International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

Financial sentiment analysis of tweets based on deep learning approach

Aswetha. M, Danwin shaju, Mrs. Sangeetha Priya
Department of Data Science, Sri Krishna Adithya College of Arts and Science, Coimbatore-41

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 feautures from big data.

Keywords: Sentiment Analysis, Natural Language Processing (NLP), Text Mining, Big Data Analytics.

I. INTRODUCTION

We can see a significant increase in user-generated comment 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 decisionsupport 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].

II. 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 tackles 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.

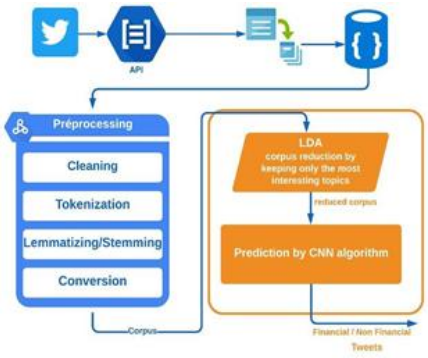

Figure 1. The overall architecture of the proposed platform

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

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 βϵ 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 Nd,d=1,...,M (nothing is changed in the algorithm). A k-dimensional Dirichlet random variable θ takes values in the (k-1)-simplex.

$$(\alpha) = \prod\Gamma(\sum ik=1\ \alpha i)\ \theta 1\alpha 1-1 \ldots \theta k\alpha k-1 \qquad (1)$$
$$k\ \Gamma(\alpha i)$$
$$i=1$$

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:

$$(\alpha, \beta) = p(\alpha) \prod Nn=1\ p(\theta)p(zn, \beta) \qquad (2)$$

where: p(z_n | θ) is simply θi for the single i such as z_ni (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. 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.

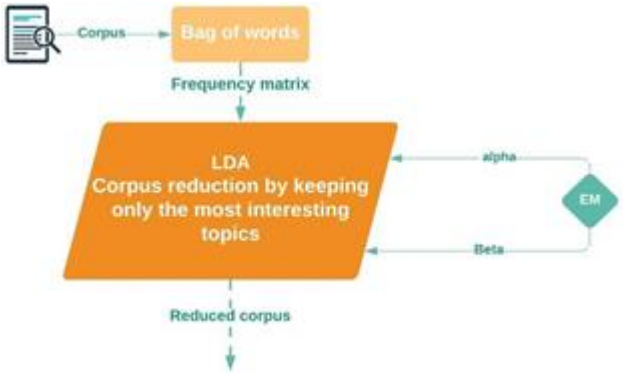

Figure 3. Topic modeling by LDA

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

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 oftesting 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.

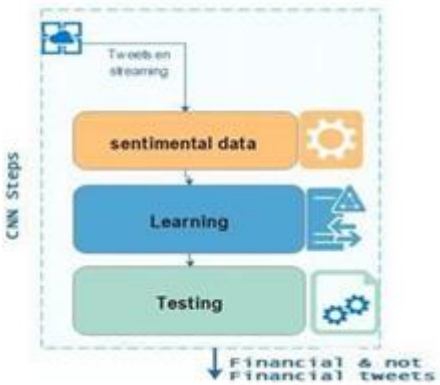

Figure 4. CNN steps

III. RESULTS AND DISCUSSION

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.

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

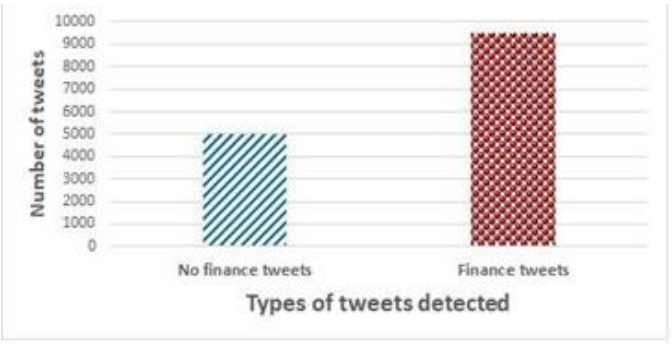

Figure 5. Results of our approach

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.

