets includes Naïve Bayes, Random Forest, XGBoost Algorithm and CNN-LSTM. Here, we focus on the Accuracy rate, Precision Rate, F1 and Recall.

B. Performance Analysis

An analysis is carried out by comparing various algorithm such as Naïve Bayes, Random Forest, XGBoost Algorithm and CNN-LSTM.

1) CNN-LSTM

Tweet	Category
Does @macleansmag still believe that Ms. Angel...	neutral
Varoufakis may have some expertise in game the...	positive
Alexis Tsipras to Angela Merkel - early Monday...	neutral
On another note, it seems Greek PM Tsipras mar...	neutral
By Michael Nienaber BERLIN, July 10 (Reuters) ...	neutral
Angela Merkel: 'Deny Marriage To Gay Couples':...	neutral
When you spend the day being Angela Merkel to ...	negative
@NiamhPuirseil Nicola S, Theresa May, Liz Kend...	neutral
Young Palestinian asylum seeker breaks down wh...	negative
Moving video illustrates Europe's refugee prob...	negative
Angela Merkel, explaining to a crying child th...	negative
Perhaps Prime Minister's Questions would be be...	neutral
wow, angela merkel tells sobbing asylum seeker...	neutral
This has just brought a tear to my eye: Angela...	negative
Not Available	neutral
Angela Merkel tells sobbing teenage refugee sl...	neutral
may not agree with Angela Merkel on a lot, the...	neutral
The best politicians tell people the truth, no...	negative
Angela Merkel and a Palestinian asylum seeker:...	negative
Not Available	neutral
Angela Merkel is a true leader . She has to te...	neutral

Fig 4.8 CNN-LSTM classifier metrics

```
Learing time 0.2639958858489995
Predicting time 0.07999992370605469s
=================== Results ===================
            Negative     Neutral     Positive
F1       [0.38949672 0.45072993 0.71369782]
Precision[0.45408163 0.48431373 0.65906623]
Recall   [0.34099617 0.42150171 0.77820513]
Accuracy 0.5795943454210203
================================================
```

Fig 4.2 Naïve Bayes Classifier metrics

VI. ADVANTAGES/LIMITATIONS OF THE PROPOSED METHOD

A. Advantage

The CNN-LSTM algorithm has garnered widespread acclaim for its consistent delivery of superior accuracy and precision across a spectrum of studies. Its efficacy stems from the fusion of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, which collectively offer a potent solution for extracting features while preserving long-range dependencies inherent in sequential data.

This integration not only facilitates efficient feature extraction but also enables the model to generalize effectively to unseen data by capturing intricate patterns and relationships. Furthermore, the modular design of CNNs enables parallel processing, enhancing the algorithm's scalability and speed, particularly beneficial for handling large-scale datasets. Beyond its technical prowess, the CNN-LSTM algorithm's versatility shines through its adaptability to diverse tasks spanning various domains. From image captioning to sentiment analysis and beyond, its ability to seamlessly navigate through different sequential data applications underscores its relevance and effectiveness in addressing real-world challenges. As research continues to unveil new possibilities and refine existing methodologies, the CNN-LSTM algorithm stands poised as a cornerstone in the realm of deep learning, offering unparalleled potential for innovation and advancement.

B. Limitations

The CNN-LSTM architecture has proven to be a robust deep learning model for a wide range of sequential data tasks. However, it also presents inherent limitations that can affect its performance and practicality. These drawbacks include the structural complexity resulting from the fusion of CNNs and LSTMs, making the model architecture intricate to manage. Additionally, training CNN-LSTM models often incurs substantial computational overhead and time investment due to their complexity, which can pose challenges for scalability. Moreover, the architecture's high memory requirements, particularly noticeable with large datasets or extended sequences, may limit its deployment in resource-constrained environments. Acknowledging these limitations is crucial for informed decision-making and the development of strategies to optimize the model's performance in specific applications.

