

Financial sentiment analysis of tweets based on deep learning approach

Aattouchi Issam¹, Ait Kerroum Mounir¹, El Mendili Saida², El Mendili Fatna³

¹Computer science research Laboratory, Faculty of Sciences, Ibn Tofail University, Kenitra, Morocco

²Engineering Sciences Laboratory, National School of Applied Sciences, Ibn Tofail University, Kenitra, Morocco

³Image Laboratory, School of Technology, Moulay Ismail University of Meknes, Meknes, Morocco

Article Info

Article history:

Received Aug 5, 2021

Revised Jan 8, 2022

Accepted Jan 13, 2022

Keywords:

Convolutional neural network

Financial tweets

Latent dirichlet allocation

Neural network

Prediction

Sentiment analysis

ABSTRACT

The volume of unstructured texts has increased dramatically in recent years due to the internet and the digitization of information and literature. This onslaught of data will only grow, and it will come from new and unusual sources. Thus, it will be necessary to develop new and inventive approaches and tools to process and make sense of this data. Investors in the financial markets can now get information faster than ever before thanks to the expansion of communication channels, in addition to the online availability of news and reports in text format through providers like Reuters and Bloomberg. This contains a plethora of information that is often overlooked by financial market data. In order to measure the sentiment of a text, predictive and deductive methods are applied, these methods aim at extrapolating new features from big data. The main objective of this study is to create and test a new system capable of predicting finance and non-finance related tweets. The convolutional neural network (CNN) and latent dirichlet allocation (LDA) algorithms are used in the proposed approach. The suggested model's correctness is tested against a benchmark financial dataset, and the results demonstrate that with a database of 1,000,000 data points, our model is 99% accurate.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

El Mendili Fatna

Image Laboratory, School of Technology, Moulay Ismail University of Meknes

Meknes, Morocco

Email: f.elmendili@gmail.com

1. INTRODUCTION

We can see a significant increase in user-generated content on the web as a result of enhanced digitization, which provides people's thoughts on various themes. The computational study of assessing people's feelings and views for an entity is known as sentiment analysis. In recent decades, this field has become a hot topic. For business intelligence, managing the flood of data produced by networks of people or devices has become increasingly important [1]. The huge range of interactions between participants that may be captured is one of the most noteworthy elements of this big data boom [2]. As a consequence of this trend, most data are becoming more unstructured, by textual data accounting for a considerable amount of the data stream. E-mail communications and tweets are examples of such data, as are company reports and regular news broadcasts. As the volume of data grows at an exponential rate, developing strategies for skimming through thousands of pages of digital texts and retrieving the critical insight hidden in basic view becomes increasingly crucial. The information explosion is particularly difficult for financial sector specialists. Websites, Twitter posts, and other social media platforms are used by a variety of companies to deliver market data and opinions. According to the effective market theory, market efficiency is dependent on timely

and accurate supply of market information to investors, thanks to the perceptive constraints of investors' thoughts and the limiting period they take to make judgments, perfectly informed and logical decisions are frequently impossible when the amount of financial data expands at a breakneck speed.

Machine learning (ML) is a multidisciplinary discipline that combines computer science methods and statistics to develop classification algorithms and predictive models. Sentiment analysis is a machine learning technique that detects polarity like positive or negative ideas in documents, texts, lines, paragraphs or subsections. Because the structure of a language influences how its speakers view their world, using text analysis [1] and similar approaches to conduct sentiment analysis of financial communications is one of the greatest strategic tendencies for coping with the present texting craze. Such methods cannot only reveal current communal attitude trends as shown in the media, but also offer insights for understanding potential implications and lowering the dangers of making trades in turbulent financial markets. Experts in the fields of financial markets and business intelligence need to know how to use text analysis to construct advanced data analytics in order to deal with today's huge data explosion. Nevertheless, the great majority of decision-support system managers have shown minuscule attention in text analysis as a viable alternate to canonical decision models that have been used in the past.

In the literature, financial sentiment analysis is regarded as a significant and difficult subject [3], [4]. In order to identify the polarity of financial text, the most current approaches on sentiment analysis use general dictionary [5], [6], domain-specific dictionary [7]-[10], or statistical/machine learning methods. Harvard global institute [11], financial polarity lexicon (FPL), SentiWordNet, SentiStrength2, SenticNet, manufacturer product quality assessment (MPQA), and LM are some of the most commonly used dictionaries in the field of financial sentiment research. The financial polarity lexicon and the Loughran and McDonald dictionaries have been employed in recent research. Hardgrove grindability index (HGI), SenticNet, and MPQA are more general, which means they are more prone to misclassify popular financial words. For instance, according to a thorough examination of the HGI [11] lexicon by [10], in a financial context, more than 75% of the negative terms used in HGI are non-negative. Also, current study findings show that utilizing a domain-specific lexicon produces better outcomes than using a general dictionary. The approach discussed above try to discover the similarity while utilizing the content filtering algorithm, which is why we can use recommendation systems to forecast tweets with a strong association with the trust domain and the set of tweets that are comparable in this context [12].

The primary goal of this study is to identify financial and non-financial tweets in order to forecast the polarity of financial tweets. We believe that sentimental analysis of tweets can predict financial tweets since it is more meaningful and closely related to how people read financial data when deciding to invest. We describe an original method for performing financial sentiment analysis that combines the latent dirichlet allocation (LDA) and convolutional neural network (CNN) algorithms. This work adds to the literature two contributions. The research begins by outlining how emotional analysis can be used to predict polarity in financial tweets. Additionally, the study offers a sentiment categorization model based on LDA and CNN algorithm. On the basis of a reference financial data set, the efficiency of the suggested sentiment categorization model is evaluated. In order to highlight the model's utility, it is also linked to other cutting-edge approaches. The following is a breakdown of the rest of the article: the next section discusses related literature. The paper subsequently gives many examples about the proposed technique in section 3. Section 4 contains an experimental evaluation of the procedure as well as a discussion of the main findings. A summary and recommendations for further study are included at the end of the present paper.

