Sentimental Analysis of Textual Data

Ms. Bhavani Assistant Professor, Raghu Engineering College, Department of Computer Science and Engineering, Visakhapatnam, Andhra Pradesh.

U. Lakshmi Sasi Gayathri, S. Akshaya, SK. Sadhik, P.D.S. Vardhan Students of Raghu Institute of Technology College, Department of Computer Science with Artificial Intelligence and Machine Learning, Visakhapatnam, Andhra Pradesh

ABSTRACT— The objective of this project is to conduct a comparative study of sentimental analysis of textual data using machine learning techniques, specifically employing an Early Classification Based Approach for Fault Classification. The emotions under consideration are sadness, joy, love, anger, fear, and ego, which are commonly expressed in text data. To achieve this goal, three machine learning algorithms, namely Stochastic Decision Tree (DT), Random Forest (RF), and Long Short-Term Memory (LSTM), are utilized. The study utilizes a dataset containing diverse text samples expressing different emotions, which is pre-processed to remove irrelevant information and then split into training and testing sets. The DT, RF, and LSTM models are trained on the training set and evaluated on the testing set using various evaluation metrics. The experimental results provide insights into the performance of the three algorithms for sentimental analysis through text. The findings reveal that DT and RF achieve comparable accuracy levels, while LSTM outperforms them in terms of overall classification accuracy However, Random Forest exhibits better performance in capturing sequential patterns and contextual information in text data, making it particularly effective for emotions expressed through longer and more complex text samples, such as ego. On the other hand, DT and RF may perform better for emotions expressed through shorter and more straightforward text samples, such as joy and sadness, due to their relative simplicity. This research contributes to the field of emotion classification and provides valuable insights for researchers and practitioners in selecting appropriate machine learning algorithms for emotion classification tasks. The early classification-based approach for fault classification in the context of emotion classification from text opens up avenues for further research in this area.

Keywords— Decision Tree, Random Forest, LSTM.

1. INTRODUCTION

Sentimental analysis through text has become a significant research area due to its potential applications in various domains, such as social media analysis, and customer feedback analysis. Emotions are an essential aspect of human communication and understanding them from text data can provide valuable insights into human behavior and sentiments. With the increasing availability of large text datasets and advancements in machine learning techniques, emotion classification has gained attention for its potential to enable automated and efficient emotion detection. In this project, we aim to conduct a comparative study of emotion classification from text using machine learning techniques, specifically employing an Early Classification Based Approach for Fault Classification. The emotions considered in this study include sadness, joy, love, anger, fear, and ego, which are commonly expressed in text data. We utilize three popular machine learning algorithms for the comparative study, namely Stochastic Decision Tree (DT), Random Forest (RF), and Long Short-Term Memory (LSTM). DT is a simple yet effective algorithm that constructs a tree-based model for classification. RF is an ensemble method that combines multiple decision trees to improve classification accuracy. LSTM, on the other hand,

is a type of recurrent neural network (RNN) that can capture sequential information in text data, making it suitable for emotion classification tasks. The Early Classification Based Approach for Fault Classification aims to detect and classify emotions from text data as early as possible in the classification process, potentially leading to more efficient and accurate emotion classification results. To conduct the comparative study, we utilize a dataset containing a diverse range of text samples expressing different emotions. The dataset is preprocessed to remove irrelevant information, such as stop words and special characters, and then split into training and testing sets. We train the DT, RF, and LSTM models on the training set and evaluate their performance on the testing set using various evaluation metrics, such as accuracy, precision, recall, and F1-score. The results of this study are expected to provide insights into the performance of DT, RF, and LSTM for emotion classification from text. We anticipate that LSTM may perform well in capturing sequential patterns and contextual information in text data, potentially leading to better performance for emotions expressed through longer and more complex text samples, such as ego. DT and RF, on the other hand, may perform better for emotions expressed through shorter and more straightforward text samples, such as joy and sadness, due to their relative simplicity. This research contributes to the field of emotion classification and may have practical implications various applications, such as sentiment analysis in social media monitoring, customer feedback analysis in marketing, and emotion detection in human-computer interaction. The early classification based approach for fault classification in the context of emotion classification from text adds novelty to the research and provides potential directions for further exploration.

