Journal of Discrete Mathematical Sciences & Cryptography
ISSN 0972-0529 (Print), ISSN 2169-0065 (Online)
Vol. 27 (2024), No. 7, pp. 2053–2065
DOI : 10.47974/JDMSC-2079

Discrete financial in sentimental analysis using exploring patterns and trends

Kamal Upreti *
Department of Computer Science
CHRIST University
Delhi NCR Campus
Uttar Pradesh 201003
India

Anmol Kapoor †
Department of Computer Science and Engineering
Maharaja Surajmal Institute of Technology
Janakpuri
New Delhi 110058
India

Akhilesh Tiwari §
School of Business and Management
CHRIST University
Delhi NCR Campus
Uttar Pradesh 201003
India

Himanshu Gupta ‡
Department of Computer Engineering
Syracuse University
Woodridge
New Jersey 13244
United States

* *E-mail:* kamalupreti1989@gmail.com (Corresponding Author)
† *E-mail:* kapoor.anmol98@gmail.com
§ *E-mail:* akhileshmatiwari1@gmail.com
‡ *E-mail:* himanshugpt13@gmail.com

Virendra Singh Kushwah [@]
School of Computer Science & Engineering
VIT Bhopal University
Bhopal-Indore Highway, Bhopal
Madhya Pradesh 466114
India

Jyoti Parashar [#]
Department of Computer Science and Engineering
Dr. Akhilesh Das Gupta Institute of Professional Studies
Shashti Park
New Delhi 110053
India

Divya Gangwar [$]
Department of Management
Babu Banarasi Das Institute of Technology and Management
Dr. A.P.J Abdul Kalam Technical University
Uttar Pradesh 226028
India

Nishant Kumar [^]
School of Business and Management
CHRIST University
Bangalore
Karnataka 560029
India

Abstract

In today's rapidly evolving financial environment, it's crucial for investors and decision-makers to effectively analyze stakeholder communications to gain valuable insights. This research conducts a comprehensive evaluation of a range of models that utilize machine learning, such as CNN (Convolutional Neural Network), LR (Logistic Regression), Doc2vec, and LSTM (Long Short-Term Memory), to determine their efficacy in interpreting investor's sentiments and predicting business assessments and trading dynamics. The justification for preferring deep neural architectures compared to conventional data analysis lies in the challenge of handling extensive amounts of diverse and unorganized data. Deep learning techniques have shown impressive capacity in automatically detecting complex characteristics

[@] *E-mail:* **kushwah.virendra248@gmail.com**
[#] *E-mail:* **jyoti.parashar123@gmail.com**
[$] *E-mail:* **divyagangwar2013@gmail.com**
[^] *E-mail:* **nishantkumar00@gmail.com**

and unveiling concealed patterns within written records, rendering them well-suited for sentiment analysis in financial dialogue. This research questions the notion that depending exclusively on data from a solitary origin leads to persistently effective investment moves. In fact, stakeholder communication is impacted by numerous influential elements, leading to diverse sentiments and sentiments. Through our comparative assessment, we aim to illuminate how various deep learning models can adeptly capture the intricate nuances of sentiment within fiscal messaging.

Subject Classification: *Primary 68R01, Secondary 97N70.*

Keywords: *Market sentiment, Investment prediction, Text mining, Financial forecasting.*

1. Introduction

The internet, serving as a large network, connects millions of devices and people worldwide, aided by the rapid rise of social media platforms. This expansion has created several possibilities for people with diverse backgrounds and experience to share ideas and interact with one another. Numerous websites, such as Wikipedia, take advantage of the internet's global reach to provide visitors with assistance and information. Furthermore, there are other platforms where visitors can communicate with specialists, with investment being a popular topic of discussion. Financial leaders like Lehman Brothers and Goldman Sachs have a long history of offering financial advice, but independent traders and investors can now readily interact and communicate online. Popular economic social networking sites such as StockTwits and Seeking Alpha appeal to stock market enthusiasts by allowing members to engage with one another and access data to help them better their investment strategy [1].

With breakthroughs in information technology, particularly the internet, a plethora of digital platforms have been widely available, making it easier for traders to stay up to date on stock market events. Analyzing the vast amount of data involved in investments can be costly, necessitating a professional understanding of economics in order to extract the most relevant information. Investors are likely to be impacted in different ways depending on the efficiency of these sources. Stock prices are chiefly affected by three main factors: firstly, foundational data derived from organization-targeted disclosures, which exert considerable influence on investors' tactics; secondly, public opinion, which can sway investors' sentiments and thus affect their judgment procedures; and third, the influence of press reporting on enterprises, which fluctuates day to day [2].

