Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 08, August-2024

NLP-Based Sentiment Analysis of Financial News

Custom Non-parametric classification method

R. Srusti
Department of ECE
PES University
Bengaluru, India

***Abstract*—This project presents a sentiment analysis system using Natural Language Processing (NLP) techniques, focusing on the Bag of Words model to classify textual data into sentiment categories. An optimal value of k was determined for the K-Nearest Neighbors (KNN) classifier to enhance the accuracy of sentiment prediction. The analysis produced a confusion matrix that illustrates the classifier's performance across different sentiment classes, revealing the relationships between these classes. An accuracy of 82.67 percent was achieved using a custom k-NN classifier. Results provide insights into the effectiveness of the chosen approach and the correlation between sentiment and vocabulary usage for financial news.**

***Keywords*—Sentimental Analysis; NLP; KNN classifier; Non-parametric; Financial news.**

I. INTRODUCTION

In today's data-driven world, understanding the sentiment behind textual information is crucial, especially in fields like finance where market sentiment can significantly impact decision-making.The system employs the Bag of Words model to represent textual data and utilizes the K-Nearest Neighbors (KNN) algorithm for classifying the data into different sentiment categories namely positive, negative and neutral. Sentiment analysis using Natural Language Processing (NLP) and K-Nearest Neighbors (KNN) has become a powerful tool for extracting subjective information from text data. NLP techniques enable computers to process and understand human language, while KNN provides a simple yet effective method for classification based on similarity to labeled examples [8, 11, 12].

Recent advancements, such as transformer-based models like BERT, have significantly improved the accuracy of sentiment analysis [3, 6, 10]. These techniques find applications in various fields, including finance, customer service, and social media analysis [1, 15].While supervised learning approaches are common, researchers have also explored unsupervised and hybrid methods [2, 7]. Ensemble learning techniques, combining multiple models, have shown promise in enhancing performance and robustness [13, 14]. As NLP and machine learning continue to evolve, sentiment analysis systems are expected to become increasingly accurate and efficient, opening up new possibilities for understanding and leveraging textual data across diverse domains.

II. METHODOLOGY

The dataset containing financial headlines and their associated sentiment labels is loaded, and appropriate column names are assigned for clear identification of variables. Following this, a preprocessing phase is conducted, which involves cleaning the data, handling missing values, and performing necessary transformations. As part of this preprocessing, the sentiment labels are converted into numeric values to facilitate machine learning model training.

The dataset is then split into training and testing sets, using an 80/20 ratio, This partition allows for model training on one subset and evaluation on an independent subset, thereby assessing its performance on unseen data. The next step involves feature engineering, where the financial headlines are converted into Bag of Words vectors, capturing the frequency of each word in the dataset. The K-Nearest Neighbors (KNN) classifier is then trained using the Bag of Words vectors from the training set, with the optimal value of k being determined to enhance model accuracy.

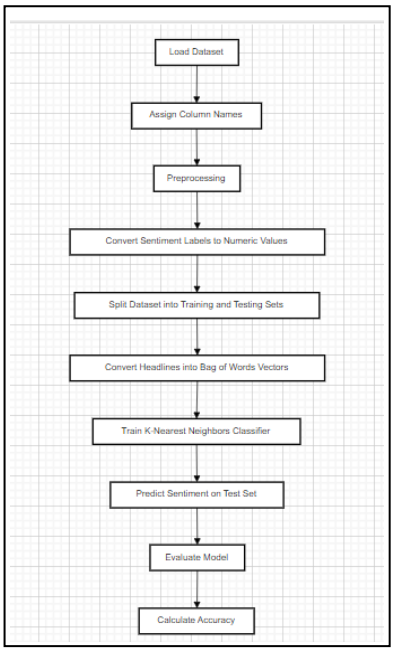

Fig. 1. Flow digram of the model

Once the model is trained, it is used to predict sentiments on the test set, assigning sentiment labels based on the learned relationships from the training phase. The evaluation process includes calculating the overall accuracy of the model and generating a confusion matrix to provide a detailed breakdown of the model's performance across different sentiment classes.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 08, August-2024

III. RESULTS

The performance of the sentiment analysis model was evaluated using a test set comprising 20% of the original dataset. The K-Nearest Neighbors (KNN) classifier, in conjunction with the Bag of Words model, was employed to predict the sentiment labels for the test data. The model achieved an overall accuracy of 82.76%, indicating a strong capability to correctly classify the sentiment of headlines. Additionally, an investigation is conducted into the relationship between the word count of headlines and their associated sentiment. This comprehensive approach allows for a thorough understanding of the model's performance and the nature of sentiment in financial headlines.