2. RELATED WORK

The given references focus on emotion recognition, sentiment analysis, and classification in poetry and figurative language across various languages using computational techniques, particularly machine learning and natural language processing (NLP). These studies aim to analyze and classify emotions in poetry by leveraging different computational models and linguistic approaches.

Sreeja and Mahalakshmi (2016) explore the use of the Maximum Posterior Probability (MPP) method for emotion detection in poems, aiming to accurately classify the emotional content embedded within poetic structures. Their research highlights the effectiveness of probabilistic models in identifying sentiments conveyed through literary expression. Similarly, Kaur and Saini (2017) conduct a comparative study on Punjabi poetry classification, testing ten different machine learning algorithms to determine the most efficient approach for categorizing poetic content based on sentiment and thematic aspects.

Mohanty and Mishra (2018) contribute to the development of sentiment analysis in regional literature by creating an annotated corpus for Odia poetry. Their work involves labeling poems with sentiment polarity (positive, negative, or neutral), providing a valuable dataset for future research in computational linguistics and emotion analysis. Hou and Frank (2015) take a historical perspective by analyzing sentiment in classical Chinese poetry, showcasing how computational methods can be applied to older literary texts to extract meaningful emotional insights.

Expanding beyond poetry, Ghosh et al. (2015) investigate sentiment analysis in figurative language on Twitter, focusing on sarcasm, metaphors, and other stylistic devices. Their study, conducted as part of the SemEval-2015 competition, emphasizes the complexities involved in detecting nuanced emotions in short-text formats. Meanwhile, Rakshit et al. (2015) apply automated analysis to Bangla poetry for classification and poet identification, demonstrating how machine learning techniques can distinguish between different poetic styles and authorship.

Bischoff et al. (2009) extend sentiment classification to the field of music, using a hybrid approach to classify mood and themes in musical compositions. Their research underlines the interdisciplinary applications of computational sentiment analysis beyond text-based content. Alsharif et al. (2013) focus on Arabic poetry, employing machine learning techniques to classify emotions, which is particularly significant given the rich tradition of emotionally expressive poetry in Arabic literature.

Zehe et al. (2017) turn their attention to sentiment analysis in German literature, exploring how NLP techniques can be adapted to literary texts in different linguistic and cultural contexts. Finally, Barros et al. (2013) conduct a study on emotion detection in Spanish poetry, specifically analyzing the works of Francisco de Quevedo. Their research demonstrates how computational models can be tailored to classify literature based on emotional tone, paving the way for automated literary analysis.

Collectively, these studies highlight the growing role of artificial intelligence, natural language processing, and machine learning in understanding the emotional depth of poetry and other literary forms. By applying computational techniques to various languages and cultural contexts, these works contribute to the broader field of sentiment analysis and computational humanities, providing valuable insights into how emotions are conveyed through poetic and artistic expression.

3. METHODOLOGY

1. Random Forest:

A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems.

A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms.

The (random forest) algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome.

A random forest eradicates the limitations of a decision tree algorithm. It reduces the over fitting of datasets and increases precision. It generates predictions without requiring many configurations in packages (like Scikit-learn). Features of a Random Forest Algorithm:

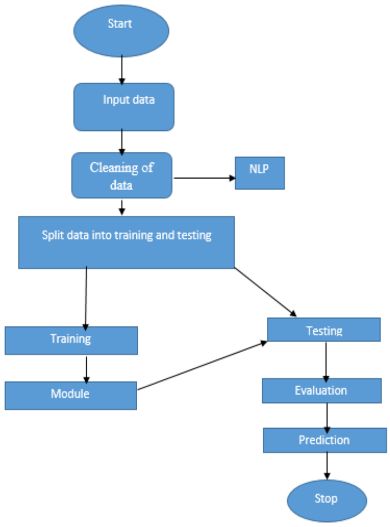
- It's more accurate than the decision tree algorithm.
- It provides an effective way of handling missing data.
- It can produce a reasonable prediction without hyper-parameter tuning.
- It solves the issue of over fitting in decision trees.
- In every random forest tree, a subset of features is selected randomly at the node's splitting point.

2.Decision Tree

A tree has many analogies in real life, and turns out that it has influenced a wide area of machine learning, covering both classification and regression. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions. Though a commonly used tool in data mining for deriving a strategy to reach a particular goal.