In accordance with the India Brand Equity Foundation (IBEF), in February 2019, the Indian banking industry alone has more than 340\$ billion in assets under managing (IBEF 2019). This figure merely offers us a rough idea of the global financial sector's full size and scope. The digitization of this fast-expanding behemoth has been made possible by technological advancements. FinTech, or financial technology, is a growing field in the financial sector that has been well-defined as a combination of information technology and finance [13]. Guo and Li [14] presented a total suspended solids (TSS) model, which includes a novel baseline correlation model that does not only have a good prediction accuracy, but also decreases computing time and allows for quick decision making without prior knowledge of previous data. The R language is used to do polynomial regression, classification modeling, and lexicon-based sentiment analysis. Using the suggested baseline criterion, the resulting TSS predicts the future stock market trend in 15-time samples (30 working hours) with an accuracy of 67.22% without reference to past TSS or market data. TSS effectiveness in predicting an upward market is shown to be significantly superior than that of a falling market. TSS has a 97.87% accuracy in predicting the rising trend of the future market using logistic regression and linear discriminant analysis.

Rafay [15] looked at the interaction between FinTech and Italian SMEs. FinTech has experienced rapid progress from the point of view of development and investment, as well as how rapidly beneficial it has shown to be for the SME sector. Using data in the financial sector has become more prevalent because of

FinTech. This information is mostly in the form of texts, both organized and unstructured. As a result, textual data has been permanently a dominating and a crucial aspect in the finance, both historically and technically.

In the financial sector, unstructured text data has risen quickly. Natural language processing (NLP) and text mining have a lot of attention here. Atefeh and Khreich [16] looked into a variety of financial applications in which text mining could be useful. They came to the conclusion that there are many trends in this field, including several forms of forecasting, cybersecurity and customer relationship management, to name a few. In recent years, many novel approaches for analysis have been proposed, with artificial intelligence allowing for the analysis and forecasting of financial results based on historical data.

Since the dawn of civilisation, finance has been a powerful component of human life. From barter systems to digital currencies, finance is linked to big data, like accounts, reporting, transactions, and pricing. The use and importance of manual data processing methods has reduced over time. When it comes to researching and analyzing financial data, researchers favor digitized and automated methods. A substantial quantity of latent information exists in financial data. It could take years to manually extract latent information from a large amount of data.

Text mining advancements have made it feasible to effectively examine textual data related to the financial field. An overview of the literature on text mining for huge finance analyzing data was published by Das *et al.* [17]. The authors have organized their research around three main questions: Financial text mining methods, financial intellectual foundation, and financial data sources sectors. Vu *et al.* [18] used sentiment analysis for stock price forecasts using text mining on Twitter posts.

A dictionary is the simplest way to represent a text. In other words, one selects all available words, labels them as positive or negative, and then utilizes the resulting set to simply apply an emotional note to the phrases. The first two common lexicons used in financial news analysis are the integrated dictionary general inquirer (GI) and the text analysis tool diction. The majority of its word lists are derived from Harvard IV-4.5 dictionaries and GI word lists. Most scholars in the financial disciplines employed GI word lists and the Harvard dictionary at first since they were the first lists easily available under the numbers.

Because the financial sector has its own vocabulary, using basic sentiment analysis algorithm in finance is not acceptable because many words have different meanings. Tax and responsibility, for example, are often negative words, yet they have a neutral meaning in finance. For example, tax and responsibility, are normally positive phrases, but they have a neutral meaning in the financial world.

The word "share" has a positive connotation in general, but in the financial world, it refers to stock or a financial benefit, which is an unbiased speech. Also, while "bull" is strictly positive in the financial field, but it's neutral in general, and "bear" is negative in the financial field, but neutral in general. These instances highlight the importance of developing specific models that allow for the extraction of emotions from financial posts.

Sentiment analysis, which combines qualitative and quantitative financial performance measures, has become a study topic in finance. According to Loughran and McDonald's fundamental research, 73.8% of the negative terms on the Harvard dictionary are not normally bad in the financial field. As a result, Loughran and McDonald have developed a practiced marked vocabulary of negative, positive and neutral financial words that express feelings in financial literature. Loughran-McDonald financial sentiment dictionary (LMFSD), which is peculiar to the field, is then used in a number of studies [19], [20].

Machine learning includes deep learning, which involves training a dataset approach to generate fresh data predictions. It features a tiered design, with input data flowing to the bottom and output data coming from the top [21]. At the intermediate levels, the input data is altered by using programs to extract useful information, convert it into indicator, and then re-enter the indicator in the deeper layer to obtain modified features. Large data sets are required for training and testing deep learning-based sentiment analysis algorithms. This technique necessitates the design and development of large data sets.

Although there are a number of huge, annotated sentiment data sets that are open to the public. These data sets are used by several sentiment analysis models [22], and they perform well in related domains. However, it is challenging to apply these models across domains because each area has its own set of vocabulary for describing emotions. The financial realm has its own terminology, which necessitates a domain-specific examination of emotions. Financial market prices take into account all available information about the assets being exchanged [23], allowing investors to make timely and well-informed assessments. Stock prices and brand reputation are influenced by feelings stated in tweets and news, necessitating continual measurement and monitoring, which has become one of the most significant activities for investors. Stock prices [24], [25], changes in foreign currency and worldwide financial markets [26], [27], and corporate earnings [28] have all been predicted using sentiment analysis based on financial news in studies. The constraints of the approaches proposed in the literature are primarily centered on the detection of feelings linked to finance and the usage of a deep learning algorithm to detect tweets containing feelings about finance vs other tweets.

2. PROPOSED APPROACH

This section explores the specifics of the suggested strategy. The model architecture can be divided into three sections, as indicated in Figure 1:

- Data filtering and pre-processing
- Topic modeling by LDA
- Prediction by CNN

The first step of our proposal is directly linked to data collection and filtering, followed by the detection of feelings in tweets with a strong relationship to the financial domain, and finally, the prediction of financial and non-financial tweets using the CNN algorithm while relying on the prediction of feelings in tweets detected by the LDA algorithm. Figure 1 illustrates the overall architecture of the proposed platform, the next sections tackle the specifics of each of these steps. The first phase is described in subsection 3.1, subsection 3.2 provides the second step, and subsection 3.3 discusses the third step.

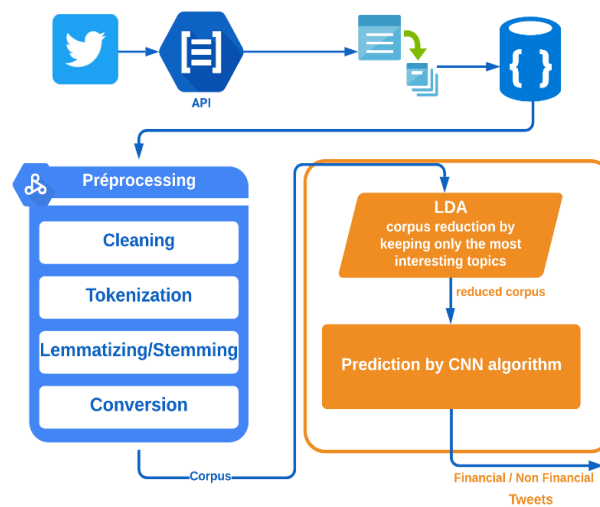


Figure 1. The overall architecture of the proposed platform

2.1. Data filtering and pre-processing

Large volumes of data are available on social media platforms, and it is expected that by 2025, this same quantity of social media data will be increased. Traditional knowledge-based data filtering systems, on the other hand, are incapable of handling large amounts of social media data, because if they can, it would take an inordinate amount of time. To retrieve relevant texts, many queries are produced in the proposed system, and then queries with a high percentage of recall are used. A first phase of automatic tagging is used to find texts and mobile content that are similar but not covered by a subject Y. This procedure improves the accuracy of document classification and retrieval of information.

The pre-processing method consists of the following steps, which present the corpus data in a more structured form to easily extract topic-related features and opinion words. In addition, these steps clean the corpus data and prepare it for word integration:

- **Cleaning:** This is a crucial step in the pre-processing process. If we are retrieving text from sources, we need to get rid of punctuation, non-alphabetic characters, and any other type of characters that might not be part of the language.
- **Tokenization:** This is the process of breaking down the raw text into smaller pieces. It breaks down the raw text into words, called tokens. These tokens help to understand the context or to develop the model for NLP.
- **Stemming and lemmatization:** Stemming and lemmatization are text normalization (or sometimes called word normalization) techniques in natural language processing that are used to prepare text, words, and documents for further processing.
- **Character conversion:** To represent generic words, the proposed system converts a sequence of characters repeated more than twice (for example, "inflaaation" becomes "inflation"). Each word is converted to lower case to avoid confusion during processing.

In summary, sentiment classification accuracy is increasing by the most prevalent words in a text, such as: is, by, and the. The uniform resource locators (URLs) in the corpus do not provide much insight into

texts or documents (tweet). As a result, the suggested method filters out URLs and commonly used terms before extracting features. Furthermore, the suggested approach employs a keyword manager to filter out stuff that isn't helpful to sentiment. Articles (a, an, the), symbols (@, date, #, and so on), and punctuation are all affected. Negation and numbers are used in some settings. Negation is important in determining how a judgment feels. According to existing research, numbers in tweets and articles are meaningless for content analysis, so they are rejected. However, numbers are used in financial text analysis to locate entities and locations (for example, \$, market, and activity). As a result, the suggested system converts negations to identify the feeling of a feature and recognizes entities using numbers. As a result, the suggested system converts negations to quickly assess a feature's mood. The Figure 2 describe step N1:

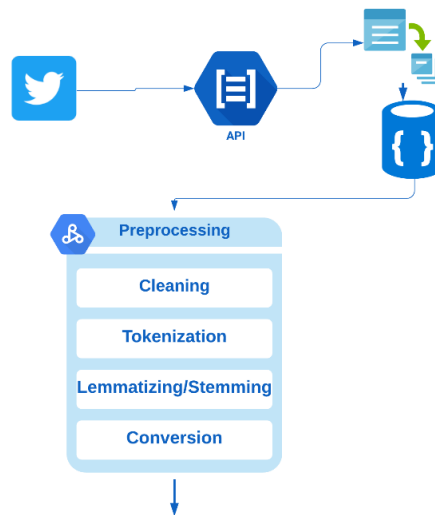


Figure 2. Data filtering and pre-processing steps

2.2. Topic modeling by LDA algorithm

The LDA model is a multiple model with each element consisting of a finite mixture of latent subjects. This technique can be used for a variety of tasks, including subject extraction, size reduction, novelty detection, summarization, similarity and relevance evaluations, and so on. The algorithm's goal is to represent short text descriptions (or any other type of data collection) that allow textual corpora to be processed while maintaining important statistical associations that are relevant for basic tasks. The LDA is used to filter tweets that contain feelings towards the field of finance. The LDA is used to filter tweets containing sentiments related to money. Although the LDA is not limited to text and may be used in other domains such as collaborative filtering, content-based picture search, and biology, it is most commonly employed for text, we will use the language of text collections to illustrate the method. Formally, we begin by defining the elements listed below:

Dictionary: consider D to be a dictionary of all possible words, with [1,...,V] as the index, and V equaling |D|. For the representation of the words, we shall use one-hot coding.

- W=(w 1,w 2,...,w n) denotes a sequence of N words, with w n denoting the nth word in the sequence.
- D=(w 1,w 2,...,w M) designates a corpus, which is a collection of M documents.

