applied sciences

Article

Comparative Study for Sentiment Analysis of Financial Tweets with Deep Learning Methods

Erkut Memiş [1], Hilal Akarkamçı (Kaya) [2], Mustafa Yeniad [1], Javad Rahebi [3] and Jose Manuel Lopez-Guede [4,*]

[1] Department of Computer Engineering, Ankara Yıldırım Beyazıt University, Ankara 06010, Türkiye; erkutmemis@gmail.com (E.M.); myeniad@aybu.edu.tr (M.Y.)
[2] Turkish Embassy Office of Educational Counsellor, 1062 Budapest, Hungary; kaya.hilal@meb.gov.tr
[3] Department of Software Engineering, Istanbul Topkapi University, Istanbul 34087, Türkiye; cevatrahebi@topkapi.edu.tr
[4] Department of Automatic Control and System Engineering, University of the Basque Country (UPV/EHU), 01006 Vitoria-Gasteiz, Spain
* Correspondence: jm.lopez@ehu.eus

Abstract: Nowadays, Twitter is one of the most popular social networking services. People post messages called "tweets", which may contain photos, videos, links and text. With the vast amount of interaction on Twitter, due to its popularity, analyzing Twitter data is of increasing importance. Tweets related to finance can be important indicators for decision makers if analyzed and interpreted in relation to stock market. Financial tweets containing keywords from the BIST100 index were collected and the tweets were tagged as "POSITIVE", "NEGATIVE" and "NEUTRAL". Binary and multi-class datasets were created. Word embedding and pre-trained word embedding were used for tweet representation. As classifiers, Neural Network, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU) and GRU-CNN models were used in this study. The best results for binary and multi-class datasets were observed with pre-trained word embedding with the CNN model (83.02%, 72.73%). When word embedding was employed, the Neural Network model had the best results on the multi-class dataset (63.85%) and GRU-CNN had the best results on the binary dataset (80.56%).

Keywords: data mining; deep learning; sentiment classification; financial; tweet; Borsa Istanbul

Citation: Memiş, E.; Akarkamçı, H.; Yeniad, M.; Rahebi, J.; Lopez-Guede, J.M. Comparative Study for Sentiment Analysis of Financial Tweets with Deep Learning Methods. *Appl. Sci.* **2024**, *14*, 588. https://doi.org/10.3390/app14020588

Academic Editor: Mustafa M. Matalgah

Received: 13 December 2023
Revised: 27 December 2023
Accepted: 7 January 2024
Published: 10 January 2024

1. Introduction

Nowadays, the most popular data sharing field is social media; therefore, social media sites accumulate huge amounts of data. Twitter [1] is a commonly used social media data sources. Twitter has nearly 700 million users and 58 million tweets on average per day [2]. People post messages that are called as tweets, which may include texts, videos, links, etc. Because of its huge popularity and usage, analyzing tweets posted by users has become more and more important. Therefore, automatically detecting tweets' sentiments is an attractive research area for many researchers.

Sentiment analysis is the outcome of people's emotions, attitudes, opinions, sentiments, etc., in their sharing's, which can be written or spoken. This concept especially focuses on polarity detection [3], which identifies negative and positive opinions in the text.

Sentiment analysis is carried out at three levels: the word or phrase level, the sentence level and the document level [4]. Generally, lexicon-based, learning-based and hybrid-based approaches [5] are used to realize sentiment classification problems. Figure 1 shows different sentiment analysis approaches and algorithms.

Tweets are, in a way, microblogs or short texts, so our sentiment analysis is performed at the sentence level. In our work, we used learning-based approaches to define sentiments in sentences. Financial news from tweets and the sentiment analysis of these tweets may contain important information or indicators for the financial or stock market. Although many studies have been conducted in English in the field of sentiment analysis and financial sentiment analysis, not many studies have been published in Turkish yet. Turkish financial tweets were collected with determined keywords from the BIST 100 index using association rule mining [6] and the tweets were tagged as “POSITIVE”, “NEGATIVE” and “NEUTRAL”. Binary datasets including only positive and negative classes, and multi-class dataset including positive, negative and neutral classes were created.