REFERENCES

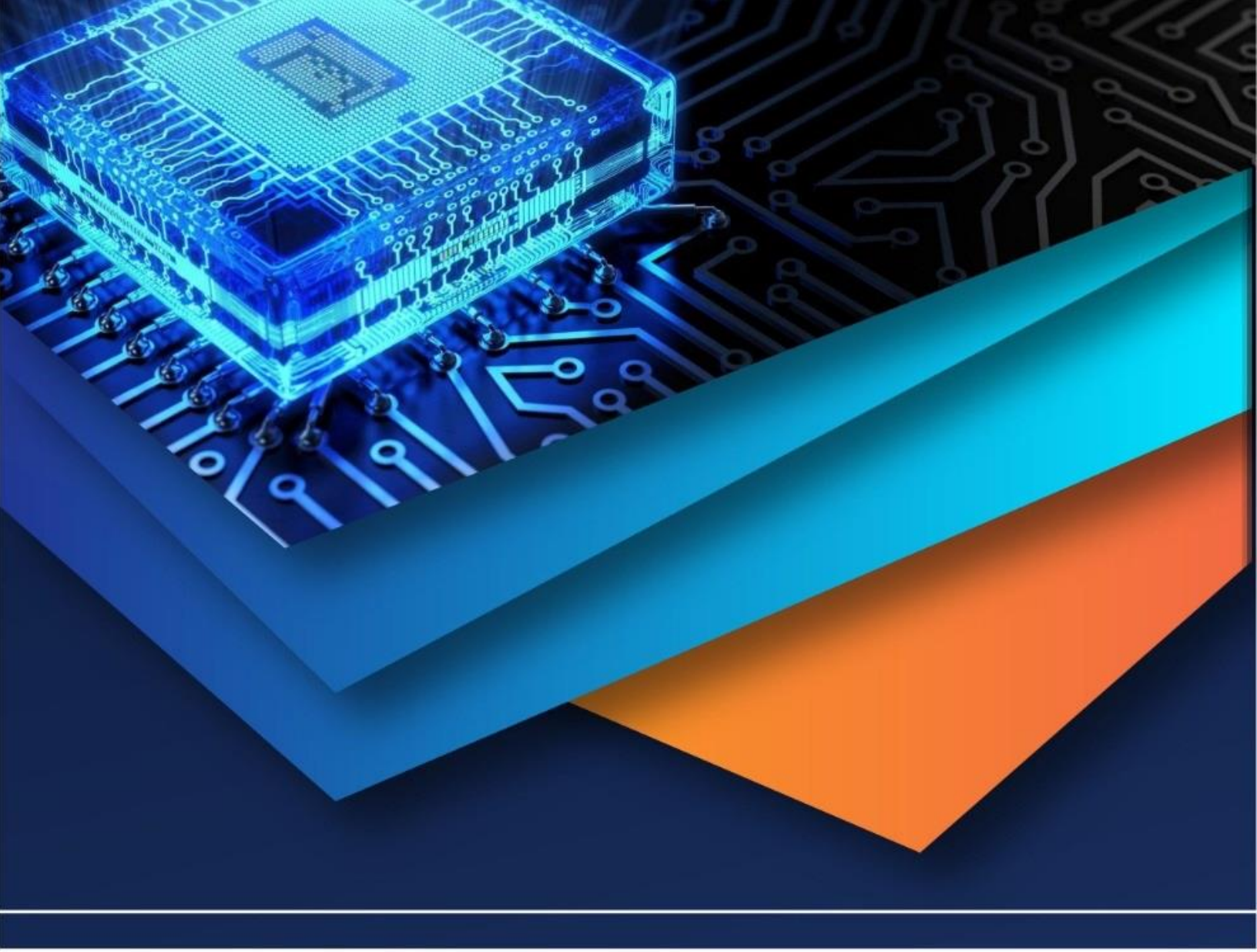
[1] Kundan Reddy Manda,2019, Sentiment Analysis of Twitter Data Using Machine Learning and Deep Learning Methods, Faculty of Computing, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden

[2] Meylan Wongkar , Apriandy Angdresey, 2019, Sentiment Analysis Using Naive Bayes, Algorithm Of The Data CrawlerTwitter,https://www.researchgate.net/publication/339176682R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[3] Muhammad Umer, Imran Ashraf, Arif Mehmood, Saru Kumari, Saleem Ullah, Gyu Sang Choi,2020, Sentiment analysis of tweets using a unified convolutional neural network-longshort-termmemory networkmodel, https://www.researchgate.net/publication/345723486I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350

[4] Padmanayana, Varsha, Bhavya K, 2021, Stock Market Prediction Using Twitter Sentiment analysis, Department of Computer Science, Srinivas Institute ofTechnology, Mangalore, Karnataka, India.

[5] Wan Setiawan, Agung Mulyo Widodo , Mosiur Rahaman , Tugiman, Muhammad Abdullah Hadi , Nizirwan Anwar , Muhammad Bahrul Ulum , Erry Yudhya Mulyani , and Nixon Erzed ,2022, Utilizing Random Forest Algorithm for Sentiment Prediction Based on Twitter Data. https://doi.org/10.2991/978-94-6463-084- 8_37



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY