Deep learning-based sentiment analysis for predicting financial movements

Hadhami Mejbri[1], Mariem Mahfoudh[1,2], and Germain Forestier[3]

[1] Kairouan University, ISIGK
Avenue Khemais El Alouini, 3100, Kairouan, Tunisia
[2] MIRACL Laboratory, University of Sfax
Route de Tunis Km 10 B.P. 242 Sfax 3021, Tunisia
[3] IRIMAS, University of Haute-Alsace
12 rue des Frères Lumière F-68093 Mulhouse Cedex France
hadhamimejbriinfo@gmail.com, mariem.mahfoudh@gmail.com
germain.forestier@uha.fr

Abstract. Sentiment analysis is a computational study of opinions, feelings, emotions, ratings and attitudes towards entities such as products, services, organizations, individuals, issues, events, subjects and their attributes. Our research is used to predict stock market movements, aims to improve the accuracy of polarity of comments, in order to accurately predict financial movements. This by creating a dictionary of emojis that contains the emoji as keys and its meanings as values. We used the dictionary at the preprocessing level to keep the meanings of emojis because they carry a lot of emotions that help us to clearly specify the polarity of the comments. We have also created a list of stopwords related to the financial field to properly clean our database. Time series and linking sequences of data is very important to properly predict stock market movements. We have therefore chosen to work with the Long short-term memory (LSTM) model. Next, we came up with two models: the first model to predict stock market movements using investor sentiment analysis of Amazon stock which gives us 93% accuracy. The second model is used to predict financial movements through historical Amazon prices. We extracted the database we used for the sentiment analysis from Twitter as the Twitter comments are up to date. As for the historical prices of Amazon stock we extracted from the most famous trading platform YahooFinance.

Keywords: · Sentiment Analysis · Stock movement prediction · Opinion Mining · Automatic Natural Language Processing (NLP) · Deep Learning · Machine Learning · Time series.

1 Introduction

Sentiment analysis have been used in diverse fields such as health, finance, sports, politics, hospitality, and consumer behavior. We are interested in our work to the financial field, more specifically in online trading platforms.

This is the author's version of an article published at KSEM 2022. The final authenticated version is available online at: https://doi.org/10.1007/978-3-031-10986-7_47

Online trading is simply buying and selling financial securities through on-Line trading platforms or mobile trading apps, to make money between buying and reselling, and vice versa. Intervening on the financial markets represents risks that can lead to financial losses. For example, market risk due to price instability linked to general economic and market fluctuations, liquidity risk and the difficulty of finding a counterparty (to sell a financial instrument at a reasonable price at a given time), etc. [9], so to minimize the risk of loss, we help through our proposed model to create a computer system that helps investors make the right decision in order to have a good trading experience in the financial markets. There are several areas of trading like commodity trading, currency trading and stock trading. Statistic shows that 90% of traders do not make money when trading the stocks [8], so we chose stock trading in the research and particularly we chose to predict the financial movements of Amazon.

Our research is divided into three main parts: the first part consists of analyzing investor sentiment towards Amazon. The reviews are taken from the official twitter page of Amazon. The second part is to predict the financial movement of Amazon stocks using sentiment analysis. Finally, the third part focuses on the use of historical prices extracted from the most famous trading site: YahooFinance.

Our article is composed of three parts. The first part is the related works in which we presented some works related to sentiment analysis in the financial field whereas in the second part, we presented our proposed approach and in the third part we discussed the results of our research.

2 Related work

Kordonis et al. [3] predicted how the market would subsequently behave via sentiment analysis on a set of tweets over the past few days, so they developed a system that collects past tweets, drafts, and examines the usefulness of various machine learning techniques such as Naive Bayes Bernoulli classification and Support Vector Machine (SVM), to provide positive or negative sentiment on the tweet corpus. The results still show that changes in public sentiment can affect the stock market.

Kalyani et al. [2] predicted market trends, and for this they used a polarity detection algorithm to label news articles. For this algorithm, a dictionary-based approach was used. Positive and negative word dictionaries are created using general and finance-specific sentiment words. They created a dictionary for stopword removal that also includes finance-specific stopwords. Based on this data, they implemented three classification models which they tested in different test scenarios. Then, after comparing their results, Random Forest performed very well for all test cases ranging from 88% to 92% accuracy. The accuracy tracked by SVM is also considerable around 86%. The performance of the Naive Bayes algorithm is around 83%.

There are some work on Deep Learning architecture for sentiment analysis. Nti et al. [6] investigated the application of the attention-based deep neural network LSTM in predicting future stock market movements. They also built an aggregate stock dataset, and an individual dataset, including stock history data, financial tweets sentiment, and technical indicators on the US stock market. The experiment investigates the temporal sensitivity of financial tweet sentiment and methods for calculating collective sentiment. The researchers also experimented with conventional LSTM and attention-based LSTM for performance comparison. The results prove that financial tweets published from the close to the open of the market have more predictive power on the movement of stocks the next day.

Liu et al. [7] proposed a model named RCNK (Recurrent Convolutional Neural Kernel) for the prediction of stock price movement. In order to improve the prediction accuracy of the model and the cumulative returns of the trading simulation, the RCNK model has been optimized in three aspects: data collection, textual data processing, and the model classifier.

Researchers have proposed several models that improve the results of predicting financial movements with several techniques.

There is researchers who analyze the temporal system (when are the periods of time which represent well the prices of stocks, and give better predictions), on the other hand there is researchers who have worked well to improve the results of sentiment analysis towards stocks. In our approach we tried to improve on both, we used “Open” and “High” respectively, as the training dataset, The “Open” column, which represents the stock’s opening price each day and the “High” column, which represents the highest stock price reached each day. Furthermore, we created a dictionary of emojis and a list of financial stopwords to improve the stock sentiment analysis result (in our research we analyzed amazon stocks).

3 Proposed approach

To improve the trading experience, it is important to know the opinion of the investors on the share price as well as on the stocks. The objective of our work is to identify the courses of action that will be descending and those that will be ascending. That’s why we offer sentiment analysis of tweets to help investors choose when and which stock to buy or sell in order to increase their earnings. In the other hand we predict the stock market movements using the historical prices extracted from the most famous trading site ”Yahoo finance”. The purpose of these two methods is to improve the prediction through comparing the results of the sentiment analysis and the historical prices methods.

Figure 1 illustrates the steps and techniques of our proposed sentiment analysis approach which includes seven main components:

1. *Data collection.* Collection of tweets from the Amazon page and the date of each. Extract the historical prices of Amazon stocks from the Yahoo finance trading site and the date of each.
2. *Data pre-processing.* it consists to clean the data and extract the relevant information in order to have an accurate classification of the sentiments.

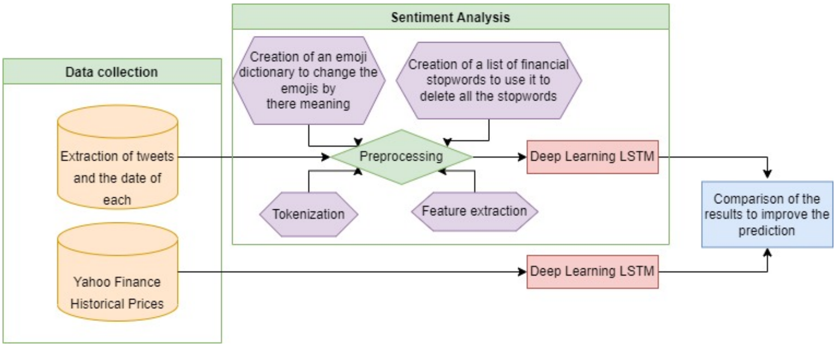
Sentiment Analysis
Data collection
Creation of an emoji dictionary to change the emojis by there meaning
Creation of a list of financial stopwords to use it to delete all the stopwords
Extraction of tweets and the date of each
Preprocessing
Deep Learning LSTM
Comparison of the results to improve the prediction
Tokenization
Feature extraction
Yahoo Finance Historical Prices
Deep Learning LSTM

Fig. 1. Summary of our proposed approach

3. *Feature extraction.* use word integration (count vectorization, tf-idf transformation, BOW) to convert tweets into digital representations.
4. *Sentiment classification.* use the LSTM long short-term memory numeric representation of tweets.
5. *Prediction of financial movements using tweets.* Representation of sentiment classification results in a curve that represents the polarity of tweets over time.
6. *Prediction of financial movements using historical stock prices.* Creation of an LSTM model to predict stock prices over time and compare the results with actual stock prices.
7. *Evaluation.* Once the model is built on the training data, we use it to predict the evaluation of the test data. To assess performance, we calculated accuracy.

3.1 Data collection

We created a *Developer* twitter account to have an API that allows us to extract tweets from the official Amazon page and the date of each one. In our experiment, we used tweets which were posted from 2021-02-12 to 2021-08-17.

We examined the tweets to see if there is any relation between the future stock price and users sentiment. In other words, we want to see if we can predict a future stock price based on the current sentiment of many users in order to improve the prediction using the historical prices of the stock, so we have loaded AMAZON's past stock price data. From the most famous trading site Yahoo-Finance, we have selected the values of the first and second columns ("Open" and "High" respectively) as training dataset. The "Open" column represents the stock's opening price each day and the "High" column represents the highest stock price reached each day.

We used historical stock prices from Amazon from 2018/08/01 to 2021/07/29 to train our lstm model, and prices from August 2021 to test our model. We used Textblob library to labbel our tweets.

3.2 Pre-processing

1. *Date standardization.* Transfer the dates of each comment to the date format necessary for the prediction of financial movements over time ***(2021-02-12T17:22:53Z to 2021-02-12 17:22:53)*** .
2. *Cleaning emojis.* Before starting the known natural language processing techniques (punctuation removal, stopword removal, tokenization, stemming and TfIdf), We noticed that our database contains a lot of emojis with a lot of feelings and which must be removed if we removes punctuation, and for this reason we decided to create a dictionary of emojis. We have chosen to extract the twitter emojis with their meanings from the French website fr.piliapp.com in order to have a dictionary that contains all the emojis with their meanings (1665 emojis) in this form 'emoji':'meaning' (the emoji as keys and the meaning as value) the dictionary we created is in French and since the tweets in our database are in English, we used the translate library to translate it into English. After creating our data dictionary, we created a function clean_text_round1 which uses the re library and the string library to replace emojis with its meaning and remove punctuations in order to apply known NLP techniques to our data.
3. *Tokenization.* Tokenization is a method of dividing a piece of text into smaller units called tokens. Here the tokens can be words, characters or sub-words. Therefore, tokenization can be roughly divided into 3 types: word, character, and subword (n-gram characters) tokenization. In our research we need the sentiments of the words so we applicated word tokenization.
4. *Remove stop words.* Stop words are words that don't add much meaning to a sentence and don't convey any emotion so they don't indicate any valuable information about the sentiment of a sentence. They can be safely ignored without sacrificing the meaning of the sentence. For example, words like the, he, have etc. We used the nltk library to import English stopwords to remove unnecessary words from our database. And like our research done in financial databases, we created a new Word stop list related to the financial domain (new_stopword), and removed all the financial stop words in our list from our database. And to be sure that our database is well cleaned, we used wordcloud to display the words, repeat the most and add the superfluous words to our list of stopwords.
5. *Feature extraction.* Machines cannot understand characters and words. So when it comes to textual data, we need to represent it in numbers for it to be understood by the machine. A simple text could be converted into features with various techniques such as Bag of Words (BOW), tf–idf wish is very often used for text features, there is also another class called TfidfVectorizer that combines all the options of CountVectorizer and TfidfTransformer in a single model, so in our research we used the TfidfVectorizer technique.

3.3 Sentiment Analysis and Prediction

In our research, we used time series and linking data sequences which are very important to predict stock market movements well. So we have chosen to work with the LSTM model which uses a technique allowing it to process information over extended time intervals.

The LSTM technique learns to store information over extended time intervals via recurrent back-propagation is time consuming mainly due to insufficient and decreasing error feedback [1].

4 Results and discussion

In this section, we will explain our results curves. At first, we have presented the result of our expected stock price using historical prices and actual stock prices in the picture 2. Then, we have presented the opinions of Internet users over time as shown in the Figure 3, 4, 5, 6 (sentiments on the vertical axis and time on the horizontal axis polarities) Although the exact prices of the predicted prices are not always close to the actual price, our model always indicates general trends, such as upwards or downwards. This research tells us that LSTMs can be effective in time series forecasting.

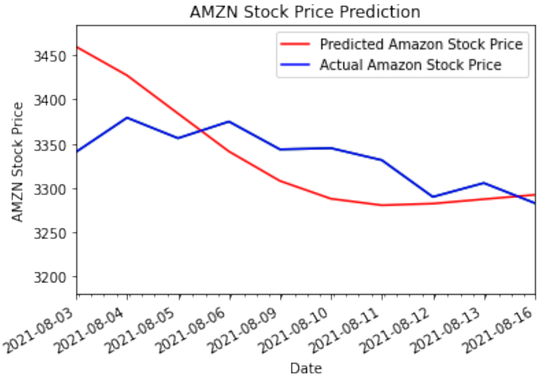

Fig. 2. Prediction of financial movements with the LSTM model using historical prices

This prediction can be improved by merging the two prediction methods used in our research: sentiment analysis of tweets and historical prices.