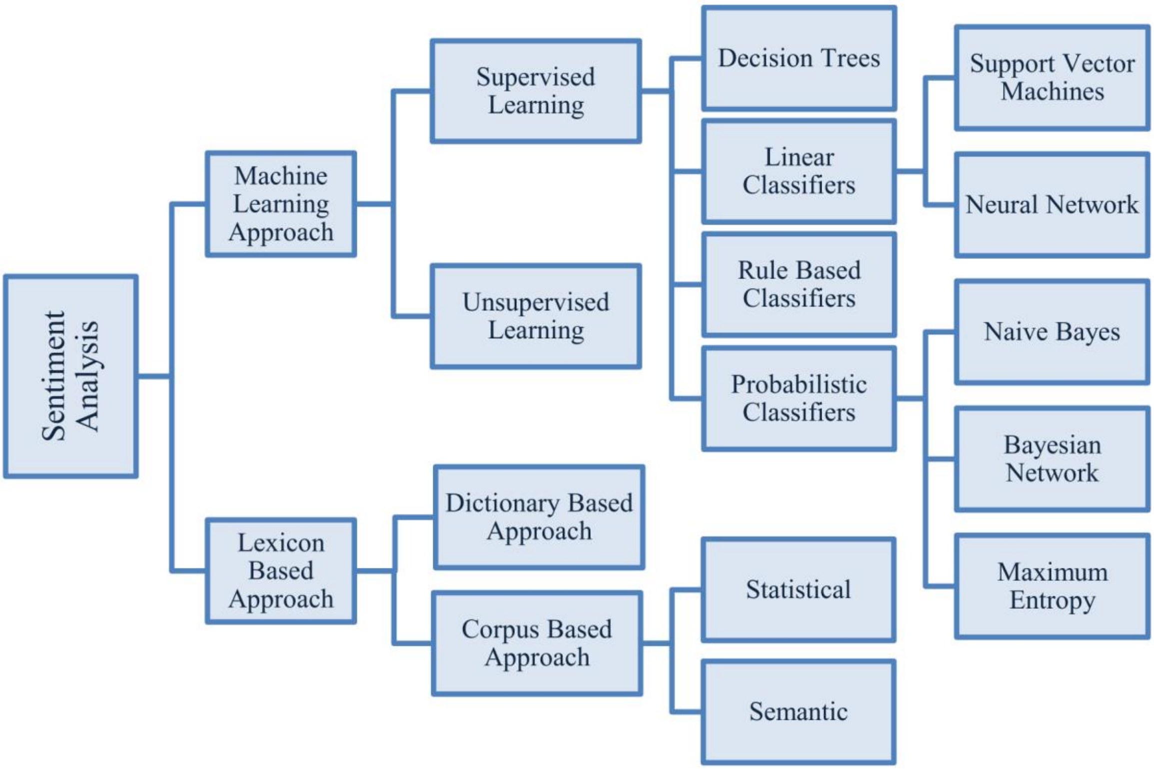

Figure 1. Sentiment analysis approaches and algorithms [5].

Noisy or unclear sentences negatively affected the sentiment classification process. In order to prepare these tweets for analysis, we used pre-processing, which included stop word removal, normalization processes, etc. The “ITU Turkish NLP Web Service API” was utilized for the Turkish text normalization process [7].

Deep learning algorithms and methods have provided great improvements in the fields of pattern recognition and image recognition. These improvements led to Natural Language Processing (NLP) researchers to be able to focus on deep learning methods. The use of dense vector representations based on Neural Networks has achieved better results for NLP tasks. The success of word embedding [8,9] and deep learning methods [10] caused the trend of using deep learning algorithms in NLP tasks. In contrast to the traditional machine-learning-based NLP systems, which use handmade features, deep learning enables automatic feature representation learning. Handmade features have several bottlenecks [11]. We used word embedding and pre-trained word embedding with fastText [12] for feature representation in our work.

Neural Network, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU) and GRU-CNN models were used as sentiment classifiers in this study. The performances of these models were evaluated based on their accuracy.

The arrangement of this paper is as follows: the introduction is included in Section 1; works related to sentiment analysis and Turkish financial tweet data are discussed in Section 2; Section 3 contains descriptions of the materials and methods used in our work; the results are presented in Section 4; and Section 5 includes our conclusions. The key highlights have been concisely outlined in Tables 1 and 2, successfully differentiating the groundbreaking contributions of this research from its practical applications for stakeholders within the financial industry.

Table 1. Originality and novelty.

Aspect	Description
Language Emphasis	Enhances sentiment analysis capabilities in Turkish, a language with limited linguistic resources, thereby broadening applicability across diverse cultural contexts.
Financial Domain	Specializes in Turkish financial tweets, offering valuable perspectives into market sentiment and investor behaviors within this specific domain.
Multiple Deep Learning Models	Evaluates and contrasts five distinct deep learning architectures tailored for Turkish financial tweets, delivering comprehensive insights into their respective performances.
Utilization of Pre-trained Embeddings	Investigates and illustrates the efficacy of utilizing pre-trained word embedding from fastText for sentiment analysis in Turkish, contributing to the optimization of the analytical process.

Table 2. Practical Implications.

Aspect	Description
Financial Market Sentiment Analysis	Assists traders, analysts and portfolio managers in making informed decisions by evaluating investor sentiment.
Risk Identification and Management	Pinpoints potential risks and opportunities in specific stocks or sectors, aiding in the formulation and execution of effective risk management strategies.
Enhanced Investor Communication	Facilitates personalized communication with investors based on social media feedback, thereby enhancing engagement and overall satisfaction.
Informed Market Research	Offers valuable insights into public perceptions of financial products, services and regulations, serving as a foundation for market research and informed product development initiatives.

2. Related Studies

The sentiment analysis of tweets related to finance can be a significant indicator for investors when analyzed and interpreted according to the stock market. Automatically determining tweets' sentiments is an attractive research area for many researchers. Feature vectors for text representation, classification techniques such as SVN, CNN, LSTM, Naïve Bayes, etc., and relations between tweets and stock markets are just a few research areas in this field. Although many sentiment analysis studies have been conducted on Twitter data, there are not enough studies on these subjects in the Turkish language and on Turkish stock markets.

Nasukawa and Yi have studied sentiment extraction for specific subjects from a document, instead of document classification [13]. Also the review of "sentiment analysis" has been reviewed in reference [14].

Almohaimeed has studied sentiment analysis on English tweets in order to predict S&P 500 index movement. He used data mining to draw out the companies affecting the S&P 500 index, in order to rank these companies and to determine patterns. In his thesis, he showed that classifier ensembles perform better than classic classifiers in the process of classifying tweets; his prediction model has an accuracy rate above 80% [15].

The relationship between the stock market index and Turkish tweets was studied by Şimşek and Özdemir. They used 113 words and eight classes for their emotion corpus. When these words were found in tweets, they count them and calculated average happiness values. They showed that the relationship between the stock market and tweet data is approximately 45% [16].

The relationship between social media and daily stock prices was investigated by Yıldırım and Yüksel. A telecommunication company from Borsa Istanbul was selected.

For a given period, daily data (opening price, closing price etc.) was collected. Sentiment analysis was applied for the same period. According the Spearman's rank correlation test results, a negative and moderate correlation exist between the daily stock price and public sentiments in tweets [17].

The prediction of exchange rate movements using tweets has been studied by Öztürk and Çiftçi. The keywords "#USD/TR", "USD/TR", "Dollar", "#Dollar" were used for tweet collection. Collected tweets' sentiments and the daily exchange rate of USD/TR were analyzed by them. They used value 1 for increasing exchange rate and 0 for the rest of the cases. They also categorized the collected tweets as Buy, Sell and Neutral. As a result, they found a remarkable relationship between the exchange rate and the sentiments of tweets [18].

Eliaçık and Erdoğan studied sentiment analysis methods on microblogging sites that use new user metrics. They proposed the measurement of the financial community's sentiment polarity on microblogging sites. In addition, they analyzed the correlation between the behavior of the Borsa Istanbul index and the mood of the financial community weekly using the Pearson correlation coefficient method [19].

Akgül, Ertano and Diri studied sentiment analysis and Twitter. They used both n-gram and lexicon methods, implementing two different models. They concluded that the lexicon method has a better performance than the n-gram method [20].