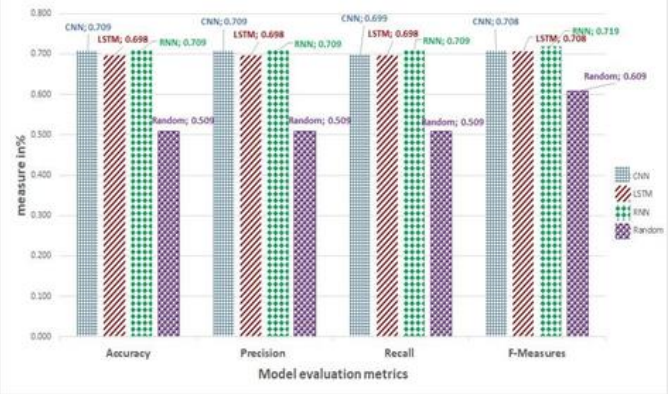

Figure 6. Performance of the feature extraction on the financial dataset 5,000

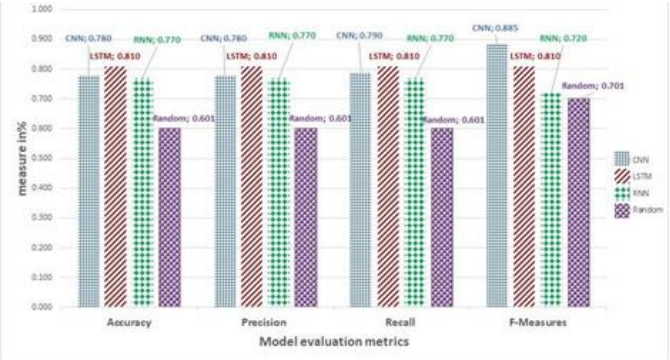

Figure 7. Performance of the feature extraction on the financial dataset 10,000

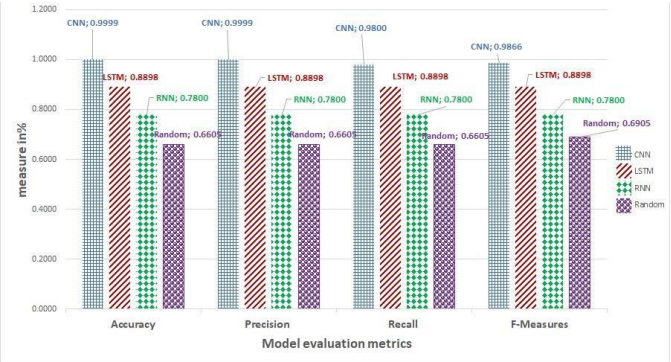

Figure 8. Performance of the feature extraction on the financial dataset 1,000,000

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%

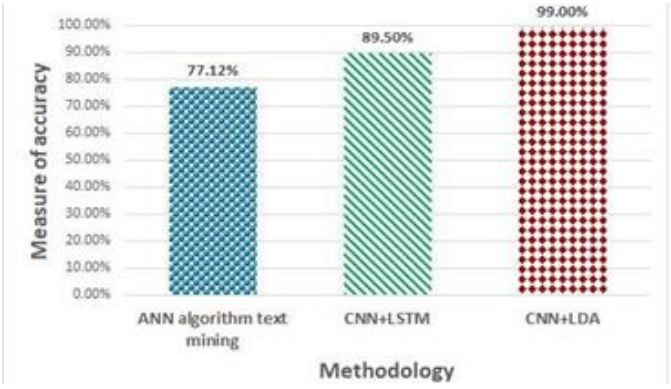

Figure 9. Comparative graph

IV. 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.knosys.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.