We evaluated our model to improve it based on what we learn to evaluate it, and whether our model is useful, whether we need to train our model on more data to improve its performance, and whether we should we include more features? The three main parameters used to evaluate a classification model are: accuracy, precision and recall. The best result we got using LSTM model is 93% accuracy. Using sentiment analysis, we found that tweets with date 2021/08/03 are positive tweets, which affects stock prices which therefore increased, which is the case as shown in Figure 2 of actual stock prices.

On the day of 2021/08/04 the comments are increasingly negative throughout the day, which clearly affects the prices of stocks which will therefore go down.

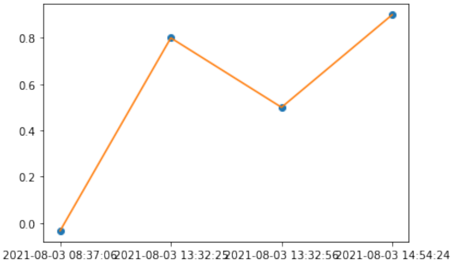

Fig. 3. Prediction of financial movements with the LSTM model using sentiment analysis on August 3, 2021.

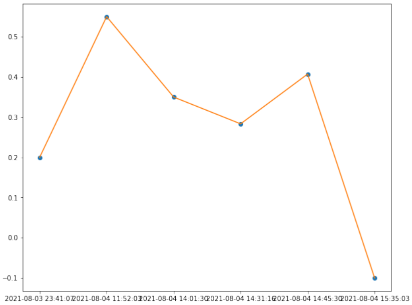

Fig. 4. Prediction of financial movements with the LSTM model using sentiment analysis on August 4, 2021.

But as Figure 5 shows at the end of the day 2021/08/04 and the beginning of the day 2021/08/05, the feedback becomes positive, which will affect the stock prices, and what is coming rise, as indicated by the curve of actual stock prices from the figure 2.

And since August 6, 2021 until 16 august 2021, the comments are increasingly negative, and at the same time, we find that the prices of actual Amazon stocks are decreasing considerably.

Our research is inspired by the works [3], [2] and [6]. Such as the work [3] the researchers created a dictionary of emojis which indicates the polarity of the emojis which carries a feeling. Which makes us think about creating a dictionary of emojis which carries the meaning of the emojis and we have put all the emojis without exception. In order to attract all the feelings of our twitters and find out if there is any sarcasm. We also adopted the idea of [2] to create a list of financial stopwords wish improve sentiment analysis results. And finally, the work [3] that analyzes the aspect of time on the prediction results, that guided us to try different stock prices on the same day and choose to use the Open and High columns as our training data set. The Open column represents the stock's opening price each day and the High column represents the highest stock price reached each day. As the table 1 indicate our model give the best accuracy.

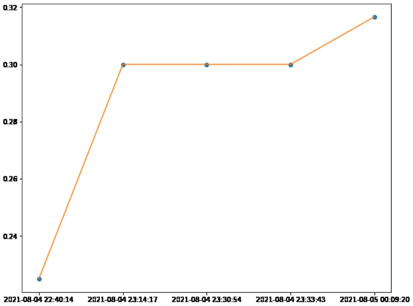

Fig. 5. Prediction of financial movements with the LSTM model using sentiment analysis on August 4 and 5, 2021.

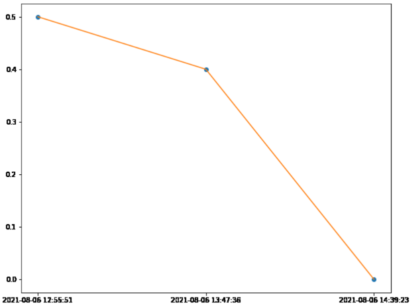

Fig. 6. Prediction of financial movements with the LSTM model using sentiment analysis on August 6, 2021.