Bollen, Mao and Zeng studied stock market predictions using Twitter moods. The text content of daily tweets were analyzed by using two mood tracking tools, OpinionFinder and Google-Profile of Mood States (GPOMS). They used a Granger causality analysis and a self-organizing fuzzy Neural Network to explore their hypothesis that public mood states could be used to predict change in DJIA closing values. They found that using specific public mood dimensions remarkably improve DJIA predictions [21].

Velioglu, Yıldız and Savas studied "sentiment analysis using learning approaches over emojis for Turkish tweets". They used bag-of-words and fastText representations for evaluated sentiment classification models, including sentiment analysis performed over emojis/emoticons. Their results show that there are no notable distinctions between these models [22].

Smailovic et al. studied stream-based sentiment analysis in the financial domain. They explored the relationship between sentiments expressed in tweets related to selected companies and their stock prices movements. They used the SVM classifier for tweet categorization based on positive, negative and neutral statements. They found that there is a relationship between company-related tweets and their stock price changes, and that tweets could be used as a measure for stock price directions [23].

Bilgin and Şentürk studied "sentiment analysis of tweets based on document vectors using supervised learning and semi-supervised learning". They carried out sentiment analysis using Turkish and English tweets [24].

Ayata, Saraçlar and Özgür studied sentiment analysis using machine learning and word embedding for Turkish tweets. They used SVM and Random Forest classifiers for sentiment classification. They also used vector embedding for Turkish tweet representation. Their results show that sectoral-based tweet classification gives better results than general or non-domain tweet classification [25].

A financial tweet refers to a message shared on the Twitter platform that delves into financial subjects, encompassing discussions on stock market trends, economic news, investment strategies, tips on personal finance and updates related to cryptocurrencies. Such tweets serve the purpose of disseminating information, offering commentary and initiating conversations among individuals with an interest in the field of finance [26].

Categories of financial tweets:

Market updates: These tweets furnish current and immediate information regarding stock prices, market indices and economic indicators [21].

Analysts' perspectives: Financial analysts frequently convey their insights and predictions on Twitter, impacting investment decisions [27].

Personal finance guidance: Authorities, bloggers and individuals disseminate practical advice and strategies for effectively managing personal finances [28].

Cryptocurrency updates: Financial Twitter frequently features news and updates on cryptocurrency prices, trading activities and regulatory developments [29].

Economic insights: Economists, policymakers and journalists often share their perspectives and analyses on various economic events and policies through financial tweets [30].

Benefits of following financial tweets:

Remaining well-informed: Following financial tweets enables individuals to stay abreast of market movements, economic trends and timely news updates [31].

Gaining knowledge from experts: Following financial tweets allows individuals to gain insights and knowledge from experienced financial professionals and analysts who share their expertise [32].

Participating in conversations: Financial Twitter serves as a platform for individuals to actively participate in discussions with like-minded individuals interested in finance, facilitating the exchange of ideas and perspectives [31]. The comparative analysis of sentiment analysis in finance, with proactive recommendations are shown in Table 3.

Table 3. Comparative Analysis of Sentiment Analysis in Finance, with Proactive Recommendations.

Aspect	Comparison	Proposal
Sentiment Analysis vs. Traditional Analysis	Provides additional layer of information.	Integrate sentiment analysis results into existing analytical frameworks.
Machine Learning Models vs. Conventional Models	Leverages word embedding techniques and pre-trained embedding.	Explore the integration of machine learning models into algorithmic trading strategies.
Social Media Influence vs. Market Fundamentals	Suggests correlation between social media sentiments and stock market movements.	Consider incorporating social media analytics into risk management strategies.
Pre-trained Word Embedding vs. Customized Approaches	Better performance compared to customized word embedding approaches.	Explore pre-trained embedding for sentiment analysis.
Real-Time Sentiment Integration	Real-time sentiment analysis tools that integrate financial tweets' sentiments into trading platforms.	Implement real-time sentiment analysis tools.
Algorithmic Trading Strategies	Algorithmic trading strategies that incorporate sentiment analysis signals.	Develop and test algorithmic trading strategies that incorporate sentiment analysis signals.
Risk Management Enhancement	Consider social media sentiment as an additional risk factor.	Enhance risk management models by considering social media sentiment as an additional risk factor.
Education and Awareness	Awareness campaigns and educational programs for market participants.	Conduct awareness campaigns and educational programs for market participants.
Collaboration with NLP Experts	Collaboration with natural language processing (NLP) experts.	Financial institutions should collaborate with NLP experts.
Cross-Disciplinary Research	Cross-disciplinary research collaborations between finance professionals, data scientists and social media analysts.	Encourage cross-disciplinary research collaborations.
Continuous Model Optimization	Continuous optimization strategies for sentiment analysis models.	Implement continuous optimization strategies for sentiment analysis models.

