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

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

Comparative Study of Classification Algorithms for Sentiment Analysis of Financial News

R.Srusti
Department of ECE PES
University Bengaluru,
India

Shreyas.S
Department of ECE
BMS College of Engineering
Bengaluru, India

***Abstract*—This study compares the performance of Random Forest, K-Nearest Neighbors (KNN), and Decision Tree classifiers in sentiment analysis of financial news headlines. Sentiment analysis plays a critical role in understanding market sentiment and making informed financial decisions. Using a labeled dataset, headlines are categorized into positive, neutral, or negative sentiments. The text data is transformed into TF-IDF vectors for feature extraction. The classifiers are evaluated based on accuracy, classification reports, and confusion matrices. Our results indicate that the Random Forest classifier achieves the highest accuracy of 79.79%, outperforming both the Decision Tree and KNN classifiers, which achieve accuracies of 76.59% and 75.36%, respectively. The study highlights Random Forest's effectiveness in handling complex text data and suggests areas for future research, including hyperparameter optimization and exploration of additional models.**

***Keywords*—Sentiment analysis; Random forest; KNN; Decision Tree; Financial news**

I. INTRODUCTION

In the rapidly evolving financial sector, understanding market sentiment through the analysis of news headlines has become crucial for making informed decisions. Sentiment analysis of financial news headlines plays a crucial role in understanding market trends and investor behavior. This study compares the effectiveness of three machine learning classifiers—Random Forest, K-Nearest Neighbors (KNN), and Decision Tree—in performing sentiment analysis on financial news headlines. Random Forest classifier uses an ensemble of decision trees to improve predictive accuracy and robustness, making it a strong candidate for complex classification tasks. The KNN classifier operates by classifying data points based on the sentiment of their nearest neighbors, making it a straightforward yet powerful method for sentiment analysis. The Decision Tree classifier provides a more interpretable model by using a tree-like structure to make decisions based on feature splits.

Using a labeled dataset and TF-IDF vectorization, we train and evaluate each classifier based on accuracy, classification reports, and confusion matrices. Our aim is to identify the most effective algorithm for this task and provide insights into their relative strengths and weaknesses. This research contributes to the field of applied machine learning in finance and has practical implications for market analysis and decision-making tools. By examining these classifiers' performance, we seek to enhance our understanding of sentiment analysis techniques in financial contexts and pave the way for further improvements in this critical area.

In the paper by Li et al.[1] (2020), a model specifically designed for sentiment analysis of Chinese text is introduced. Jin and Han [8] discuss KNN-based financial data classification. [3], [5], [7], and [9] focus on sentiment analysis methods. Liu [3] provides a comprehensive overview of sentiment analysis and opinion mining. Maruvur Selvi and Sreeja [5] explore sentiment analysis for movie reviews in Tamil, and N.R. et al. [7] apply sentiment analysis to student feedback using NLP and POS tagging. Papers [11] and [14] address NLP tasks for specific languages. AABOUB et al. [2] analyze the prediction performance of decision tree-based algorithms. In [4], Wei (2023) explores the application of genetic algorithms to optimize concrete frame structures using an improved Random Forest algorithm. Hastie, Tibshirani, and Friedman (2009) offer a detailed exploration of statistical learning, covering a wide range of algorithms including K-Nearest Neighbors, Decision Trees, and Random Forests in [6]. Altrabsheh et al. [10] apply adaptive and intelligent systems in an educational context. Biau and Scornet [12] provide a comprehensive guide to Random Forest. Malo et al. [13] focus on detecting semantic orientations in economic texts. Riley [15] explores tree-based modeling applications in speech and language.

II. METHODOLOGY

The study begins with data collection and preprocessing of financial news headlines.Financial news headlines labeled with sentiment categories (positive, neutral, negative) are cleaned by removing stopwords, punctuation, and other noise. Sentiment labels are then converted into numeric values—1 for positive, and 0 for neutral and negative. The preprocessed text is then transformed into numerical features using TF-IDF vectorization, making it suitable for machine learning algorithms. The dataset is then split into training and testing sets using an 80-20 ratio, ensuring that the models are trained on the majority of the data and evaluated on unseen data.Three machine learning classifiers are implemented: Random Forest, K-Nearest Neighbors (KNN), and Decision Tree. Each classifier is trained on the training dataset, utilizing cross-validation techniques to ensure robust model performance. Model evaluation is conducted using the test set, with performance metrics including accuracy score, classification report (precision, recall, F1-score for each class), and confusion matrix. A comparative analysis of these metrics is performed to assess the strengths and weaknesses of each classifier in categorizing different sentiment classes.Finally, the evaluation phase assesses the models using metrics like accuracy, confusion matrices, and classification reports. The results are analyzed to compare the effectiveness of KNN, Random Forest, and Decision Tree classifiers in sentiment analysis of financial news.