The approach's primary idea is to describe the texts as random mixes on latent subjects z n, each of which is characterized by a word distribution. The hypothesis is that k, the Dirichlet's dimensionality, is known and fixed. The probability of the word where is encoded by the matrix β∈ k×V (βi, j=p (wj=1 | zi=1), i.e., the probability of having the word if index j under the condition of being in the ith subject. Obviously, this parameter must be estimated. For the sake of simplicity, we consider that all texts in the corpus have a fixed length N_{d,d}=1,...,M (nothing is changed in the algorithm). A k-dimensional Dirichlet random variable θ takes values in the (k-1)-simplex.

$$(\alpha) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} \dots \theta_k^{\alpha_k-1} \tag{1}$$

where: α is a k-dimensional vector with α_i > 0 for each i. The distribution of dirichlets is conjugated to the multinomial distribution. According to the assumptions about the text generation process and given α and β

(aka the model parameters), the conjugate distribution of a mixture of θ topics, a set of N topics z , and a set of N words w is given by:

$$(\alpha, \beta) = p(\alpha) \prod_{n=1}^N p(\theta) p(z_n, \beta) \quad (2)$$

where: $p(z_n | \theta)$ is simply θ_i for the single i such as z_n^i (we choose the subject of each word using a multinomial distribution as discussed in the model assumptions). We can obtain the marginal distribution of documents by simply marginalizing the last distribution at θ and z .

LDA begins with the "bag of words" assumption, that the order of words in a document is not important. As shown in Figure 3. The primary goal of LDA is to understand the subject as well as the distribution of words in the data. It ignores syntax and groups semantically similar words in the same subject according to how they appear in various publications. The LDA of the machine language learning toolbox (MALLET) is used in the research to create topics on texts related to a topic Y , and different numbers of topics K are examined in the topic modeling.

The subjects developed using the LDA include common words as well as words of opinion. However, when other exams linked to topic Y include them, they also contain words that are not typical of topic Y . When a manuscript provides sufficient words, the LDA eliminates the subject-document association. You may learn semantic associations between words using the LDA. In addition, each subject is given a set quantity of words, and each document is given a set number of topics. As a result, a document's vector is insufficient.

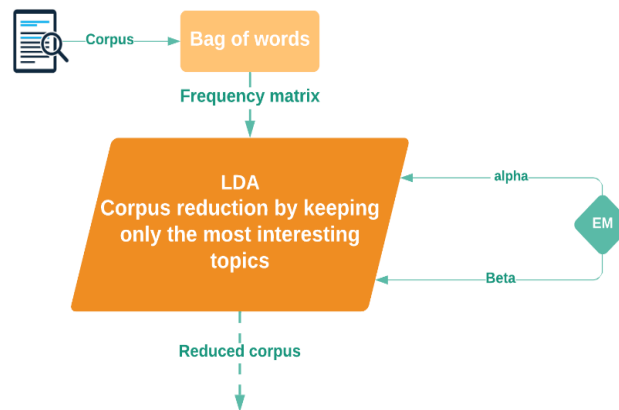


Figure 3. Topic modeling by LDA

2.3. Prediction by CNN algorithm

A convolutional neural network is a particular type of neural network. This type of network is generally used for image processing but we apply it to a text classification task [29]. These challenges can be amplified due to the enormous number of neurons required in text processing. Furthermore, working with high-dimensional input like photos or text is costly due to deep learning's hierarchical learning process. On the other side, when dealing with large amounts of data, these deep learning algorithms might become stuck. Convolutional neural networks have a number of features that make them an excellent solution for these problems. To begin with, rather than connecting to all neural network, every neuron in the first hidden layer just connects to a sample of them. This decrease in interconnection complexity decreases the risk of computing difficulties.

Second, the same feature can be detected in multiple portions of the input text by using the same weights for each of the hidden neurons. The data from the convolutional layers to the output is streamlined by a pooling layer at the network's end [30]. The convolutional neural network is one of the approaches that may be used effectively for big data analysis. In the convolutional neural network, which is one of the most powerful deep learning models, convolutional layers are utilized to filter inputs for relevant information. CNNs are a type of sensor with several layers MLP. The CNN structure in Algorithm 1 offers several advantages, including:

- Layers become deeper than usual.
- Activation functions such as ReLU, dropout, and batch normalization improve the system's calculation performance.

- With the development of the backpropagation method, the number of connections between network levels has increased.

Algorithm 1: Algorithm of CNN

```

Input: Sentence Matrix x (L × d), F filters
Output: Most Important Characteristics
for each filter f ∈ {1, ..., F} do //Get the most Important Characteristics
wj = [xj + xj+1 + ... + xj+k-1]
cj = ReLUe (wj n + b)
c = (c1 ⊕ c2 ⊕ ... ⊕ cj )
end for
    
```

- We used CNN's technique to detect financial and non-financial tweets. The application of this method is divided into two steps:

Step N 1: Learning, this step consists of learning the tweets related to the finance field.

Step N 2: Testing, this step consists of testing and detecting financial and non-financial tweets. The use of CNN for classification is made by a multilayer perception variant designed for minimal pretreatment, with little hyper parameter tuning and static vectors, the Algorithm 1 provides excellent results. In our model we use CNN on tweets and filters of different sizes to find the number of filters suitable for achieving good results. For further clarification we present the following example. The Figure 4 shows the process of this step.

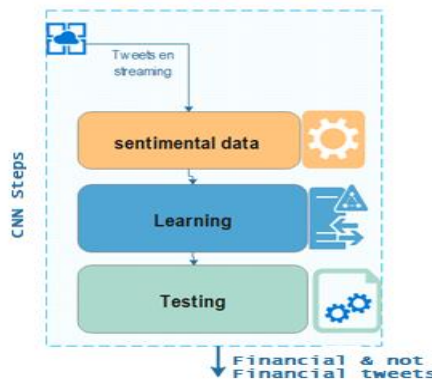


Figure 4. CNN steps

3. RESULTS AND DISCUSSION

3.1. Dataset

The data used in this experiment is taken from Twitter which provides a good environment for sentiment analysis. Twitter is microblogging, which allows a maximum of 280 characters per tweet; it's one of the best social networks. Twitter has 330 million active users per month [31]. It provides a good platform to share user's opinions and views about the trending topics. In this experiment, we thoroughly investigate over 1 million tweets collected from June 1, 2019 to June 20, 2020. In this part, we start our implementation by building an application in Twitter application programming interface (API) that reads Twitter's online feeds, this will provide us with the "keys" we will need to use the Twitter APIs. The tweet object has a long list of 'root-level' attributes, including fundamental attributes such as id, created_at, and text. Tweet objects will also have nested objects to include the user, entities, and extended_entities. Knowing that to achieve higher accuracy, several sizes of training data were tested, starting with the size 5,000 tweets to a size of 1,000,000 tweets.

