Literature review: sentiment analysis using machine learning for stock market prices prediction.

Words: 2145

1. Introduction

Sentiment analysis, also known as opinion mining, is a computational approach to the analysis of subjective text in order to capture the sentiment expressed by the writer (Pang & Lee, 2008). The increasing availability of large volumes of data from social media, news articles, and financial reports has been an opportunity for sentiment analysis to become a popular technique for predicting stock market trends (Zhang, Chen, & Leung, 2020) and with the development of advanced techniques machine learning inevitably became a powerful tool for sentiment analysis (Lee & Lin, 2021).

The purpose of this review is to analyse the use of machine learning in the sentiment analysis domain, specifically when applied to stock market prediction and make a point of the current state of research on the topic. The review will provide an overview on datasets, sentiment analysis techniques, machine learning models and performance evaluation metrics and will also identify challenges and opportunities to organize the current discoveries for the finance and computing public to as much easily as possible understand the state of the art at once.

Recent studies have demonstrated the promising potential of machine learning combined with sentiment analysis in the domain of stock market prediction. For example, Gaurav and Saini (2021) used a combination of sentiment analysis and machine learning techniques to predict stock prices, achieving an accuracy of 72.67%. Similarly, Lee and Lin (2021) developed a sentiment analysis model using a combination of natural language processing and machine learning algorithms, achieving an accuracy of 65.29%.

However, also challenges and limitations obstacle the use of these techniques, for example the integrity and quality of datasets, the techniques and models used (Rao & Yarlagadda, 2018). Furthermore, financial data is extremely complex and subject to many external factors, making it difficult to accurately predict stock prices (Wu, Zhang, & Xu, 2021).

1. Sentiment analysis techniques

The review has discerned three main categories in sentiment analysis techniques: rule-based, lexicon-based, and machine learning-based (Rao & Yarlagadda, 2018). Rule based techniques involve the use of predetermined and crafted rules to judge the sentiment of a text. Lexicon-based, on the other hand, assign specific sentiment scores to specific words that are then used to calculate the overall sentiment of a text. Machine learning techniques, lastly, make use of specific algorithms to learn from data and classify text sentiment based on said training data and usually divide it into positive, negative and neutral (Pang & Lee, 2008).

Several studies have utilized machine learning techniques to evaluate the performance and utility of sentiment analysis in the stock market prediction field. For example, Zhang, Chen, and Leung (2020) used and compared the performance of multiple machine learning paradigms when it comes to sentiment analysis, such as Naïve Bayes, Support Vector Machines (SVM), and Random Forest. However, SVM was found to outperform the other models. Similarly, Lee and Lin (2021) compared the performance of several natural language processing and machine learning algorithms, including Word2Vec, Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN), and found that LSTM outperformed the other models.

Machine learning models have proved to be a valuable resource in sentiment analysis, nonetheless they induced doubts in researchers. Apart from the strong dependence on the quality of the training data used, often machine learning models are regarded as “black boxes” in the sense that it is difficult to determine how they make predictions (Rao & Yarlagadda, 2018). This makes it also difficult to interpret their outcomes.

1. Datasets

As appointed in previous paragraphs, machine learning depends heavily on the quality and relevance of the training and testing datasets. Various datasets have been used to evaluate the analysis performance. Some of the most used include financial news headlines or full articles that contain information about stock market trends and sentiment (Zhang et al., 2020). Another important, easily accessible and rich dataset is Twitter’s collection of Tweets. This particular social network is the favourite for business individuals and provides an invaluable resource of data, it in fact contains millions of tweets about financial sentiment (Kumar et al., 2020).

The advantage of tweet is the enormous amount of data and polarizing sentiment expressed in tweets. On the other hand, tweets may contain irony, sarcasm and words that are difficult to interpret. Meanwhile, financial news is a more stable and trustable source, but it doesn’t always reflect market sentiment and is often focused on a specific sector or company. Wu et al also highlight how common are fake news or “clickbait news titles” and social media bots, all these factors contribute to dirty datasets and difficult text interpretation.

1. Machine learning models

Machine learning models have widely been used for sentiment analysis due to their ability to learn from data without the need of extensive human interaction (Yao et al., 2020). Some of the most used machine learning models in the specific field of stock market prediction include logistic regression, support vector machines (SVM), decision trees, and artificial neural networks (ANN) (Kumar et al., 2020).