5 Conclusion

The stock market is often volatile and changes abruptly due to economic conditions, the political situation and major events for the country. To improve the trading experience: it is important to know the opinion of investors on the share price as well as on stocks, and thanks to social networks now, we can know the opinions of Internet users. In our research, we used Amazon page tweets to do Amazon customer sentiments analysis to identify which courses of action are going to be bottom-up. Our task consists in two main steps : in the first step, we cleaned our database by creating a dictionary of emojis that replaces emojis with their meanings, and we created a list of stopwords related to the financial field, then we applied the TF / IDF, to represent the words numerically. Time series and linking data sequences are very important for predicting stock market movements well. And for that we then chose to use several models of the "LSTM" classifier and a model of the "decision tree" classifier. After a comparison between these models, the results of our proposed approach, the LSTM classifier that demonstrates the best reliability demonstrates a performance of 93%. The second part of our proposed approach is to use our investor sentiment analysis

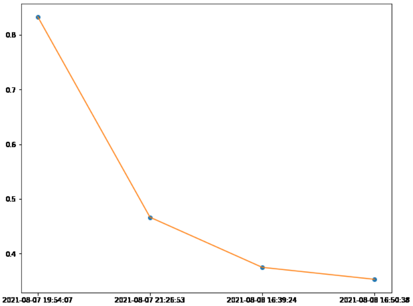

Fig. 7. Prediction of financial movements with the LSTM model using sentiment analysis on August 7 and 8, 2021.

Table 1. Comparison of our work with previous works

Works	Database	The technique used	Algorithm	Accuracy
[3]	Yahoo-Finance, Twitter	They have created a dictionary of emojis that indicates the polarity of emojis that carry sentiment	Naive Bayes	80%
[2]	Yahoo-Finance, Reauters	They have created a dictionary for stopword removal which also includes finance-specific stopwords.	Random Forest	92%
[6]	Yahoo-Finance, Twitter	The experiment investigates the time sensitivity of financial tweet sentiment and methods for calculating collective sentiment.	LSTM	62%
[7]	guba eastmoney	The different from previous studies is that The researchers treated the text data as sequential data and they used the RCNK model to train sentiment embeddings with the temporal features.	Recurrent convolutional neural kernel	66.62%
Our work	Yahoo-Finance, Twitter	We created a dictionary of emojis that contains the emoji as keys and its meanings as values, this dictionary we used at the preprocessing level to keep the meanings of emojis , we have also created a list of stopwords related to the financial field to properly clean our database.	LSTM	93%

to predict market movements of Amazon stocks. As for the other method, we used to predict the stock market movement, the method using historical prices extracted from the most famous trading site Yahoo Finance for this prediction we used LSTM which gives us good results as showing Figure 2. The perspectives of our work consist in changing our database to make it dynamic in real time, and in uniting the results of the sentiment analysis with the results of historical prices, in order to develop and exploit a computer system allowing to invest in the financial markets automatically (without human intervention). We also aim to solve the problem of ironic sentences using our dictionary of emojis because an opposition between the emotion of the emoji and the emotion of the text usually constitutes sarcasm.

References

1. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation Journal, 9, 1735-1780.
2. Kalyani Joshi, Prof. Bharathi H. N, Prof. Jyothi Rao, F.:stock trend prediction using news sentiment analysis, International Journal of Computer Science & Information Technology (IJCSIT), 2016.
3. Kordonis, John and Symeonidis, Symeon and Arampatzis, Avi, F.: Stock Price Forecasting via Sentiment Analysis on Twitter, Proceedings of the 20th Pan-Hellenic Conference on Informatics November 2016.
4. Kaihui Zhang and Lei Li and Peng Li and Wenda Teng, F.:Stock trend forecasting method based on sentiment analysis and system similarity model, T.: Proceedings of 2011 6th International Forum on Strategic Technology, vol. 2, pp. 890-894, 2011,
5. Aparna Bhat and S. Kamath, F.: Automated stock price prediction and trading framework for Nifty intraday trading, J. 2013 Fourth International Conference on Computing, Communications and Networking Technologies, pp. 1-6, 2013
6. Y. Xu and V. Keselj, T. 2019 IEEE International Conference on Big Data (Big Data), F.: Stock Prediction using Deep Learning and Sentiment Analysis, 2019.
7. Liu, Suhui and Zhang, Xiaodong and Wang, Ying and Feng, Guoming, J. plos one, 2nd edn. Public library of science, F.: Recurrent convolutional neural kernel model for stock price movement prediction, June 2020.
8. Dale Gillham, Trading the Stock Market – Why Most Traders Fail, 2021. https://www.wealthwithin.com.au/learning-centre/share-trading-tips/trading-the-stock-market.
9. Hugo, Tom, what are the different types of fianacial risks, 2020, https://study.com/academy/lesson/financial-risk-types-examples-management-methods.html