Economic technological platforms act as bridges between individuals, corporations, and organizations, promoting the interchange of ideas and skills. These systems generate a large amount of unstructured information, known as Big data, that can be used in decision-making processes. In today's world, Big Data is regarded as a valuable instrument for predictive analysis due to its massive production and relatively low acquisition cost [3]. Our study focuses on leveraging deep learning approaches to analyze stakeholder communications and forecast company valuations and market developments.

2. Related work

In the realm of deep learning applications, minimal focus has been dedicated to exploring the potential of deep learning frameworks specifically for assessment. Numerous expansive fields of data, such as Visual recognition in machines and voice authentication, have illustrated the effectiveness of employing deep learning to optimize classification modeling outcomes. The multi-level fine-scaled sentiment detection and ambivalence management approach, which Wang et al. [4] introduced, is a novel sentiment analysis scheme that emphasizes the nuanced analysis of both positive and negative attitudes.

The capacity of RNNs to handle gradients in longer texts has greatly accelerated the field's progress in Natural Language Processing (NLP). Examples of RNNs that meet these demands are LSTM and GRU. However, Basiri et al. [5] brought to light problems that occurred in earlier sentiment analysis systems using RNNs, such as high-dimensional outputs and evaluating all words identically. Because of their effective learning and representation capabilities, pre-trained Transformer based architectures such as GPT-3 and BERT have gained significant traction and revolutionized language problems as illustrated by Devlin et al. [6]; Floridi and Chiriatti [7]. Li et al. [8] introduced Neural Tensor Networks (NTWs) have garnered interest because to their capacity to identify various associations among elements, making them useful tools for tasks related to natural language processing (NLP) for instance sentiment analysis.

Picasso et al. [9] combined sentiment elements taken from news articles, market data, and fundamental data to create a time series classification method for NASDAQ100 portfolio movements. This strategy proved successful, especially in High-Frequency Trading simulations. In addition, several recent investigations have explored the use of deep learning models with tweets in accordance with Jeong et al. [10] research;

sentiment analysis of customer-generated social networking information to identify opportunities for advancement of products as demonstrated by Pham et al. [11]; and the sentiment analysis of movie reviews to improve recommendation systems as illustrated by Wang et al. [12]. Beyond these discrete techniques, Stai et al.'s [13] work portrays holistic models known as lifetime learning algorithms have surfaced, which update knowledge bases continually to boost the reliability of sentiment analysis.

3. Methodology

The structured flowchart depicted in Figure 1 establishes a systematic methodology for leveraging analytics-driven observations in equity market assessment. The procedure begins with "News Gathering," in which relevant media coverage is routinely compiled. Subsequently, the data is subjected to "Text Cleaning" to guarantee consistency and suitability for further analysis. The "Sentiment Analysis" step assesses the emotional tone of news material, providing useful information on market sentiment. Following that, "Feature Creation" entails generating new information characteristics that capture critical market indices, whereas "Temporal Aggregation of Features" investigates how these attributes evolve across time periods. The following step, "Data Integration," is combining all pertinent datasets into a single entity. Moving forward, "Classification of Stock Movement" categorizes fluctuations in stock prices utilizing a variety of ML approaches, such as LSTM, Doc2Vec, LR and CNN.

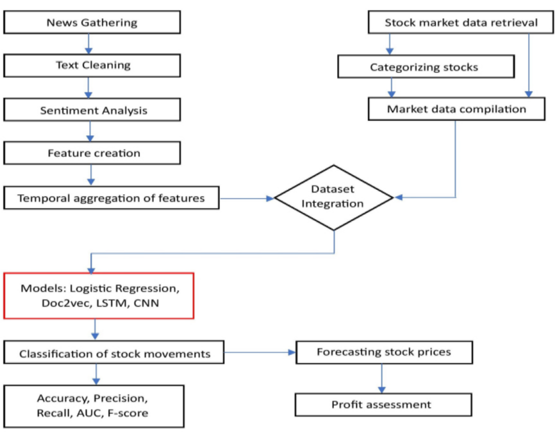

Figure 1

Stock Market Analysis Pipeline

i) *Logistic Regression*

In the time of data abundance, the synergy between AI-driven approaches and LR is essential due to the unprecedented volumes of data being generated. LR is a critical link in unlocking the potential of deep learning while resolving some of its limitations. Furthermore, its accessibility aids to understand the impact of numerous variables on model outcomes, which is significant in sectors such as finance and healthcare where full disclosure is essential. Performance metrics for the model are presented in Table 1.