3.2. Results

Twitter is known for allowing users to share their opinions in short text messages of no more than 280 characters, which are referred to as Tweets and are available to the general public. Many attributes are included in the data obtained via Twitter API, such as the message identification number and the ID number of tweets. Our research focuses on several classifiers in order to compare different classification algorithms and determine which one produces the best results. To collect the 1,000,000 tweets, we used the API streaming Twitter from June 1, 2019 to June 20, 2020. To put our strategy to the test. Table 1 displays the overall outcomes of our system. In order to determine which classifier is the most effective and efficient in

terms of efficiency measures, we looked at a number of CNN on a tweet which consists of 4 words each word is presented in 3 dimensions. The different steps of CNN's architecture for identifying tweets are depicted in Figure 4. In order to evaluate the proposed approach, we compared it with the "classifiers": random forest, recurrent neural network (RNN), long short term memory (LSTM), using the same preprocessed corpus. And to obtain higher accuracy, several data sizes were tested, Table 1 illustrates the variation of the accuracy according to the size of the training data. The results obtained show that with a size that amounts to 1,000,000 tweets, the classification using our approach achieves a very good accuracy of (99%) compared to other classifiers. This confirms the superiority of our approach using the CNN Classifier. This comparison leads to an improvement in the performance of the result. Our technique of choice, is based on metrics that produce useful results, and this classification, in particular, tests the full data set with a 1.6 percent error rate. The results of our technique are depicted in Figure 4. The Figure 5 represents the results of our approach. The red color signifies the finance tweets detected by the proposed approach and the blue represents the rest.

Table 1. Performance of the feature selection on the financial dataset

Features	Algo	Accuracy	Precision	Recall	F-Measures
5,000	CNN	0.7088	0.7088	0.6988	0.7080
10,000	CNN	0.7796	0.7796	0.7896	0.8845
1,000,000	CNN	0.9999	0.9999	0.9800	0.9866
5,000	LSTM	0.6978	0.6978	0.6978	0.7078
10,000	LSTM	0.8098	0.8098	0.8098	0.8098
1,000,000	LSTM	0.8898	0.8898	0.8898	0.8898
5,000	RNN	0.7090	0.7090	0.7090	0.7190
10,000	RNN	0.7700	0.7700	0.7700	0.7200
1,000,000	RNN	0.7800	0.7800	0.7800	0.7800
5,000	Random	0.509	0.509	0.509	0.609
10,000	Random	0.6007	0.6007	0.6007	0.7007
1,000,000	Random	0.6605	0.6605	0.6605	0.6905

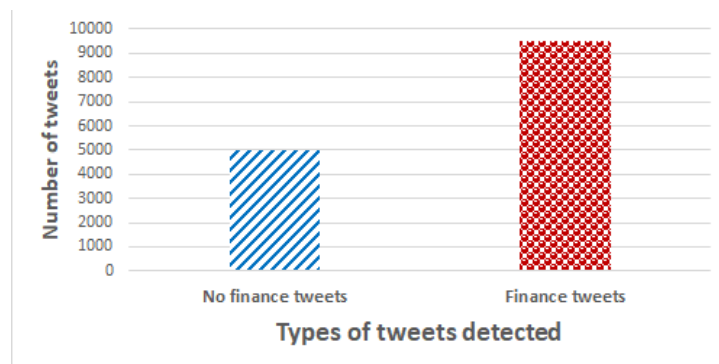


Figure 5. Results of our approach

3.3. Discussion

We tested our proposal on three example data of 5,000 (as shown in Figure 6), 10,000 (as shown in Figure 7), 100,000 (as shown in Figure 8) respectively. We noticed that the application of the CNN algorithm especially for good prediction requires voluminous data, if the test data is voluminous, the accuracy is better. The three graphs present the set of performance measures on the different datasets. We observed that each time we increase the data, the performance measures become more powerful. In terms of accuracy, our suggested solution using CNN and LDA algorithms outperformed all previous studies, Table 2 and Figure 9, describe a comparative study between existing approaches conducted by various researchers and our contribution, then we notice that the application of our approach gives better results in the calculation of accuracy 99%, compared to the other two approaches, the first based on CNN and LSTM had an accuracy of 77.12% and the second based on ANN had an accuracy of 89.5%. The utility of a detection system based on CNN and LDA algorithm for detecting financial tweets and non-financial tweets is proposed. The implementation of a new system, which is made up of three layers, results in a model of financial tweets in social networks. We can say that the proposed method improves the performance of research work in the field of sentiment analysis and brings innovation compared to the existing systems in the literature.