In summary, market participants stand to gain advantages by incorporating sentiment analysis into their decision-making workflows, utilizing machine learning models and adjusting their strategies to align with the instantaneous insights offered by financial tweets.

3. Materials and Methods

We collected Turkish financial tweets discretely between 13 January 2019 and 10 March 2020 using Python, Tweepy library, Twitter API and MySQL. Collected tweets were manually tagged as positive, negative, neutral and irrelevant using our Java-based tagging program.

In the tweet pre-processing phase, using our Python code, we removed unnecessary sections of tweets, transformed tweet text to lowercase, and fixed spelling/writing errors (normalization) and restored popular abbreviations to their full forms (e.g., mrb to merhaba). ITU Turkish NLP Web Service API [7] was used for the normalization process.

Word embedding and fastText's pre-trained word embedding [12] were used as feature extractors. Deep learning algorithms—Neural Network, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU) and GRU-CNN—were used for sentiment classification. The configuration of Neural Networks, encompassing factors such as the number of hidden layers, the dimensions of layers and the choice of activation functions, was contingent upon the unique requirements posed by the task at hand and the characteristics of the dataset. Nevertheless, Table 4 furnishes broad insights into the prevalent architecture commonly employed across different categories of Neural Networks.

Table 4. A brief overview of the specific architectural features of each type of Neural Network.

Neural Network Type	Number of Hidden Layers	Layer Sizes	Activation Functions
Simple Neural Network (Binary Classification)	Typically 1–2	Varies, depends on problem complexity	Hidden: ReLU, Output: Sigmoid
Convolutional Neural Network (CNN)	Multiple convolutional and pooling layers, followed by fully connected layers	Convolutional layers: filter size determines neuron count, fully connected layers: variable	Convolutional: ReLU, Output: Sigmoid or Softmax (depending on task)
Recurrent Neural Networks (RNN)	One or more recurrent layers	Number of recurrent units (neurons) per layer	tanh or ReLU
Long Short-Term Memory (LSTM)	Multiple layers of memory cells	Number of memory cells (neurons) per layer	Specialized within memory cells (sigmoid, tanh)
Gated Recurrent Unit (GRU)	Multiple layers possible	Number of gated units (neurons) per layer	Specialized gating mechanisms (sigmoid, tanh)

3.1. Datasets

In this study, we worked on a newly created Turkish tweet dataset, tagged by us, that included 2313 tweets. The dataset had 992 POSITIVE, 629 NEGATIVE and 691 NOTR labelled tweets. We created two datasets: binary ("0-NEGATIVE", "1-POSITIVE") and multi-class ("NEGATIVE", "POSITIVE" and "NEUTRAL") datasets. Dataset distributions are shown in Figure 2.