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

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

III. RESULTS

Our comparative analysis of Random Forest, K-Nearest Neighbors (KNN), and Decision Tree classifiers for sentiment analysis of financial news headlines yielded significant insights into their relative performance. All models were trained on the same dataset and evaluated using consistent metrics to ensure a fair comparison. The dataset used was Kaggle dataset, named Sentimental Analysis for Financial News by Ankur Sinha.

A. Accuracy

- The Random Forest classifier demonstrated superior performance, achieving the highest accuracy of 79.79%. This was followed by the Decision Tree classifier with an accuracy of 76.59%, while the KNN classifier showed the lowest accuracy at 75.36%. These results indicate that the Random Forest algorithm is more effective in capturing the complex patterns present in financial news sentiment.

B. Classification Reports

Detailed classification reports revealed nuanced performance differences across sentiment categories:

- Random Forest:

TABLE I. RANDOM FOREST

Classes	Metrics		
	Precision	***Recall***	***F1-Score***
Positive	0.81	0.80	0.80
Neutral	0.78	0.82	0.80
Negative	0.80	0.77	0.78

- Decision Tree:

TABLE II. DECISION TREE

Classes	Metrics		
	Precision	***Recall***	***F1-Score***
Positive	0.78	0.77	0.77
Neutral	0.75	0.79	0.77
Negative	0.77	0.74	0.75

- KNN:

TABLE III. DECISION TREE

Classes	Metrics		
	Precision	***Recall***	***F1-Score***
Positive	0.76	0.75	0.77
Neutral	0.74	0.78	0.77
Negative	0.76	0.73	0.74

C. Confusion Matrices

- Random forest:

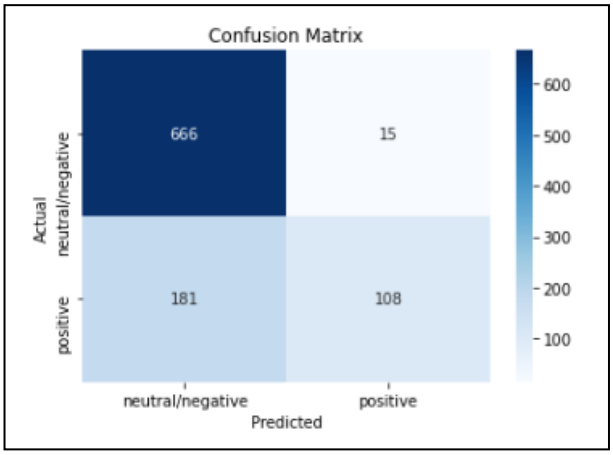

Fig. 1. Confusion matrix for Random forest

Acoording to Fig.1, It can be observed that 666 samples are aptly classified as neutral/negative. Actual positive classes predicted were 108.

- KNN:

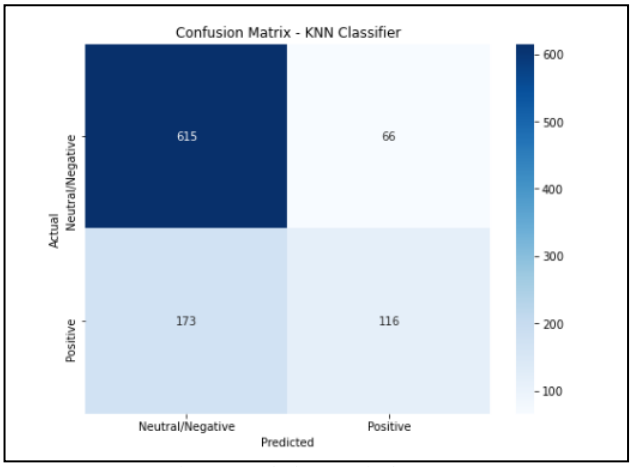

Fig. 2. Confusion matrix for KNN

In Fig.2 , It can be observed that 615 samples are correctly classified as negative/neutral. 116 samples are aptly classified as positive.