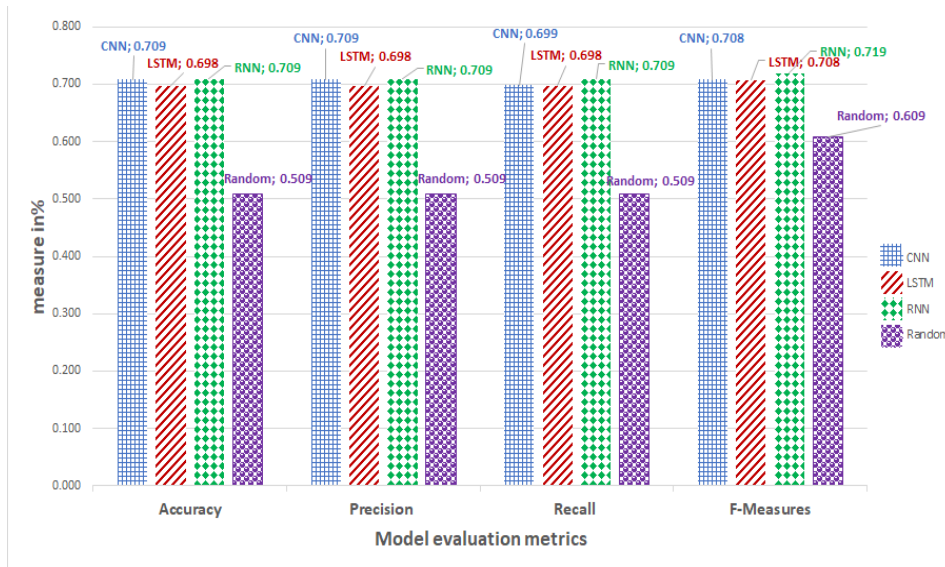


Figure 6. Performance of the feature extraction on the financial dataset 5,000

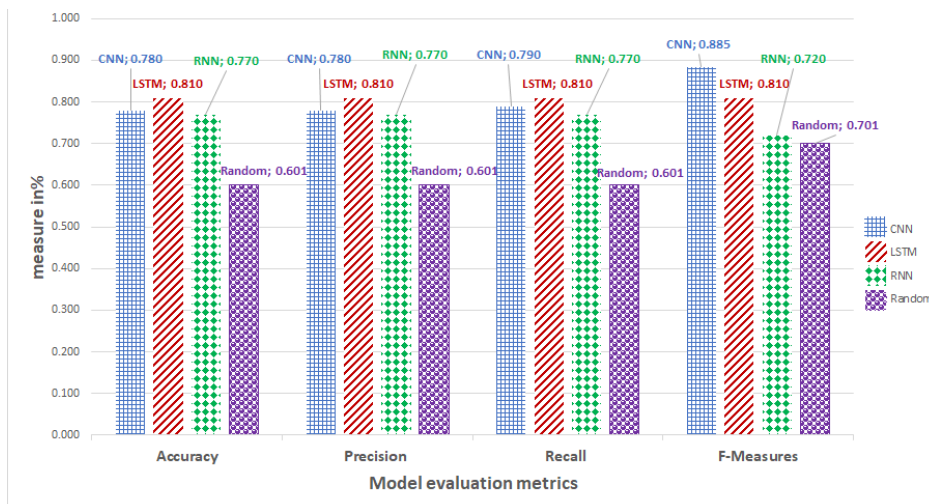


Figure 7. Performance of the feature extraction on the financial dataset 10,000

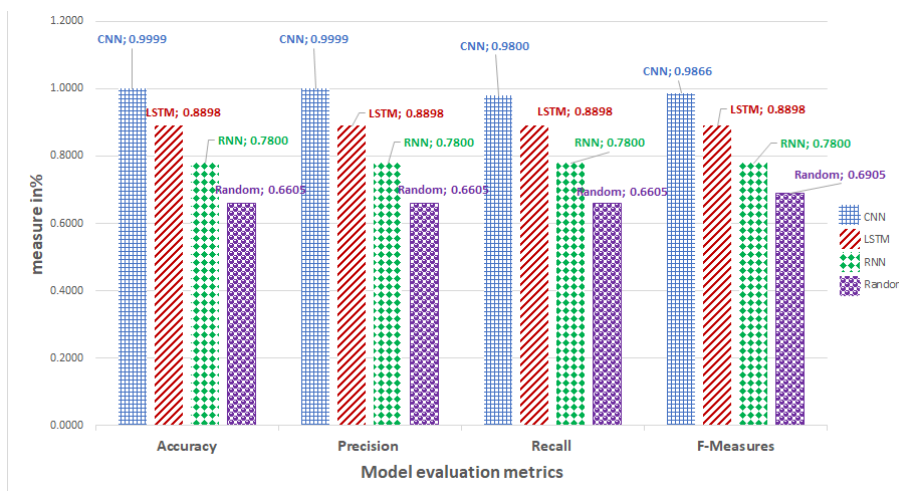


Figure 8. Performance of the feature extraction on the financial dataset 1,000,000

Table 2. Methods of comparison

Title	Methodology	Accuracy
[20]	ANN algorithm text mining	77.12%
[22]	CNN+LSTM	89.5%
Our proposal system	CNN+LDA	99%

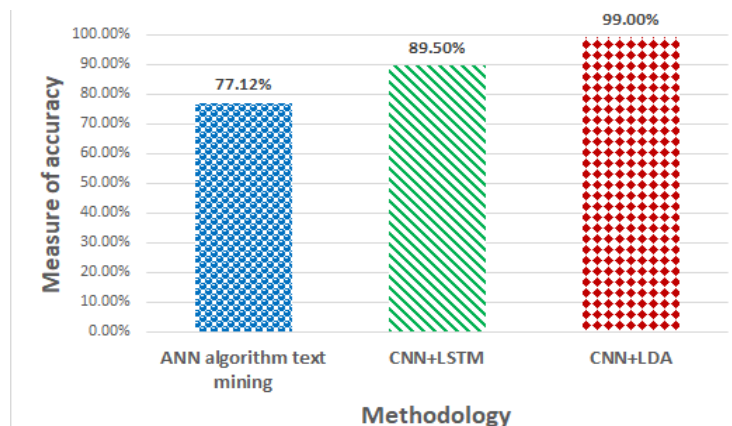


Figure 9. Comparative graph

4. CONCLUSION

The capacity to find word semantics and associations is enabled by deep learning's hierarchical learning process. Because of these qualities, deep learning is among the most recent brands for sentiment analysis. Following our findings, we demonstrate that convolutional neural networks (CNN) is able to outperform data mining in sentiment analysis. Using CNN, we can extract sentiment from a document quickly using n-grams. Convolution layers, in which each computer unit responds to a limited portion of incoming data, take advantage of the inherent data structure that exists in a document. Among the numerous deep learning approaches used in sentiment analysis, our results prove that the Convolutional Neural Network outperforms other algorithms. Convolutional neural networks have significantly higher accuracy than other models. We may use CNN to retrieve tweets linked to finance based on our findings. Only a few people in the social finance industry have the ability to correctly predict the stock market. We can forecast the market's future evolution by utilizing CNN to anticipate their attitude. Indeed, we have produced better predictability of tweets, the intelligent model is developed using Python language. Machine learning classifiers are used to evaluate the proposed word embedding system, the method achieves 99% accuracy, which shows that the proposed approach is effective for sentiment classification.




