FINANCIAL NEWS SENTIMENTS: A COMPUTATIONAL LINGUISTICS ANALYSIS

[1]ABAYOMI T. Samuel and [2]OFODU Graceful Onovughe
[1]732A Bolton Road, Swinton, Manchester M27 6EW, United Kingdom
[2]Department of Arts and Language Education,
Faculty of Education, Ekiti State University, Ado-Ekiti.

Abstract
Financial news is crucial for forecasting market trends and making informed investment decisions, but analyzing large volumes of such news is time-consuming and energy sapping. Computational linguistics, or specifically known as Natural Language Processing (NLP) addresses this by automatically detecting sentiments through either heuristic/lexicon-based approaches or machine learning techniques. This study developed a system that uses both machine learning and heuristic-based techniques to analyze the sentiment of financial news articles. The research utilized a dataset consisting of 4,846 records of financial headlines. The dataset was cleaned and preprocessed by removing duplicates, lemmatizing the words, and removing stop words. Then, four machine learning algorithms were experimented with, including logistic regression, linear SVC, random forest, and multilayer perception. The results of the experimentation showed that logistic regression performed best with an accuracy of 76.65% and an F1 score of 71.37% when utilizing the bag of words representation of the text and heuristic feature to learn. The conclusion of the research showed that computational linguistics has potential in analyzing the sentiments of financial news. It was recommended, among other things, that combining lexicon-based approaches and machine learning techniques improves sentiment detection accuracy from financial news compared to using either technique alone. Future researchers should consider this as part of their research.

Keywords: *Sentiment Analysis, Financial News, Machine Learning, Lexicon-Based Approach, Computational Linguistics.*

Introduction

News is everywhere, and many people listen to it every day in order to be informed and make decisions. People's life choices are often driven by how others see the world. For example, individuals may base their actions on the opinions of others (Shuhidan, Hamidi, Kazemian, Shuhidan & Ismail 2018). These opinions may sometimes, be communicative, political, religious, technological, spiritual, educative and financial.

Financial news contains information and reports about economic events, market trends, financial markets, investments, corporate and financial performance. Sentiment analysis of financial news is important because it helps analysts and investors forecast market trends and make informed decisions. Market sentiment indicators help predict future stock performance. If sentiment about a stock is positive, its price will likely increase or stay the same. Conversely, negative sentiment can

signal a potential drop in prices (Agrawal, 2020). Financial news is critical to life, business and global sustainability.

However, analyzing a large volume of financial news to gauge sentiment is time-consuming. This challenge can be addressed through computational linguistics or, more specifically known as, Natural Language Processing (NLP). NLP has been used for various tasks to analyze text, including sentiment analysis, text classification, and topic modeling.

NLP tasks can be solved through two major approaches: the heuristic or lexicon-based approach and the machine-learning approach. The heuristic or lexicon-based approach involves defining a set of rules to achieve a particular task. For example, a rule-based technique in sentiment analysis might involve defining a dictionary containing lists of positive and negative words. The text is then analyzed using this dictionary to determine the overall sentiment based on the aggregate values of the words.

On the other hand, the machine learning approach involves using statistical techniques to automatically detect patterns in a text without explicitly defining rules. This approach uses algorithms to learn from data and make predictions or classification based on learned patterns.

NLP is important in analyzing financial news sentiment because it can automatically detect the sentiment of a text. Several researchers have utilised NLP for sentiment detection, as evident in the work of Sahayak, Shete & Pathan (2015), where the researchers utilised machine learning techniques to analyze the sentiment of Twitter data.

The aim of this research is to develop a system that uses both machine learning and a lexicon-based approach system to analyse the sentiment of financial news.