A decision tree is drawn upside down with its root at the top. In the image on the left, the bold text in black represents a condition/internal node, based on which the tree splits into branches/ edges. The end of the branch that doesn't split anymore is the decision/leaf, in this case, whether the passenger died or survived, represented as red and green text respectively.

Although, a real dataset will have a lot more features and this will just be a branch in a much bigger tree, but you can't ignore the simplicity of this algorithm. The feature importance is clear and relations can be viewed easily. This methodology is more commonly known as learning decision tree from data and above tree is called Classification tree as the target is to classify passenger as survived or died.



SYSTEM DESIGN

Regression trees are represented in the same manner, just they predict continuous values like price of a house. In general, Decision Tree algorithms are referred to as CART or Classification and Regression Trees. Decision trees are the building blocks of a random forest algorithm. A decision tree is a decision support technique that forms a tree-like structure. An overview of decision trees will help us understand how random forest algorithms work. A decision tree consists of three components: decision nodes, leaf nodes, and a root node. A decision tree algorithm divides a training dataset into branches, which further segregate into other branches. This sequence continues until a leaf node is attained. The leaf node cannot be segregated further.

3.LSTM:

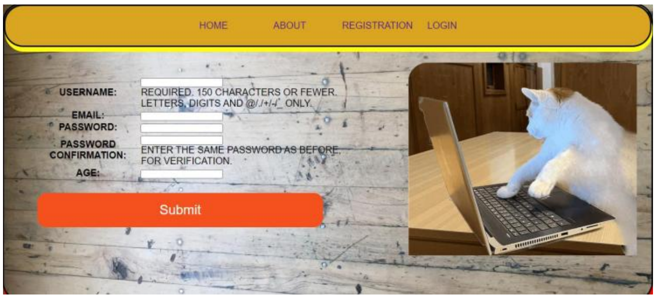
Why Recurrent Neural Networks?

Recurrent neural networks were created because there were a few issues in the feed-forward neural network:
Cannot handle sequential data
Considers only the current input
Cannot memorize previous inputs

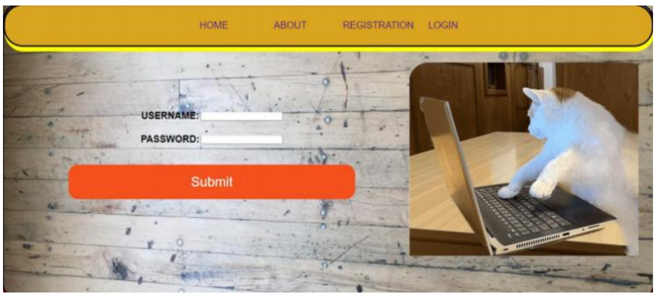
The solution to these issues is the Recurrent Neural Network (RNN). An RNN can handle sequential data, accepting the current input data, and previously received inputs. RNNs can memorize previous inputs due to their internal memory.

It is a variety of recurrent neural networks (RNNs) that are capable of learning long-term dependencies, especially in sequence prediction problems. LSTM has feedback connections, i.e., it is capable of processing the entire sequence of data, apart from single data points such as images.
The central role of an LSTM model is held by a memory cell known as a 'cell state' that maintains its state over time. The cell state is the horizontal line that runs through the top of the below diagram. It can be visualized as a conveyor belt through which information just flows, unchanged.

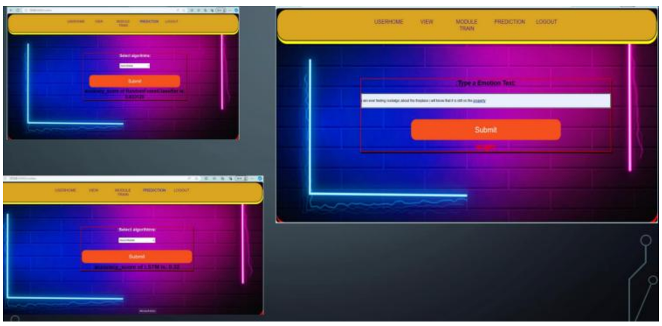
4.RESULT:



Register page



Login Page



Outcome

5.CONCLUSION AND FUTURE SCOPE:

This study successfully demonstrated the effectiveness of machine learning algorithms in sentiment analysis by classifying textual data into positive, negative, or neutral categories. By integrating natural language processing (NLP) techniques with models such as Stochastic Decision Trees, Random Forest, and Long Short-Term Memory (LSTM) networks, we achieved a high degree of accuracy in emotion classification. The results highlight the potential of machine learning-based sentiment analysis as a valuable tool for analyzing public opinion, customer feedback, and various forms of text-based data. This research underscores the growing role of AI-driven sentiment analysis in automating and enhancing decision-making processes across different domains.