Then, we evaluated the performance of our model by performing a comparative performance analysis against different classifiers. The results obtained reveal that the classification using our approach reaches a good accuracy (99%) compared to the classifiers: random (66.05%), LSTM (88.98%), RNN (78%). In this experiment, this confirms the superiority of our approach using the filter and the CNN model. In future work, we will consider the use of articles and reports that deal with the field of finance to predict the evaluation of resources as well as the application of different data sources for prediction. Furthermore, we will develop applications using the proposed model to predict the stock prices value and movement. Whilst most data sources used for predicting the stock price trend are quantitative, prediction by analyzing financial news articles in order to improve prediction efficiency, will be our next future work.

REFERENCES




- [1] N. Sinha, S. Saxena, and K. Joshi, "Sentiment Analysis of Facebook Posts using Hybrid Method," *International Journal of Recent Technology and Engineering*, vol. 8, no. 2, pp. 2421–2428, Jul. 2019, doi: 10.35940/ijrte.B1969.078219.
- [2] S. E. Mendili, "Big data processing platform on intelligent transportation systems," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 4, pp. 1099–1109, Sep. 2019, doi: 10.30534/ijatcse/2019/16842019.
- [3] T. Loughran and B. McDonald, "Textual Analysis in Accounting and Finance: A Survey," *Journal of Accounting Research*, vol. 54, no. 4, pp. 1187–1230, Jun. 2016, doi: 10.1111/1475-679X.12123.
- [4] T. Loughran and B. McDonald, "The Use of Word Lists in Textual Analysis," *Journal of Behavioral Finance*, vol. 16, no. 1, pp. 1–11, Jan. 2015, doi: 10.1080/15427560.2015.1000335.
- [5] P. C. Tetlock, M. Saar-Tsechansky, and S. Macskassy, "More Than Words: Quantifying Language to Measure Firms' Fundamentals," *Journal of Finance*, vol. 63, no. 3, pp. 1437–1467, May 2008, doi: 10.1111/j.1540-6261.2008.01362.x.