Statement of Problem

Analyzing the sentiment of financial news is important because it informs stakeholders about what action to take. However, given the large volume of data, manual analysis of the data is inefficient, so there is a need to develop an automated analysis system. Existing approaches, such as lexicon-based and machine-learning methods, each have limitations. The lexicon-based approach can be biased in terms of word in different domain and limited in several vocabularies, and the machine learning model often struggles with nuances of financial terminology (Wankhade, Rao & Kulkarni 2022). Combining both approaches is necessary to enhance sentiment analysis accuracy in the financial domain.

Related Works

A work by Agrawal (2020) developed a lexicon-based system to analyze the sentiment of financial news headlines. The researcher utilized data scraped from Finviz. The researcher utilized the VADER lexicon library from NLTK to determine the sentiment value of financial news headlines. The VADER library was tuned to account for specific keywords found in financial headlines, such as fall and crashes, etc., by assigning custom values to these words so that the analyzer can understand the words in the news headlines in their financial sense and return a specific value. The limitation of this work is that despite the researcher creating custom values for specific keywords, the values assigned to those words are enormous. For example, the word falls was assigned a value of -100. In a sense, in VADER, any word can have a value within the range of -4 to +4. So, giving a single word a -100 is potentially a biased approach. Also, the VADER lexicon only accounts for a few

words in English. So, this means it doesn't capture the sentiment of the text well. Another work by Taj et al. (2019) developed a lexicon-based approach system to analyze news sentiment. The researchers utilized a BBC news dataset that contains 2225 news documents spanning from 2004 to 2005. The researchers utilized the SentiWordNet 3.0.0 dictionary to classify news articles into positive, negative, or neutral sentiments. The drawback of this research is that they didn't report how their proposed model performed in classifying news into sentiment. Another work by Im Tan, San Phang, Chin & Patricia, (2015) developed a rule-based sentiment analyzer system to detect the sentiment of news articles, whether positive or negative. The researchers utilized a dataset with 200 news articles collected from various Malaysian newspapers. The result of their experimentation showed that their sentiment analyzer achieved an f-score of 75.6%. One limitation of the research is that their proposed system cannot address the word ambiguity issue, where words in the subjectivity lexicon are assigned multiple polarities. Another work by Shudinan et al. (2018) developed systems to analyze the sentiment of Malaysian financial news headlines. The systems developed by the researchers include an opinion-lexicon algorithm-based system and a machine learning algorithm system based on Naïve Bayes. The researchers extracted 356 Malaysian online financial news articles from News Trait Times spanning 12 months, from January 2017 to December 2017. They then performed necessary preprocessing, including stop word removal and stemming of the words using Snowball. The limitation of this research is that they did not report the performance of the machine learning model used in terms of any metrics such as accuracy, precision, etc.

The main goal of this research is to expand the body of knowledge of analyzing sentiment in financial news using computational linguistics; hence, the research aims to build a system that uses a lexicon approach feature and a vectorization feature to train a machine learning to analyze the sentiment of financial news headlines. This study aims to provide insights into how various machine learning algorithms perform in this task and find the most suitable algorithm. Furthermore, it explores how utilizing a vectorization approach to create features and the lexicon feature approach improves the model's performance to generalize well.

Proposed System Architecture

This section outlines the proposed methodology for sentiment analysis of financial headlines using machine learning. Four different machine learning algorithms were experimented with to see which model would finally be utilized for the sentiment analysis. The system framework followed a series of steps to achieve the goal of this research, as depicted in Figure 1. The proposed system consists of six stages.

1- Data Collection: A dataset consisting of 4,846 records of financial headlines was collected from Kaggle. The dataset has two columns: the text column, which contains the text of the news headline, and the sentiment column, which indicates whether the text is neutral, negative, or positive.

2- Data Cleaning and Preprocessing: The financial news headline dataset was cleaned by removing duplicates, and the text column was then preprocessed by converting the text to lowercase, removing stop words, lemmatizing the words, and

removing punctuation. Also, the sentiment column was initially in categorical form, so it was then encoded by converting the categories, i.e. neutral, negative, and positive, to 0, 1, and 2, respectively.

