**Social Media Sentiment Analysis Using Twitter Dataset**

Thanuj R   
*Department of Computer Science & Engineering*

*Amrita School of Computing Bengaluru*

*Amrita Vishwa Vidyapeetham, India*

Bl.en.u4aie22140@bl.students.amrita.edu

Mani Sankar D  
*Department of Computer Science & Engineering*

*Amrita School of Computing Bengaluru*

*Amrita Vishwa Vidyapeetham, India*

Bl.en.u4aie22112@bl.students.amrita.edu

Prem Reddy YS   
*Department of Computer Science & Engineering*

*Amrita School of Computing Bengaluru*

*Amrita Vishwa Vidyapeetham, India*

Bl.en.u4aie22166@bl.students.amrita.edu

*Abstract* — this project is developing an automated sentiment analysis of Twitter data for a chosen theme. It is composed of the following: collecting various Twitter datasets, preprocessing texts, initial data analysis, and feature engineering. The specific sentiment analysis model, which varies from traditional to deep learning techniques, is learned and assessed for precision and efficiency. The goal is to offer a valuable data analytics tool with plug-and-play capabilities and an interactive interface for the purpose of real-time public sentiment analysis.

*Keywords- Sentiment analysis, Social Media, Hyperparameters, feature fusion, lifelong learning, Noisy Data, Supervised learning, Artificial Intelligence.*

# **Introduction**

The prevalence of various social media platforms, especially Twitter, has revolutionized the means through which people express their opinions, share stories, and partake in global discussions from the daily life point of view. The topic of this document is a comprehensive platform for sentiment analysis in terms of its dynamics, which depend on the temporal characteristic of user-generated content on Twitter. The skill of distinguishing and decoding sentiments enclosed in posts is an essential competence for companies, scientists and individuals, who have the task of processing the finer details of public opinion covering various topics, products and events.

While social media and Twitter, in particular, are invaluable archives of the most natural and pristine expressions, they are really amazing sources of getting an understanding of people’s moods. Analyzing this endless and continuously changing data source needs advanced analytical methods, including Natural Language Processing and Machine Learning. This paper recommends a three level system, which starts with the data collection, then goes to the preprocessing, exploratory data analysis, feature engineering, model selection, training and evaluation. The objective is to supply a systematic facility for understanding emotions hence bringing out information on the transient environment of public talk in social media.

The following paragraphs will detail the contents of sections, which expounds on the methodology of the sentiment analysis framework, covering data acquisition processes, in-depth pre-processing steps, the array of modelling techniques, and the possibility of real-time implementation. With this extensive examination, we strive to cast some light on the hitherto neglected topics in the sentiment analysis methods for the Twitter, giving the exact resource needed for the persons analysing public sentiment in the digital era.

# **Literature** **Survey**

The paper explains the use of a lexicon-based approach for analyzing Turkish shares in social media with a similarity around 80% in the classification of the sentiment aspect. [1] The offered paper covers the topic on the efficacy of Bidirectional Encoder Representations From Transformers (BERT) in regard to the sentiment analysis on bulk social media datasets. Leveraging a state of the art language model, is able to handle the varying dimensions of sentiment expressions and also context under the fast dynamic and diverse nature of the social media contents. [2] The paper study looks at the application of sentiment analysis to examine the communication of Indian companies on the social media platforms like Twitter. The instrument does not explicitly refer to assessing social responsible stances and attitudes with the help of sentiment analysis. [3] The paper proposes a multimodal sentiment analysis approach based on a top-layer fusion strategy in social networks that outperforms the baselines on the MVSA-Single dataset. This approach allows the use of multiple models in one, combining different types of data, showing its advantage in the context of sentiment analysis accuracy improvement for social media content. [4] The paper analyzes the impact of sentiments expressed in tweets on the stock market using the VADER model for sentiment analysis. [5] The paper proposes a multilingual sentiment classifier to analyze how Malaysians react on social media during disasters, achieving 0.862 accuracy and 0.864 F1-score. [6] A paper is presented that suggests a distantly supervised lifelong learning scheme for the social media sentiment analysis on the large scale. This approach, therefore, seeks to reconfigure the ways of dealing with a large amount of data from the social media platforms. [7] This paper is about exploring the significance of sentiment analysis and examines existing research on evaluating and comparing relevant data mining techniques. [8] The paper contributes to a growing body of research on sentiment analysis within the context of social media using some different classifiers like support vector machines and random Forest etc. [9] This paper is about Identifying Fake News in Social Media Using Sentiment Analysis using the relations negative sentiment and the likelihood of news being fake. [10]

# **METHODOLOGY**

1. Data Pre-processing(A1) :
   * The data which we used for this project is taken from Kaggle. The data consist of Categorical Variables like Date, Symbol, Series, Prev Close, Open, High, Low, Last, Close etc.
   * Now we are only using ’Open’ as categorical variable and ’Close’ as target value. It handles missing values and encodes the data of target variable like ’Open’. After it calculates centroid, spreads and distance between vectors.
2. Data Visualization and Analysis(A2) :
   * Using histogram it visualizes and analyse the distribution of target variable(’Open’) price and calculates the mean and variance of the given dataset.
3. Distance Calculation(A3) :
   * Calculation of Distance is calculated by using Minkowski distance between two selected feature variable for different values of r.
4. Model Training and Evaluation(A4 - A8) :
   * The model is divided into 2 tests. One is ’Training Tests’ and other one is ’Testing Tests’. It trains by K- nearest neighbours (KNN) classifier and computes the accuracy.
   * This model also explores shows the effect or impact on changing the number of neighbours (K).
5. Performance Evaluation(A9) :

The code implements the Banker’s algorithm for resource allocation.