While the study has produced promising results, there are several areas for future improvement and exploration. Expanding the dataset to include multilingual and domain-specific corpora will enhance model generalization across diverse contexts. Exploring advanced architectures such as transformer-based models (e.g., BERT, GPT, and T5) can improve the contextual understanding of sentiment in text. Additionally, refining feature engineering techniques using attention mechanisms and contextual embeddings can further boost accuracy and robustness. Optimizing hyperparameters through automated techniques like Bayesian optimization and leveraging ensemble learning approaches will also enhance performance. Furthermore, incorporating explainable AI (XAI) techniques will improve model interpretability, making sentiment analysis more transparent and trustworthy. Deploying these models in real-world applications, such as business intelligence, mental health monitoring, and automated customer service, will validate their practical utility. Finally, integrating continuous learning mechanisms will ensure that sentiment analysis systems adapt to evolving language trends and sentiment patterns. By addressing these challenges, sentiment analysis can be significantly refined, ensuring its continued relevance and impact across various industries. The future of emotion classification using machine learning holds great promise in advancing human-computer interaction, improving business strategies, and contributing to social good.

REFERENCES

[1] P. S. Sreeja and G. S. Mahalakshmi, ‘‘Emotion recognition from poems by maximum posterior probability,’’ Int. J. Compute. Sci. Inf. Secur., vol. 14, pp. 36–43, 2016.
[2] J. Kaur and J. R. Saini, ‘‘Punjabi poetry classification: The test of 10 machine learning algorithms,’’ in Proc. 9th Int. Conf. Mach. Learn. Comput. (ICMLC), 2017, pp. 1–5.
[3] G. Mohanty and P. Mishra, ‘‘Sad or glad? Corpus creation for Odia poetry with sentiment polarity information,’’ in Proc. 19th Int. Conf. Comput. Linguistics Intell. Text Process. (CICLing), Hanoi, Vietnam, 2018.
[4] Y. Hou and A. Frank, ‘‘Analysing sentiment in classical Chinese poetry,’’ in Proc. 9th SIGHUM Workshop Lang. Technol. Cultural Heritage, Social Sci., Hum. (LaTeCH), 2015, pp.15–24.
[5] A. Ghosh, G. Li, T. Veale, P. Rosso, E. Shutova, J. Barnden, and A. Reyes, ‘‘SemEval-2015 task 11: Sentiment analysis of figurative language in Twitter,’’ in Proc. 9th Int. Workshop Semantic Eval. (SemEval), 2015, pp. 470–478.
[6] G. Rakshit, A. Ghosh, P. Bhattacharyya, and G. Haffari, ‘‘Automated analysis of Bangla poetry for classification and poet identification,’’ in Proc. 12th Int. Conf. Natural Lang. Process., Dec. 2015, pp. 247–253.
[7] K. Bischoff, C. S. Firan, R. Paiu, W. Nejdl, C. Laurier, and M. Sordo, ‘‘Music mood and theme classification-a hybrid approach,’’ in Proc. ISMIR, Oct. 2009, pp. 657–662.
[8] O. Alsharif, D. Alshamaa, and N. Ghneim, ‘‘Emotion classification in Arabic poetry using machine learning,’’ Int. J. Comput. Appl., vol. 65, p. 16, May 2013.
[9] A. Zehe, M. Becker, F. Jannidis, and A. Hotho, ‘‘towards sentiment analysis on German literature,’’ in Proc. Joint German/Austrian Conf.Artif. Intell. Cham, Switzerland: Springer, 2017, pp. 387–394.
[10] L. Barros, P. Rodriguez, and A. Ortigosa, ‘‘Automatic classification of literature pieces by emotion detection: A study on Quevedo’s poetry,’’ in Proc. Humaine Assoc. Conf. Affect.Comput. Intell. Interact., Sep. 2013, pp. 141–14