3- Feature Extraction: Since machine learning cannot understand text data, vectorization approaches are utilized to convert the text to numerical representation. The Bag of Words (Bow) and TF-IDF Vectorization approaches were compared in conjunction with the custom VADER sentiment (lexicon feature) to see which vectorization method, combined with the custom VADER, would improve model performance. VADER is a lexicon sentiment approach that output polarity value of -4 to +4. This VADER library was tuned to account for certain keywords that exist in financial news headlines and convey sentiment. The figure 2 shows the dictionary of words that were added to the VADER sentiment library.

```
sentiment_dict = {
    "growth": 4, "profit": 4, "gain": 4, "increase"
"bullish": 4, "high": 4, "beats":4,
    "strong": 4, "improvement": 4, "exceed expectati
4, "robust": 4, "recovery": 4,
    "loss": -4, "decline": -4, "drop": -4, "decrease
"crash": -4, "weak": -4,
    "downturn": -4, "shortfall": -4, "miss expectati
"underperform": -4, "volatile": -4,
    "recession": -4
}

vader = SentimentIntensityAnalyzer()
vader.lexicon.update(sentiment_dict)
```

Figure 2: creation of custom VADER lexicon.

4- Dataset Splitting: The dataset was split into training and test sets. 80% of the dataset was allocated for training and 20% for testing. Cross-validation with five folds was performed to evaluate the model's performance on different folds. The random state for splitting the dataset and performing cross-validation was set to 42.

5- Model Selection, Training, and Evaluation: Four machine learning models were selected and trained on the dataset to determine which could accurately classify the financial news headlines into different sentiments. The selected 6models include logistic regression (LR) (Peng, Lee & Ingersoll 2002), linear SVC (L.SVC), multilayer perceptron (MLP), and random forest (RF) (Breiman, 2001). Evaluation metrics such as accuracy and F1-score were used to assess the performance of the selected models. The F1-macro score was the primary metric for choosing the best-performing model in the selection phase (Grandini, Bagli, &Visani, 2020).

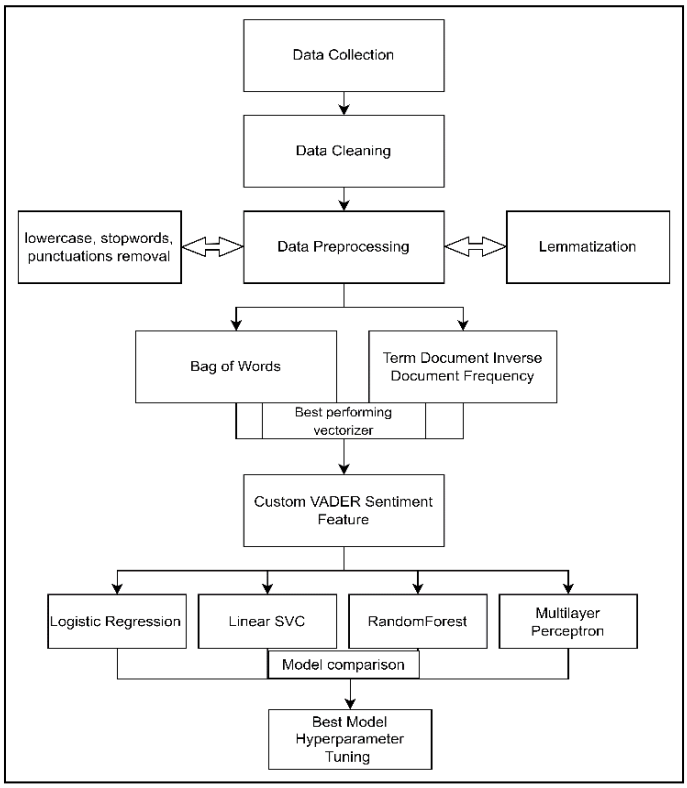

Figure 1: The proposed system framework

Analysis and Results

Exploratory data analysis was performed on the dataset. Firstly, upon analyzing the distribution of words in each news headline, it is seen, as shown in Figure 3, that the majority of news headlines have 20 words on average.