Table 1

Performance Metrics for LR Model

Precision(P)	0.7133
Accuracy(A)	0.7089
Recall(R)	0.6979
AUC	0.7087
F1 score	0.7055

Recurrent Neural Network

RNNs function on the premise that input information is interrelated, allowing for more accurate predictions by taking into account previous iterations of the data. RNNs accomplish this by combining information from prior steps in a sequential order, establishing a buffer to hold data from previous computations. However, vanishing gradients, a prevalent issue in deep learning, limit the efficiency of RNNs by allowing them to only grasp dependencies over small distances [14]. Using Rectified Linear Units (ReLU) as an activation function rather than sigmoid or tanh can help to tackle this challenge. LSTM (Long Short-Term Memory) is another prominent option that can overcome the limits caused by vanishing gradients.

For sequential data:

Forward Propogation: $h_t = \sigma(W_{hx}x_t + W_{hh}h_t + b_h)$, where the hidden state is represented by h_t, the input is represented by x_t, the weight matrices are t, W_{hx}, W_{hh}, the bias term is b_h, and the activation function is σ.

Backpropagation through time (BPTT): A type of backpropagation that trains RNNs by unrolling the network over time.

ii) *Convolutional neural network*

Text processing requires a huge number of neurons, which exacerbates these difficulties. Furthermore, words that appear frequently in an expression are more likely to be connected than those that do not. However, fully connected neural networks ignore the contextual relationship between distant and adjacent words. Deep Learning, with its multilayer learning approach, can be computationally demanding, particularly for complex data such as photos or text. In contrast, such deep learning algorithms may struggle to handle enormous amounts of Big Data [15, 16, 17].

Furthermore, utilizing the same values for all hidden neurons enables the network to detect similar features across different regions of the input data [18]. Furthermore, a subsampling layer located at the architecture's end simplifies the flow of information from the filtering layers, allowing information to be processed from input to output.

For text Classification:

Convolution Operation: $z_i = \Sigma_{j=1}^{k} w_j . x_{i+j-1} + b$, where the convolution output is denoted by z_i while the filter weights are represented by w_j, respectively. The input characteristics x_{i+j-1}, b is the bias term, and the filter size is represented by the variable k.

Activation functions generally include sigmoid or ReLU (Rectified Linear Unit).

Pooling operations include maximum and average pooling to minimize the dimensionality of feature maps.

The softmax function converts the output from the last layer into classification probabilities.

Loss Functions:

Cross Entropy Loss: $L(y, \hat{y}) = -\Sigma_i^{-} y_i \log(\hat{y}_i)$, where y represents the genuine label vector and $\hat{y}$ is the anticipated probability vector.

Mean Squared error: $L(y, \hat{y}) = \frac{i}{n} \Sigma_{i=2}^{n} (y_i - \hat{y}_i)^2$, where the true label is denoted by y, whereas the anticipated value is denoted by $\hat{y}$.

4. Findings and Discussion

This segment outlines our work on the usage of Deep Learning algorithms to analyze sentiment using the StockTwits corpus. We aimed to determine whether employing advanced machine learning algorithms for sentiment analysis on StockTwits improves its precision. Deep learning aims to emulate the sequential instructional process of the human mind. It detects complicated features and allows for nonlinear evaluation of Big Data. The following section summarizes the findings of 3 notable deep learning algorithms for natural language processing (NLP).

i) *Doc2vec*

To evaluate whether or not the doc2vec technique could improve the reliability of emotional tone evaluation for financial market commentators, we conducted preliminary testing on the StockTwits database. In our study, a single paragraph vector consists of two vectors: one for the distributed BOW technique and one for the DM (distributed memory) technique. The size of the window employed can also influence the doc2vec technique's accuracy, with larger windows frequently providing more precise outcomes. We considered windows of popular sizes, including 5 and 10, to analyze this aspect. Doc2vec was executed using Python's Gensim package, with sentences with a cumulative frequency< 2 being disregarded. Table 2 shows the outcomes of the analysis.