* Linear regression is a widely known linear model that is used to forecast the probability of an event to occur (Wu et al., 2021).
* SVM instead is a non-linear model that differentiates data into classes based on their features and is used by Zhang et al (2021) to demonstrate the effectiveness of machine learning in sentiment analysis with promising results.
* Decision trees are models that use a tree-like structure to represent decisions and their possible consequences. The use of decision trees has been shown to improve the accuracy of sentiment analysis models in several studies, including Lee and Lin (2021) and Gaurav and Saini (2021). Decision trees have also been used in combination with other algorithms, such as random forest, a model that yielded outstanding results in the studies of Zhang et al (2020) and Lee and Lin (2021) (64,6% accuracy), and gradient boosting (the combination of weaker models to generate a stronger model and minimize losses), to further improve the performance of sentiment analysis models, as demonstrated by Kumar et al. (2020). For example, the study by Lee and Lin (2021) used Gradient Boosting to classify the sentiment of news articles and achieved an accuracy of 62.5%. Similarly, the research conducted by Zhang et al. (2020) used Gradient Boosting along with other machine learning algorithms for sentiment analysis and achieved an accuracy of 73.25%.
* Lastly ANN are a particular type of models inspired by the structure of human brain. They are capable of learning complex non-linear relationships between input features and output targets, making them suitable for sentiment analysis tasks. There are different types of ANNs, such as feedforward neural networks (FFNNs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs).
  + FFNNs have been used to classify news articles into positive, negative, or neutral sentiments based on their impact on stock prices (Kumar et al., 2020).
  + CNNs have been used to extract sentiment features from news articles for predicting stock price movements (Wang et al., 2021).
  + RNNs, such as long short-term memory (LSTM) networks, have been used to capture temporal dependencies in news articles for predicting stock prices (Yao et al., 2020).

However, it is important to note that the performance of machine learning models depends on several factors, such as the quality and quantity of the training data, the choice of features, and the hyperparameters of the model (Wu et al., 2021). Therefore, careful selection and optimization of these factors is necessary to ensure accurate predictions and is difficult to compare the results of each study and say which model is the best performing. Nonetheless results from the studies will be presented in the conclusive paragraph of this literature review.

1. Performance evaluation

Assessing the performance and establish appropriate metrics is fundamental to evaluate and compare different models. Below are some of the key factors and metrics used in the research examined:

* Accuracy: The accuracy of a model is the proportion of correct prediction out of the total prediction made. It is an extremely common metric often used as a starting point due to its generality (Kumar et al, 2020). While accuracy is a useful metric, it has some limitations. For example, in imbalanced datasets where one class is much more frequent than the other. A model that always predicts the majority class will have a high accuracy, even though it is not useful in practice.
* Precision and Recall: Precision measures the proportion of true positive predictions out of all positive predictions, while recall measures the proportion of true positive predictions out of all actual positive instances (Zhang et al., 2020). Both metrics are especially useful when dealing with imbalanced datasets, counterbalancing the accuracy’s fault.
* F1 score: F1 score is a commonly used evaluation metric in sentiment analysis tasks, which measures the balance between precision and recall. It is the harmonic mean of precision and recall, ranging from 0 to 1, with higher values indicating better performance. Wu, Zhang, and Xu (2021) used F1 score to evaluate the performance of different sentiment analysis techniques in the domain of market stock prediction and found that it provided a more balanced assessment of model performance than accuracy alone.

Other factors to consider include AUC-ROC (Area Under the Receiver Operating Characteristic Curve) (Kumar et al, 2020) and the confusion matrix. AUC-ROC measures the area under the ROC curve, which is a plot of the true positive rate (TPR) against the false positive rate (FPR) at different classification thresholds. The scores can range between 0 and 1 with 1 being a perfect score. AUC-ROC is particularly fitted in distinguishing positive and negative samples. Confusion matrix is used by many of the studies mentioned and it is a tool to calculate some of the precedently mentioned metrics such as accuracy, precision, recall and F1 (Yao et al, 2020).