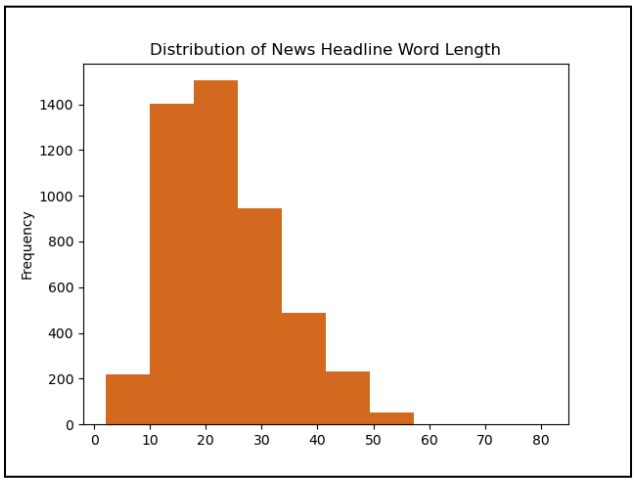

Figure 3: Distribution of news headline word length.

The analysis of the target variable, as shown in Figure 3, shows that close to 60% of the dataset contains neutral financial news, close to 30% consists of positive financial news, and the remaining proportion stands for negative news.

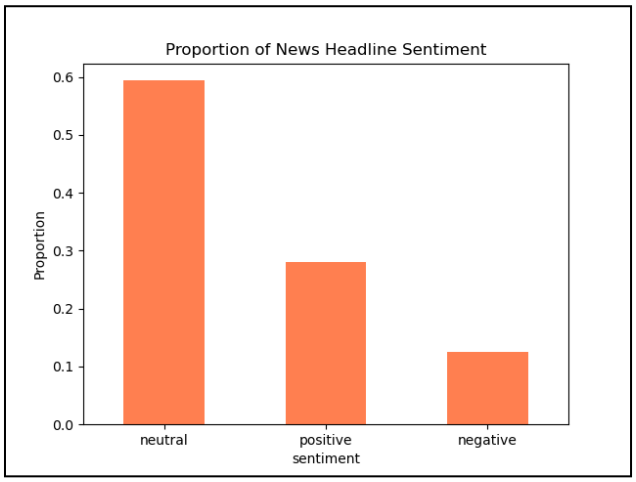

Figure 4: Proportion of news headline sentiment.

Figure 5 shows the word cloud analysis of the financial headlines. As shown, the words that occur mostly are "company," "service," etc.

Figure 5: Word cloud analysis of financial news

The selected models were first trained on each vectorization approach and evaluated to see the models' performance. Then, each vectorization approach was combined with the lexicon VADER sentiment polarity feature to see which combination would yield the best result on the models. Upon experimenting, it was seen that the Bag of Words vectorization and the VADER sentiment polarity feature yielded the best F1 macro average score on the logistic regression model. Tables 1, 2, 3, and 4 show the models' performance when trained on different sets of features.

Table 1: Accuracy and f1-score performance of the models when trained on TF-IDF vectorization.

TF-IDF	LR	L.SVC	MLP	RF
Accuracy	0.7355	0.7531	0.7056	0.7531
F1	0.6424	0.6921	0.6479	0.6682

Table 2: Accuracy and f1-score performance of the models when trained on bag of words vectorization.

BoW	LR	L.SVC	MLP	RF
Accuracy	0.7541	0.7366	0.7107	0.7531
F1	0.6966	0.6811	0.6500	0.6682

Table 3: Accuracy and f1-score performance of the models when trained on TF-IDF vectorization and VADER polarity.

TF-IDF + Vader polarity	LR	L.SVC	MLP	RF
Accuracy	0.7521	0.7634	0.7169	0.7500
F1	0.6662	0.7045	0.6584	0.6530

Table 4: Accuracy and f1-score performance of the models when trained on BoW vectorization and VADER polarity.