Table 2

Performance Metrics for Doc2Vec with Different Window Sizes

Window-size	5	10
P	0.6096	0.6688
A	0.6203	0.6722
R	0.6681	0.6831
AUC	0.6203	0.6722
F1 score	0.6375	0.6758

As forecasted, utilizing doc2vec with a window size of ten results in somewhat higher accuracy than using a window size of five, though the difference is insignificant. However, when we compare the outcomes of LR as the base model using the StockTwits corpus from table 1 to those related to doc2vec from table 2, we observe that doc2vec may not be a viable sentiment prediction model in the StockTwits database. The ROC curve, demonstrating performance across window size of 5 and 10, is depicted and its results are compared to the ROC curve associated with LR.

ii) *Long short-term memory*

Proceeding from the previous subsection where doc2vec demonstrated limited effectiveness in gauging sentiment towards the stock market, we turn to Recurrent Neural Networks (RNNs) to potentially improve sentiment analysis accuracy on StockTwits. We investigate the utilization of LSTM networks, which have better memory capacity than regular RNNs. Using the Python Theano module [18], we implemented RNNs with an average pooling method. The pooled output was then processed through a LR layer to determine the objective category label linked alongside existing input stream. The findings are summarised in Table 3.

Table 3

Performance Metrics for LSTM

P	0.8519
A	0.6924
R	0.6572
AUC	0.7108
F1 score	0.7420

Using LSTM instead of doc2vec improves accuracy, although still remains below required levels. A careful examination of Tables 1 and 3 reveals LSTM's limitations in reliably predicting sentiment on the basis of StockTwits database, especially when LR is used as its foundation. Juxtaposes the outcomes of LSTM and LR in relation to the area under the ROC curve.

iii) *CNN*

Our technique begins by embedding sentences in low-dimensional vectors. These vectors are then convolutionalized using three, four, and five filter sizes. Dropout regularization and max pooling are also used to improve the model's robustness. Finally, we use a softmax layer for data

Table 4

Performance Metrics at Different Steps for Model Evaluation

Steps	100	2K	4K	6K	8K	10K	70K
P	0.6348	0.7789	0.7829	0.8779	0.8773	0.9169	0.9908
A	0.5701	0.7942	0.8211	0.8650	0.8892	0.9094	0.9896
R	0.3293	0.8220	0.8884	0.8483	0.9047	0.9003	0.9886
AUC	0.5701	0.7944	0.8209	0.8652	0.8890	0.9092	0.9898
F1 score	0.4339	0.7998	0.8324	0.8628	0.8909	0.9087	0.9896

classification. Table 4 summarizes performance metrics collected at various stages of the procedure.

When we compare the accuracy of LR as a benchmark from Table 1 to the outcomes obtained from implementing CNN as shown in Table 4, we can see that CNN outperforms, especially beyond the 2K step threshold. Around the 6K-step level, CNN achieves an accuracy of about 86 percent, well exceeding other systems. The AUC as CNN goes through several levels. Notably, as CNN progresses, the ROC curve continually shifts to the upper-left quadrant in the visualization plot, demonstrating an increase in prediction accuracy over time while analyzing the StockTwits dataset.

Table 5 provides an in-depth comparison of various machine learning models based on performance indicators. Each model is assessed according to important measures such as precision-P, recall-R, accuracy-A, F1 score and area under the curve (AUC). This extensive examination helps to comprehend each model's respective strengths and limitations, which aids in determining their usefulness in handling the specific issue at hand.

Upon meticulous assessment of the outcomes illustrated in Table 5, it is apparent that the CNN model exhibited superior performance across all

Table 5

Overall Performance Comparison of Different Models

Model	LR	Doc2vec	LSTM	CNN
P	0.7133	0.6688	0.8519	0.9169
A	0.7089	0.6722	0.6924	0.9094
R	0.6979	0.6831	0.6572	0.9003
AUC	0.7087	0.6722	0.7108	0.9092
F1 score	0.7055	0.6758	0.7420	0.9087

measured parameters when compared to alternative approaches. Remarkably, in Precision, Accuracy, Recall, AUC, and F1 score, the CNN model consistently outperformed other methodologies-LR, Doc2vec, and LSTM. This unequivocal dominance underscores the effectiveness of CNN in predicting sentiment within the StockTwits dataset. Therefore, the compelling results validate CNN's proficiency as a robust framework for anticipating sentiment among writers in this particular dataset.