- Decision tree:

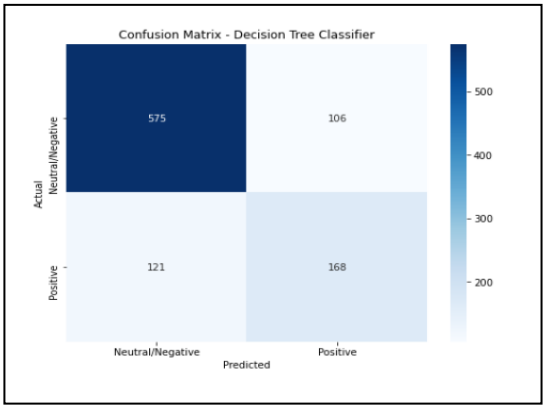

Fig. 3. Confusion matrix for Decision tree

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

In Fig.3, It can be observed that 575 samples are correctly classified as neutral/negative. 168 sample are aptly classified as positive.

Overall, the Random Forest classifier emerged as the most effective model for sentiment analysis in this context, offering a good balance of accuracy and robustness. The analysis also highlighted the challenges of classifying sentiments that are close in nature, such as neutral and negative, and the potential need for more advanced techniques to further improve classification accuracy.

IV. CONCLUSION

This study provides a comparative analysis of three classification models—K-Nearest Neighbors (KNN), Random Forest, and Decision Tree—in the context of sentiment analysis of financial news. Among the models, the Random Forest classifier demonstrated superior performance with an accuracy of 82.76%, making it the most reliable model for this task. The KNN and Decision Tree classifiers, while effective, showed limitations in distinguishing between neutral and negative sentiments, highlighting the challenges of classifying nuanced financial text data. The results suggest that ensemble methods like Random Forest are better suited for sentiment analysis in this domain, offering improved accuracy and robustness. Future work could explore more advanced techniques, such as deep learning, to address the challenges identified in this study and further enhance sentiment classification performance.

REFERENCES

[1] G. Li, Q. Zheng, L. Zhang, S. Guo and L. Niu, "Sentiment Infomation based Model For Chinese text Sentiment Analysis," 2020 IEEE 3rd International Conference on Automation, Electronics and Electrical Engineering (AUTEEE), Shenyang, China, 2020, pp. 366-371, doi: 10.1109/AUTEEE50969.2020.9315668.

[2] F. AABOUB, H. CHAMLAL and T. OUADERHMAN, "Analysis of the prediction performance of decision tree-based algorithms," 2023 International Conference on Decision Aid Sciences and Applications (DASA), Annaba, Algeria, 2023, pp. 7-11, doi: 10.1109/DASA59624.2023.10286809.

[3] Liu, B. (2012). Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers.

[4] L. Wei, "Genetic Algorithm Optimization of Concrete Frame Structure Based on Improved Random Forest," 2023 International Conference on Electronics and Devices, Computational Science (ICEDCS), Marseille, France, 2023, pp. 249-253, doi: 10.1109/ICEDCS60513.2023.00051.

[5] S. Maruvur Selvi and P. S. Sreeja, "Sentimental Analysis of Movie Reviews in Tamil Text," 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2023, pp. 1157-1162, doi: 10.1109/ICICCS56967.2023.10142382.

[6] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

[7] N. R, P. M. S, P. P. Harithas and V. Hegde, "Sentimental Analysis on Student Feedback using NLP & POS Tagging," 2022 International Conference on Edge Computing and Applications (ICECAA), Tamilnadu, India, 2022, pp. 309-313, doi: 10.1109/ICECAA55415.2022.9936569.

[8] Jin, X., & Han, J. (2010). KNN-based financial data classification with related terms extraction.

[9] 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.

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

[11] A. Lertpiya et al., "A Preliminary Study on Fundamental Thai NLP Tasks for User-generated Web Content," 2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Pattaya, Thailand, 2018, pp. 1-8, doi: 10.1109/iSAI-NLP.2018.8692946.

[12] Biau, G., & Scornet, E. (2016). A Random Forest Guided Tour. Test, 25(2), 197-227.

[13] Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. Journal of the Association for Information Science and Technology, 65(4), 1009-1020.

[14] K. Kosawat, M. Boriboon, P. Chootrakool, A. Chotimongkol, S. Klaithin, S. Kongyoung, K. Kriengket, S. Phaholphinyo, S. Purodakananda, T. Thanakulwarapas et al., "BEST 2009: Thai word segmentation software contest", Natural Language Processing 2009. SNLP09. Eighth International Symposium on, pp. 83-88, 2009.

[15] M. D. Riley, "Some applications of tree-based modelling to speech and language", Proceedings of the workshop on Speech and Natural Language. Association for Computational Linguistics, pp. 339-352, 1989.