BoW + Vaderpolarity	LR	L.SVC	MLP	RF
Accuracy	0.7665	0.7448	0.7221	0.7593
F1	0.7137	0.6947	0.6598	0.6823

As stated, the logistic regression model yielded the best performance when trained on the BoW and the VADER polarity feature. Then, the logistic regression model was tuned for improved performance using GridSearch CV. After tuning, the logistic regression performance increased from a 76.65% to a 77.27% accuracy and 71.37% to 71.89% F1 scores. The optimum parameters that improved the performance of the logistic regression were (C=1.1, penalty='l1', random _state=42, solver='saga').

Table 5 shows the cross-validation performance of the tuned logistic regression when evaluated using accuracy and f1-score

Table 5: cross validation performance of tuned logistic regression.

Cross Val	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Accuracy	0.7582	0.7655	0.7634	0.7570	0.7446
F1	0.7045	0.7136	0.7032	0.7040	0.6898

Figure 6 below shows the performance of the tuned logistic regression compared with other models when utilizing Bag of Words and VADER polarity as the features of the model.

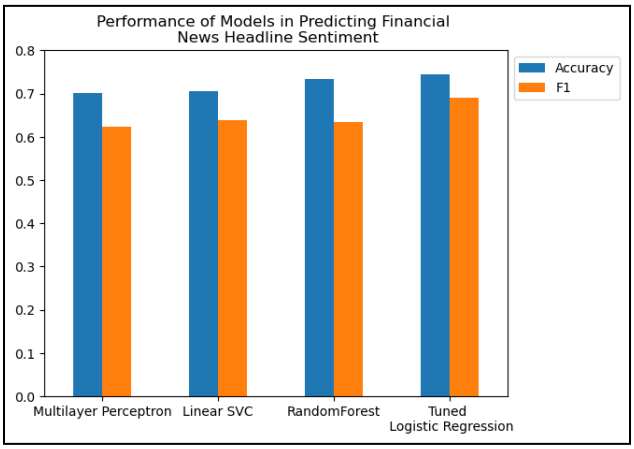

Figure 6 performance of model when utilizing BoW and VADER polarity

Discussion

The result of this study shows the potential of using machine learning to predict the sentiment of financial news. When vectorization approach of the text was used as the features into the machine learning model, the best result that was found was 69.66% f1-score from logistic regression. However, it was seen that combining a lexicon-based approach with machine learning performs better than just utilizing machine learning. Combining machine learning with a lexicon-based approach increases the F1 score from 69.66%, which is just utilizing machine learning with vectorization of the text, to 71.37. The findings of this study also align with the findings of previous research that highlights the potential of using machine learning and lexicon-based approach to determine the sentiment of Face-book comments (Mahmood, Kamaruddin, Naser & Nadzir, 2020).

Comparison of the four machine learning models, including logistic regression, linear SVC, multilayer perception, and random forest when utilizing vectorization and lexicon feature shows that logistic regression achieved the best results, with an accuracy of 76.65% and an F1 score of 71.37%. The model was then further tuned for improved performance, resulting in an improved accuracy score of 77.27% and an F1 score of 71.89%.

Out of all the machine learning models utilised, logistic regression is the model that performed best on the average. Then linear SVC is the second-best performing model, followed by random forest, and lastly multilayer perception.