5. Conclusion

In the realm of data analytics, deep learning emerges as a transformative force, particularly in natural language processing, owing to its adeptness at handling vast datasets. Unlike conventional techniques, deep learning models leverage complex layers of abstraction to discern intricate patterns within data, making them well-suited for tasks like detecting nuanced word meanings essential for sentiment analysis. This study underscores the superior performance of convolutional neural networks (CNNs) in gauging sentiment within the stock market domain. By employing convolution layers, CNNs efficiently capture contextual nuances embedded in textual data, surpassing traditional methods in terms of precision. This indicates CNN's ability to extract valuable insights from the language employed by authors, thereby facilitating accurate sentiment assessment towards equities. The study's outcomes not only corroborate CNN's effectiveness in sentiment classification but also underscore the pragmatic deep-learning applications in enhancing decision formulation within the ever-changing realm of financial market evaluation. Moreover, CNN's success in this study opens up intriguing avenues for leveraging sentiment analysis in predicting market behavior, leveraging the insights gleaned from skilled participants in financial networks.

References

[1] Thomas Renault, "Sentiment analysis and machine learning in finance: A comparison of methods and models on one million messages," *Digital Finance* (2019).

[2] C. Y.-H. Chen, R. Despres, L. Guo, and T. Renault, "What makes cryptocurrencies special? Investor sentiment and return predictability during the bubble," Working Paper (2019).

[3] K. Mishev, A. Gjorgjevikj, I. Vodenska, L. T. Chitkushev, and D. Trajanov, "Evaluation of sentiment analysis in finance: From lexicons to transformers," *IEEE Access*, vol. 8, pp. 131662-131682 (2020).

[4] Z. Wang, S. B. Ho, and E. Cambria, "Multi-level fine-scaled sentiment sensing with ambivalence handling," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 28, pp. 683–697 (2020).

[5] M. E. Basiri, S. Nemati, M. Abdar, E. Cambria, and U. R. Acharya, "ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis," *Future Generation Computer Systems*, vol. 115, pp. 279–294 (2021).

[6] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186 (2019).

[7] L. Floridi and M. Chiriatti, "GPT-3: Its nature, scope, limits, and consequences," *Minds and Machines*, vol. 30, pp. 681–694 (2020).

[8] Ankit Kumar, Abhishek Gupta, and Ramesh Kumar, "An improved quantum key distribution protocol for verification," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 22, no. 4, pp. 491-498 (2019).

[9] A. Picasso, S. Merello, Y. Ma, L. Oneto, and E. Cambria, "Technical analysis and sentiment embeddings for market trend prediction," *Expert Systems with Applications*, vol. 135, pp. 60–70 (2019).

[10] Ankit Kumar, Abhishek Gupta, and Ramesh Kumar, "An enhanced quantum key distribution protocol for security authentication," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 22, no. 4, pp. 499-507 (2019).

[11] D. H. Pham and A. C. Le, "Learning multiple layers of knowledge representation for aspect based sentiment analysis," *Data & Knowledge Engineering*, vol. 114, pp. 26–39 (2018).

[12] Y. Wang, M. Wang, and W. Xu, "A sentiment-enhanced hybrid recommender system for movie recommendation: A big data analytics framework," *Wireless Communications and Mobile Computing* (2018).

[13] E. Stai, S. Kafetzoglou, E. E. Tsiropoulou, and S. J. Papavassiliou, "A holistic approach for personalization, relevance feedback & recommendation in enriched multimedia content," *Multimedia Tools and Applications*, vol. 77, pp. 283–326 (2018).

[14] Mohammad Ehsan Basiri, Shahla Nemati, Moloud Abdar, Erik Cambria, and U. Rajendra Acharya, "ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis," *Future Generation Computer Systems*, vol. 115, pp. 279–294 (2021).

[15] G. S. N. Murthy, Shanmukha Rao Allu, Bhargavi Andhavarapu, Mounika Bagadi, and Mounika Belusonti, "Text based sentiment analysis using LSTM," *International Journal of Engineering Research & Technology* (IJERT), vol. 9, no. 05 (2020).

[16] Anmol Kapoor, Shreya Kapoor, Kamal Upreti, Prashant Singh, Seema Kapoor, Mohammad Alam, and Mohammad Nasir, "Cardiovascular disease prognosis and analysis using machine learning techniques," (2023).

[17] M. Rhanoui, M. Mikram, S. Yousfi, and S. Barzali, "A CNN-BiLSTM model for document-level sentiment analysis," *Machine Learning and Knowledge Extraction*, vol. 1, no. 3, pp. 832-847 (2023).

[18] Archana Kanwade, Kamal Upreti, Mohini Sardey, Sarika Panwar, Milind Gajare, and Monali Chaudhari, "Combined weighted feature extraction and deep learning approach for chronic obstructive pulmonary disease classification using electromyography," *International Journal of Information Technology* (2023).

Received March, 2024