1. Conclusion

Machine learning has been proven as a promising technique combined with sentiment analysis for stock market prediction. This literature has taken into consideration various studies in this domain with a focus on models, techniques and performance metrics.

Most of the studies have proven that these techniques can improve stock market prediction performances and give useful insights.

However, many disagreements are present between the studies examined, especially regarding the choice of the machine learning model used and performance evaluation metrics. With respect to the latter some studies evaluated the performance of machine learning models using accuracy and area under the receiver operating characteristic curve (AUC-ROC) (Kumar et al., 2020), metric that Wu et al (2021) considered as inappropriate due to the imbalanced nature of financial data, others used precision, recall, and F1-score (Zhang et al., 2020) while others like Gaurav and Saini (2021) used only accuracy.

Overall, the choice of performance evaluation metrics may depend on the specific research question and context, as well as the availability and quality of data.

When it comes to one of the main focuses of this literature review, which is machine learning, various models, features and techniques have been used by various studies with different outcomes about their effectiveness, here is proposed a brief summary:

* In Zhang et al. (2020), the best performing model for predicting stock price movements based on financial news sentiment was the Random Forest model with a precision score of 0.6 and F1-score of 0.52.
* Lee and Lin (2021) employed a hybrid model consisting of sentiment analysis, feature selection, and machine learning algorithms, achieving an accuracy of 56.92% for predicting stock price movements using news headlines.
* Kumar et al. (2020) surveyed several studies and reported that Artificial Neural Networks (ANNs) have shown promising results in predicting stock prices, with some studies achieving accuracy rates above 90%.
* Yao et al. (2020) conducted a survey on the use of deep learning in sentiment analysis for stock market prediction and reported that Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) have shown promising results in capturing the temporal and spatial dependencies of the sentiment signals.

It is worth noting that the choice of the best performing model may depend on the specific dataset, features, and performance metrics used in each study. Therefore, a thorough evaluation and comparison of different models may be necessary but beyond this review’s scope.

Finally, this review highlights the importance of sentiment analysis and machine learning for the stock market implementation. Further studies on the models, evaluation metrics and quality of datasets could greatly improve performances and application in this domain, especially when factoring the new generation machine learning models that made great advances in 2022. For example, Imran et al (2022) utilized state-of-the-art synthetic text generation to refine imbalanced datasets regarding educational feedback, which is a common problem in financial datasets as expressed by Wu et al (2021), finding improved performances for the models trained with refined datasets. This scenario is particularly important for deep learning techniques that are particularly data-hungry (Imran et al, 2022).

References

* Gaurav, S., & Saini, R. (2021). Sentiment analysis and prediction of stock market trends using machine learning. Intelligent Systems Reference Library, 200, 119-128.
* Kumar, R., Singh, P., & Kumar, A. (2020). Sentiment analysis using machine learning in stock market prediction: A survey. In Advances in Machine Learning and Data Science (pp. 219-229). Springer, Singapore.
* Lee, T. K., & Lin, Y. C. (2021). Predicting stock price direction with sentiment analysis and machine learning. Technological Forecasting and Social Change, 165, 120574.
* Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1-2), 1-135.
* Rao, D., & Yarlagadda, R. (2018). A comprehensive review of sentiment analysis techniques in text mining. Journal of King Saud University-Computer and Information Sciences, 30(4), 431-441.
* Wang, Y., Gao, S., Liu, J., & Wang, Y. (2021). A hybrid deep learning model for sentiment analysis in stock market prediction. Applied Intelligence, 51(1), 212-222.
* Wu, H., Zhang, B., & Xu, C. (2021). Sentiment analysis for financial market prediction: A survey. IEEE Transactions on Computational Social Systems, 8(2), 430-445.
* Yao, S., Ma, S., Chen, L., & Liu, X. (2020). Deep learning for sentiment analysis in stock market prediction: A survey. IEEE Transactions on Neural Networks and Learning Systems, 31(10), 3782-3808.
* Zhang, Y., Chen, C., & Leung, S. C. (2020). Sentiment analysis of financial news using machine learning algorithms. Journal of Business Research.
* Imran et al. (2022). The impact of synthetic text generation for sentiment analysis using GAN based models. Egyptian Informatics journal, (23), 547-557.