The performance of the proposed model is also compared to the work of Im et al. (2015), who obtained a f1-score of 75.6%. Even though our result is slightly lower than the result i.e. f1-score of 71.89%, it cannot be judged due to the limitations of the researchers' dataset. In contrast to this work, which included nearly 5000 financial news samples, the researchers only used a dataset of 200 samples. 200 samples of financial news cannot capture the diverse dynamics of words found in financial news. Furthermore, this study extends the work of Agrawal's (2020), who developed a lexicon-based approach to predicting the sentiment of financial news. The researcher took a custom lexicon approach, tuning the VADER to account for specific words. It faced limitations as a result of the bias involved in assigning values to specific words. This study

addresses these limitations by carefully assigning values to specific words and using a hybrid approach that incorporates lexicon-based sentiment scores as features into machine learning to improve performance. Shudinan et al. (2018) also developed a system to analyze the sentiment of financial news using an opinion lexicon-based approach and naive bayes machine learning. However, one limitation of this study is that they did not report the performance of their machine learning. This study not only reports the performance metrics of the various machine learning methods used, but it also emphasizes the advantages of combining lexicon-based features with machine learning for sentiment analysis of financial news.

Conclusion

This research has successfully aimed to develop a system that uses both lexicon and machine learning approaches to predict the sentiment of financial news. The research built a custom VADER library to get the sentiment polarity of a text to use as a feature in conjunction with the vectorization of the text using a bag of words or TF-IDF into the machine learning model. It was seen, however, that the bag of words vectorization in conjunction with the lexicon sentiment feature performed the best in improving the model's performance. After the experimentation, the model that performed better was logistic regression.

Recommendations

Base on the findings of this study, the following recommendations are suggested.

- Using a hybrid approach to detect sentiment: The findings of the research showed that, combining lexicon-based approaches and machine learning techniques improves sentiment detection accuracy from financial news. Future research should adopt a hybrid approach to leverage the strengths of both methods.
- Improve lexicon customization: Creating and fine-tuning custom lexicons, like the modified VADER library in this study, can address the nuances of financial terminology. To accurately capture the evolving language of financial news, it is recommended that new lexicons be created or refined on a regular basis.
- API integration: It is recommended that the proposed system be integrated with an API so that stakeholders can receive real-time sentiment on financial news.
- Exploration of advanced technique: Future work including expanding the dataset to include a wide range of financial news with positive and negative sentiments and utilizing more advanced deep learning models such as Bidirectional Encoder Representations from Transformers (BERT) and recurrent neural networks (RNNs) for better performance is recommended.

References

Agarwal, A. (2020). Sentiment analysis of financial news. In *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)* (pp. 312-315). IEEE.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media* (Vol. 8, No. 1, pp. 216-225).

Grandini, M., Bagli, E., &Visani, G. (2020). Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*.

Im Tan, L., San Phang, W., Chin, K. O., & Patricia, A. (2015, October). Rule-based sentiment analysis for financial news. In *2015 IEEE International Conference on Systems, Man, and Cybernetics* (pp. 1601-1606). IEEE.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR). [Internet]*, *9*(1), 381-386.

Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The journal of educational research*, *96*(1), 3-14.

Mahmood, A. T., Kamaruddin, S. S., Naser, R. K., & Nadzir, M. M. (2020). A combination of lexicon and machine learning approaches for sentiment analysis on Facebook. *Journal of System and Management Sciences*.

Sahayak, V., Shete, V., & Pathan, A. (2015). Sentiment analysis on twitter data. *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, *2*(1), 178-183.

Shuhidan, S. M., Hamidi, S. R., Kazemian, S., Shuhidan, S. M., & Ismail, M. A. (2018). Sentiment analysis for financial news headlines using machine learning algorithm. In *Proceedings of the 7th International Conference on Kansei Engineering and Emotion Research 2018: KEER 2018, 19-22 March 2018, Kuching, Sarawak, Malaysia* (pp. 64-72). Springer Singapore.

Taj, S., Shaikh, B. B., & Meghji, A. F. (2019, January). Sentiment analysis of news articles: a lexicon based approach. In *2019 2nd international conference on computing, mathematics and engineering technologies (iCoMET)* (pp. 1-5). IEEE.

Vajjala, S., Majumder, B., Gupta, A., & Surana, H. (2020). *Practical natural language processing: a comprehensive guide to building real-world NLP systems*. O'Reilly Media.

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(7), 5731-5780.