- [6] P. C. Tetlock, "Giving content to investor sentiment: The role of media in the stock market," *Journal of Finance*, vol. 62, no. 3, pp. 1139–1168, May 2007, doi: 10.1111/j.1540-6261.2007.01232.x.
- [7] N. J. Ferguson, D. Philip, H. Y. T. Lam, and J. M. Guo, "Media Content and Stock Returns: The Predictive Power of Press," *Multinational Finance Journal*, vol. 19, no. 1, pp. 1–31, Mar. 2015, doi: 10.17578/19-1-1.
- [8] Q. Li, T. Wang, P. Li, L. Liu, Q. Gong, and Y. Chen, "The effect of news and public mood on stock movements," *Information Sciences*, vol. 278, pp. 826–840, Sep. 2014, doi: 10.1016/j.ins.2014.03.096.
- [9] X. Li, H. Xie, L. Chen, J. Wang, and X. Deng, "News impact on stock price return via sentiment analysis," *Knowledge-Based Systems*, vol. 69, no. 1, pp. 14–23, Oct. 2014, doi: 10.1016/j.knsys.2014.04.022.
- [10] T. Loughran and B. McDonald, "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *Journal of Finance*, vol. 66, no. 1, pp. 35–65, Jan. 2011, doi: 10.1111/j.1540-6261.2010.01625.x.
- [11] P. J. Stone, R. F. Bales, J. Z. Namenwirth, and D. M. Ogilvie, "The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information," *Behavioral Science*, vol. 7, no. 4, pp. 484–498, Jan. 2007, doi: 10.1002/bs.3830070412.
- [12] F. Elmendili and Y. E. Bouzekri El Idrissi, "A framework for spam detection in twitter based on recommendation system," *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 5, pp. 85–96, Oct. 2020, doi: 10.22266/ijies2020.1031.09.
- [13] L. Zavolokina, M. Dolata, and G. Schwabe, "The FinTech phenomenon: antecedents of financial innovation perceived by the popular press," *Financial Innovation*, vol. 2, no. 1, Dec. 2016, doi: 10.1186/s40854-016-0036-7.
- [14] X. Guo and J. Li, "A Novel Twitter Sentiment Analysis Model with Baseline Correlation for Financial Market Prediction with Improved Efficiency," in *2019 6th International Conference on Social Networks Analysis, Management and Security, SNAMS 2019*, Oct. 2019, pp. 472–477, doi: 10.1109/SNAMS.2019.8931720.
- [15] A. Rafay, "Preface: FinTech as a Disruptive Technology for Financial Institutions," *SSRN Electronic Journal*, 2019, doi: 10.2139/ssrn.3376358.
- [16] F. Atefeh and W. Khreich, "A survey of techniques for event detection in Twitter," *Computational Intelligence*, vol. 31, no. 1, pp. 133–164, Sep. 2015, doi: 10.1111/coin.12017.
- [17] S. Das, X. Sun, and A. Dutta, "Text mining and topic modeling of compendiums of papers from transportation research board annual meetings," *Transportation Research Record*, vol. 2552, no. 1, pp. 48–56, Jan. 2016, doi: 10.3141/2552-07.
- [18] T. T. Vu, S. Chang, T. H. Quang, and N. Collier, "An Experiment in Integrating Sentiment Features for Tech Stock Prediction in Twitter," *Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data*, vol. 3, pp. 23–38, 2012, Accessed: Jan. 19, 2022. [Online]. Available: <https://aclanthology.org/W12-5503>.
- [19] M. Azmi Shabestari, K. Moffitt, and B. Sarath, "Did the banking sector foresee the financial crisis? Evidence from risk factor disclosures," *Review of Quantitative Finance and Accounting*, vol. 55, no. 2, pp. 647–669, Nov. 2020, doi: 10.1007/s11156-019-00855-y.
- [20] I. K. Nti, A. F. Adekoya, and B. A. Weyori, "A systematic review of fundamental and technical analysis of stock market predictions," *Artificial Intelligence Review*, vol. 53, no. 4, pp. 3007–3057, Aug. 2020, doi: 10.1007/s10462-019-09754-z.
- [21] O. J. Ying, M. M. A. Zabidi, N. Ramli, and U. U. Sheikh, "Sentiment analysis of informal malay tweets with deep learning," *IAES International Journal of Artificial Intelligence*, vol. 9, no. 2, pp. 212–220, Jun. 2020, doi: 10.11591/ijai.v9.i2.pp212-220.
- [22] A. Yenter and A. Verma, "Deep CNN-LSTM with combined kernels from multiple branches for IMDb review sentiment analysis," in *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, UEMCON 2017*, Oct. 2017, vol. 2018-Janua, pp. 540–546, doi: 10.1109/UEMCON.2017.8249013.
- [23] B. G. Malkiel, "The efficient market hypothesis and its," *Journal of Economic Perspectives*, vol. 17, no. 1, pp. 59–82, Feb. 2003, doi: 10.1257/089533003321164958.
- [24] B. R. Upreti, P. M. Back, P. Malo, O. Ahlgren, and A. Sinha, "Knowledge-driven approaches for financial news analytics," in *Network Theory and Agent-Based Modeling in Economics and Finance*, Springer Singapore, 2019, pp. 375–404, doi: 10.1007/978-981-13-8319-9_19.
- [25] S. M. Grant-Muller, A. Gal-Tzur, E. Minkov, S. Nocera, T. Kuflik, and I. Shoor, "Enhancing transport data collection through social media sources: Methods, challenges and opportunities for textual data," *IET Intelligent Transport Systems*, vol. 9, no. 4, pp. 407–417, May 2015, doi: 10.1049/iet-its.2013.0214.
- [26] S. F. Crone and C. Koeppl, "Predicting exchange rates with sentiment indicators: An empirical evaluation using text mining and multilayer perceptrons," in *IEEE/IAFE Conference on Computational Intelligence for Financial Engineering, Proceedings (CIFER)*, Mar. 2014, pp. 114–121, doi: 10.1109/CIFER.2014.6924062.
- [27] C. Curme, H. E. Stanley, and I. Vodenska, "Coupled Network Approach To Predictability of Financial Market Returns and News Sentiments," *International Journal of Theoretical and Applied Finance*, vol. 18, no. 7, p. 1550043, Nov. 2015, doi: 10.1142/S0219024915500430.
- [28] J. Lee, D. Jang, and S. Park, "Deep learning-based corporate performance prediction model considering technical capability," *Sustainability (Switzerland)*, vol. 9, no. 6, p. 899, May 2017, doi: 10.3390/su9060899.
- [29] S. Aich, S. Chakraborty, and H.-C. Kim, "Convolutional neural network-based model for web-based text classification," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 6, pp. 5785, Dec. 2019, doi: 10.11591/ijece.v9i6.pp5785-5191.
- [30] J. Gu *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 354–377, May 2018, doi: 10.1016/j.patcog.2017.10.013.
- [31] K. Arun and A. Srinagesh, "Multi-lingual Twitter sentiment analysis using machine learning," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 6, pp. 5992–6000, Dec. 2020, doi: 10.11591/ijece.v10i6.pp5992-6000.




BIOGRAPHIES OF AUTHORS

Prof. Dr. Aattouchi Issam    is an assistant Professor in National School of Business and Management, Ibn Tofail University, Morocco. He obtained his state engineer diploma in Computer Engineering from Cadi Ayyad University, Morocco, and he gained his Ph.D in computer science from Ibn Tofail University. His research interests include artificial intelligence, big data analytics, machine learning, human machine interfacing, and application of machine learning in finance. He can be contacted at email: aattouchi.issam@uit.ac.ma.






Prof. Dr. Ait Kerroum Mounir    is a Higher education Professor in National School of Business and Management, Ibn Tofail University, Morocco. He obtained his M.Eng. in Computer Science and Telecommunications, and he gained his Ph.D. in IT and Telecommunications from Mohammed V University of Rabat, Morocco. His research interests include artificial intelligence, pattern recognition, deep learning, classification of hyperspectral images, classification of medical images, recognition of arabic manuscript text. He can be contacted at email: aitkerroum.mounir@uit.ac.ma.



Prof. Dr. El Mendili Saida    is an assistant Professor in Institute of Sport Professions, Ibn Tofail University, Morocco. she obtained her state engineer diploma in Computer Engineering from Cadi Ayyad University, Morocco, and she gained her Ph. D in computer science from Ibn Tofail University. Her research interests include artificial intelligence, Big Data analytics, machine learning, smart city. She can be contacted at email: elmendili.saida@uit.ac.ma.



Prof. Dr. El Mendili Fatna    is an assistant Professor at University Moulay Ismail of Meknes, she received the MSc and Ph. D degrees in Computer Sciences from University Ibn Tofail in 2015 and 2020 respectively. Member of Image laboratory in higher school of technology and associate Member of Engineering Science Laboratory in ENSA Kenitra. Technical program member and chair on several international conferences. her research area concentrates on security in social network, Big Data analytics, recommendation Systems and data security. She can be contacted at email: f.elmendili@gmail.com.