**Hyperparameter Tuning:**

For our project, we start by selecting the dataset pertinent to our task, ensuring it contains the necessary features and classes. Identifying two crucial features (Feature\_X and Feature\_Y) and their corresponding classes, we perform any needed preprocessing, like handling missing values or encoding categorical variables. The data is then split into training and testing sets (e.g., 80% for training, 20% for testing). We integrate machine learning library, scikit-learn, to train the kNN classifier on the training set, utilizing Feature\_X and Feature\_Y as inputs. Going beyond the basic steps, we delve into feature selection, handle imbalanced classes if necessary, and explore hyperparameter tuning using techniques like RandomizedSearchCV() or GridSearchCV(). Optimization of the training process and visualizing decision boundaries follow, aiding in understanding model behavior. Model interpretation and analysis, along with iterative refinement based on initial results, lead to an optimized kNN classifier. Documentation encompasses the chosen features, hyperparameters, challenges faced, and a comprehensive report presenting the methodology, results, and insights gained.Top of Form

**E. Model Evaluation**

The performance of tuned models was rigorously evaluatedusing a variety of evaluation metrics including accuracy,precision, recall, and F1 score. Cross-validation techniqueswere utilized to ensure robustness and mitigate overfitting. Statistical tests and visualizations were employed to analyzeand compare model performance, providing valuable insights into their effectiveness in addressing the research objectives.

**F. Sensitivity Analysis**

A sensitivity analysis was conducted to assess the robustness of our findings and evaluate the impact of key parameters on model performance. Variations in data preprocessing techniques and model hyperparameters were systematically explored, enabling a comprehensive understanding of the stability and reliability of our results.

III. Data Preparation

For logical gate implementation tasks (A1 to A4), synthetic data is generated to represent input-output mappings for AND and XOR gates, ensuring a diverse and comprehensive dataset for experimental purposes. The synthetic data creation involves systematically varying input parameters to cover a range of scenarios, enhancing the robustness of the perceptron models.

Additionally, for task A5, where transaction data is utilized, a detailed preprocessing pipeline is implemented to ensure the quality and suitability of the dataset for perceptron model training:

Handling Missing Values:

Identified and addressed any missing values in the transaction data by employing appropriate imputation techniques, ensuring a complete dataset for model training.

Normalization of Features:

Implemented feature normalization to scale numerical attributes, preventing biases due to differing magnitudes of features. Techniques such as Min-Max scaling or Z-score normalization may be applied based on the characteristics of the data.

Encoding Categorical Variables:

Addressed categorical variables within the transaction data by employing encoding methods like one-hot encoding or label encoding. This facilitates the incorporation of categorical information into the perceptron models, enhancing their ability to handle diverse types of input features.

These data preparation steps collectively contribute to the creation of a well-structured and standardized dataset for the subsequent tasks, promoting the effectiveness and generalizability of perceptron models in both logical gate implementation and transaction classification.

IV. Model Implementation and Training

A5. Classifying Transactions as High or Low Value

Built a perceptron model using the provided customer data. Used the Sigmoid activation function and initialized weights and learning rate with a predetermined choice.

Trained the model with the provided data, focusing on classifying transactions as high or low value.

Computed evaluation metrics such as accuracy, precision, recall, and F1-score to assess model performance.

A6. Comparing Results with Matrix Pseudo-Inverse

Compared results obtained from perceptron learning in task A5 with those obtained using matrix pseudo-inverse.

Included metrics such as convergence time and accuracy in the comparison.

A7. Developing a Neural Network for AND Gate Logic using Backpropagation

Developed a neural network to implement AND gate logic using the backpropagation algorithm.

Utilized a Sigmoid activation function and set the learning rate (α) to 0.05.

Monitored convergence during training, considering the learning process converged if the error was less than or equal to 0.002.

Limited the learning process to 1000 iterations if the convergence error condition was not met.

A8. Repeating A1 Experiment for XOR Gate Logic

Repeated the experiment for XOR gate logic as outlined in task A1.

Maintained the learning rate and activation function consistent with the A1 experiment.

A9. Repeating A1 and A2 with Two Output Nodes

Modified perceptron models to have two output nodes and repeated experiments A1 and A2 to investigate performance.

Mapped zero output of the logic gate to [O1 O2] = [1 0], and one output to [0 1].