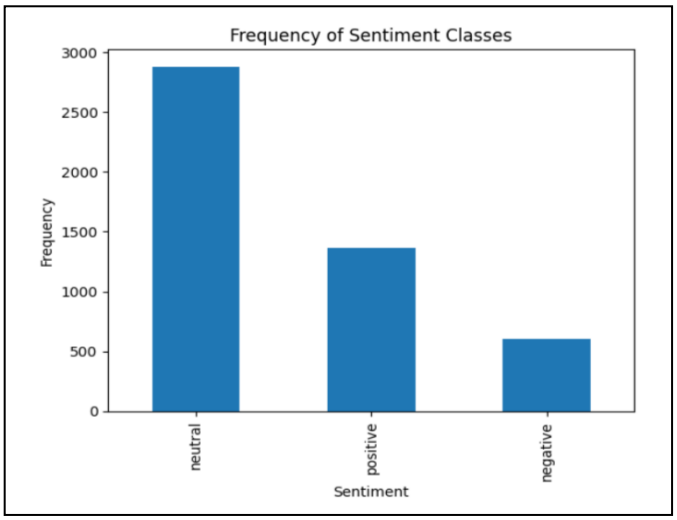

Fig. 2. The frequency of sentiment classes

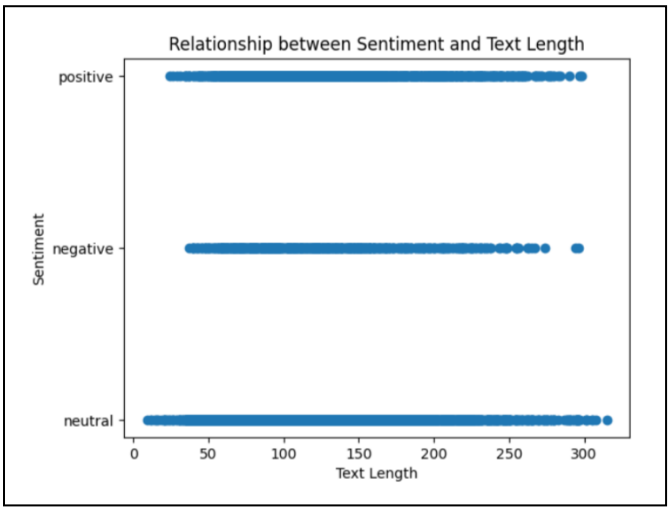

Fig. 3. Relationship between sentiment and text length

IV. CONCLUSION

In conclusion, the sentiment analysis project successfully employed the K-Nearest Neighbors (KNN) classifier alongside the Bag of Words model to classify the sentiment of textual data. we can conclude that the K-Nearest Neighbors classifier, combined with a Bag of Words approach, demonstrates promising performance in sentiment analysis of financial headlines. Overall, this study contributes to our understanding of sentiment expression in financial news and provides a foundation for future research aimed at enhancing the accuracy and reliability of sentiment analysis in the financial sector.

REFERENCES

[1] J. Kim, J. Seo, M. Lee and J. Seok, "Stock Price Prediction Through the Sentimental Analysis of News Articles," 2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN), Zagreb, Croatia, 2019, pp. 700-702, doi: 10.1109/ICUFN.2019.8806182.

[2] M. Fernandez-Gavilanes, T. Alvarez-Lopez, J. Juncal-Martınez, E. Costa-Montenegro and F. J. Gonzalez-Castano, "Unsupervised method for sentiment analysis in online texts", Expert Systems with Applications, vol. 58, pp. 57-75, 2016.

[3] Z. Yang et al., "FinBERT: A Pretrained Language Model for Financial Communications," in Proc. 28th Int. Conf. Comput. Linguistics, 2020, pp. 4513-4523.R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[4] N. Altrabsheh, M. Cocea and S. Fallahkhair, Adaptive and Intelligent Systems, Springer, pp. 40-49, 2014.

[5] M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2004, pp. 168-177

[6] T. Brown et al., "Language Models are Few-Shot Learners," in Adv. Neural Inf. Process. Syst., 2020, vol. 33, pp. 1877-1901.

[7] O. Appel, F. Chiclana, J. Carter and H. Fujita, "A hybrid approach to the sentiment analysis problem at the sentence level", Knowledge-Based Systems, vol. 108, pp. 110-124, 2016.

[8] A. Sharma and S. Dey, "A comparative study of feature selection and machine learning techniques for sentiment analysis," in Proc. 2012 ACM Res. Appl. Comput. Symp., 2012, pp. 1-7.

[9] J. Doshi, "Chatbot User Interface for Customer Relationship Management using NLP models," 2021 International Conference on Artificial Intelligence and Machine Vision (AIMV), Gandhinagar, India, 2021, pp. 1-4, doi: 10.1109/AIMV53313.2021.9670914.

[10] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," in Proc. 2020 Conf. Empir. Methods Nat. Lang. Process.: Syst. Demonstrations, 2020, pp. 38-45.

[11] G. Pradeepa and R. Devi, "Malicious Domain Detection using NLP Methods — A Review," 2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART), Moradabad, India, 2022, pp. 1584-1588, doi: 10.1109/SMART55829.2022.10046882.

[12] A. Soni, B. Amrhein, M. Baucum, E. J. Paek and A. Khojandi, "Using Verb Fluency, Natural Language Processing, and Machine Learning to Detect Alzheimer’s Disease," 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Mexico, 2021, pp. 2282-2285, doi: 10.1109/EMBC46164.2021.9630371.

[13] A. Athar, S. Ali, M. M. Sheeraz, S. Bhattachariee and H. -C. Kim, "Sentimental Analysis of Movie Reviews using Soft Voting Ensemble-based Machine Learning," 2021 Eighth International Conference on Social Network Analysis, Management and Security (SNAMS), Gandia, Spain, 2021, pp. 01-05, doi: 10.1109/SNAMS53716.2021.9732159.

[14] G. Wang, J. Sun, J. Ma, K. Xu and J. Gu, "Sentiment classification: The contribution of ensemble learning", Decis. Support Syst., vol. 57, pp. 77-93, Jan. 2014.

[15] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov, "SemEval-2016 Task 4: Sentiment analysis in Twitter," in Proc. 10th Int. Workshop Semantic Eval., 2016, pp. 1-18.