A10. Learning Using MLP Network from Sci-Kit Manual

Utilized the MLPClassifier function from scikit-learn to train perceptron models for AND and XOR gate logic.

Repeated exercises A1 and A8 using the MLPClassifier() function.

A11. Applying MLPClassifier() on Project Dataset

Used the MLPClassifier() function on the project dataset to explore its performance.

Assessed the results and compared them with previous experiments to evaluate the effectiveness of scikit-learn's implementation.

G. Decision Tree

In the development of a Decision Tree for autograding descriptive answers, the feature labeled "1.0" was identified as the root node due to its highest Information Gain of approximately 1.375, indicating it as the most informative feature for the initial decision-making process. This approach leverages entropy reduction to systematically partition the dataset, aiming to maximize homogeneity within subsets relative to the target variable. The recursive construction of the tree results in a hierarchical structure that facilitates interpretability, allowing for an intuitive understanding of grading determinations through a series of if-else decision rules based on feature values. Such a model not only highlights the critical features influencing grading outcomes but also underscores the importance of feature selection in enhancing model accuracy and interpretability, demonstrating the Decision Tree's utility in providing transparent and explainable predictions for the autograding task.

1. Output :

Finally after training the model, it calculates the different types of metrics like confusion matrix, precision, recall and F1 score to know about the models performance.  
Adapt the kNN classifier to our project by selecting relevant features (Feature\_X and Feature\_Y), preprocessing data, splitting into training/testing sets, utilizing scikit-learn, exploring feature selection, handling imbalanced classes, and optimizing through hyperparameter tuning and decision boundary visualization.

V. Conclusion

Through systematic experimentation and analysis, this study aims to gain insights into the performance and characteristics of perceptron models in various machine learning tasks. The findings will contribute to the understanding of perceptron models and their applications in practical scenarios, utilizing the provided customer data for transaction classification and logical gate implementation.

# **REFERENCES**.

1. H. Karamollaoğlu, İ. A. Doğru, M. Dörterler, A. Utku and O. Yıldız, "Sentiment Analysis on Turkish Social Media Shares through Lexicon Based Approach," 2018 3rd International Conference on Computer Science and Engineering (UBMK), Sarajevo, Bosnia and Herzegovina, 2018, pp. 45-49, doi: 10.1109/UBMK.2018.8566481.
2. T. Bikku, J. Jarugula, L. Kongala, N. D. Tummala and N. Vardhani Donthiboina, "Exploring the Effectiveness of BERT for Sentiment Analysis on Large-Scale Social Media Data," 2023 3rd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2023, pp. 1-4, doi: 10.1109/CONIT59222.2023.10205600.
3. M. V. S, P. Mastan Rao and S. Babu, "Evaluating Social Responsible Attitudes and Opinions using Sentiment Analysis – An Indian Sentiment," 2022 3rd International Conference on Computing, Analytics and Networks (ICAN), Rajpura, Punjab, India, 2022, pp. 1-7, doi: 10.1109/ICAN56228.2022.10007315.
4. C. Li and Z. Hu, "Multimodal Sentiment Analysis of Social Media Based on Top-Layer Fusion," 2022 IEEE 8th International Conference on Computer and Communications (ICCC), Chengdu, China, 2022, pp. 1-6, doi: 10.1109/ICCC56324.2022.10065604.
5. A. Singh, J. Singh and A. Ghosal, "Impact Of Social Media On Stock Market- A Case Of Sentiment Analysis Using Vader," 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), Faridabad, India, 2022, pp. 300-304, doi: 10.1109/COM-IT-CON54601.2022.9850668.
6. M. J. Fuadvy and R. Ibrahim, "Multilingual Sentiment Analysis on Social Media Disaster Data," 2019 International Conference on Electrical, Electronics and Information Engineering (ICEEIE), Denpasar, Indonesia, 2019, pp. 269-272, doi: 10.1109/ICEEIE47180.2019.8981479.
7. R. Xia, J. Jiang and H. He, "Distantly Supervised Lifelong Learning for Large-Scale Social Media Sentiment Analysis," in IEEE Transactions on Affective Computing, vol. 8, no. 4, pp. 480-491, 1 Oct.-Dec. 2017, doi: 10.1109/TAFFC.2017.2771234.
8. Farogh, Sakhawia. (2023). A Comprehensive Evaluation and Comparative Analysis of Data Mining Techniques for Sentiment Analysis in Social Media. International Journal of Advanced Research in Science, Communication and Technology. 40-45. 10.48175/IJARSCT-11609.
9. (2024). Sentiment Analysis using Support Vector Machine and Random Forest. Journal of Informatics and Web Engineering. 3. 67-75. 10.33093/jiwe.2024.3.1.5.

J. Wilson, S. E. Hosseini and S. Pervez, "Identification of Fake News in Social Media Using Sentimental Analysis," 2023 IEEE Industrial Electronics and Applications Conference (IEACon), Penang, Malaysia, 2023, pp. 220-224, doi: 10.1109/IEACon57683.2023.10370300.