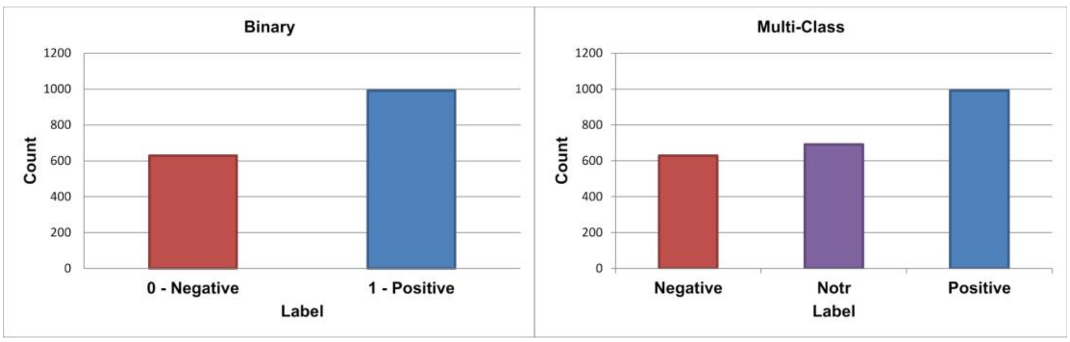

Figure 2. Binary and multi-class datasets.

3.2. Tweet Pre-Processing Phase

Before using tweets as an input in our Neural Network models, the tweets needed pre-processing. Tweet pre-processing included:

- Removing unnecessary sections of tweets (external links and usernames (signified with @sign), URL (http://...), stop words, #tags, retweets (starts with "RT"), punctuations, unnecessary whitespaces, etc.);
- Transforming characters to lowercase;
- Removing numbers;
- Correcting spelling/writing errors (normalization) and restoring popular abbreviations to their full forms (e.g., mrb to merhaba). ITU Turkish NLP Web Service API [7] was used for the normalization process.

We developed a tweet pre-processing program with Python, which processed the tweets as shown in Figure 3.

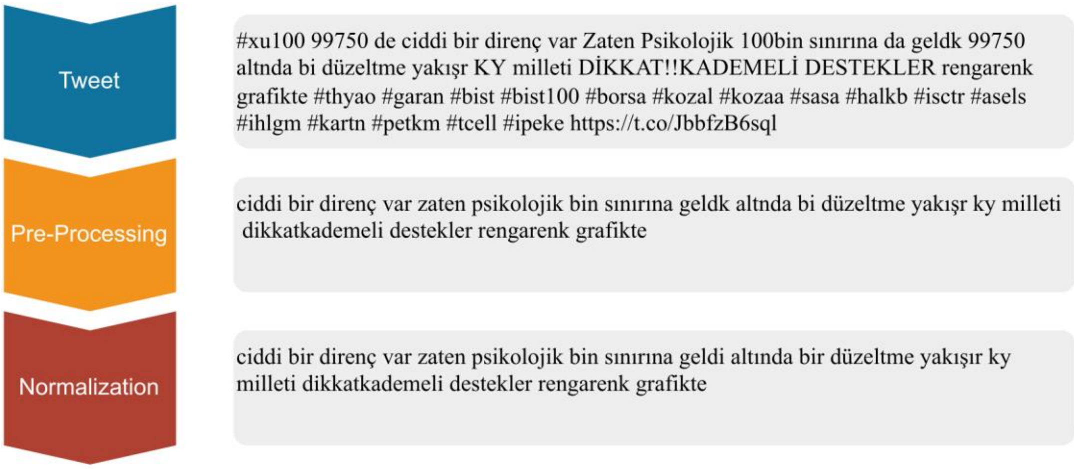

Figure 3. Pre-processing and normalization steps.

3.3. Feature Extraction

Machine learning algorithms, needing numerical values as inputs, cannot directly run on text data. The process of converting text to numerical values is called feature extraction. There are numerous types of feature extraction methods. Some popular feature extraction methods for text are Bag of Words (BoW) and word embedding. We used the word embedding approach in our work.

3.3.1. Bag of Words (BoW)

Each document is represented as a vector **d**, and each dimension of vector d consists of a unique term in the term spaces of the document collection. We express each vector **d** as

$$\mathbf{d} = (w_1, w_2, w_3, \ldots, w_n)$$

where w_i is the weight of the term of document **d**. **Boolean weighting** and **TF-IDF** are the most commonly used weighting algorithms.

Boolean weighting has a binary representation for term weight. Its weight is considered as 1 if the document consists of the term, otherwise it is considered as 0. The equation of Boolean weighting is

$$w_i = \begin{cases} \mathbf{1}, & if\ tf_i > \mathbf{0} \\ \mathbf{0}, & otherwise \end{cases}$$

where tf_i is the frequency of term i in the document [33].

The TF-IDF (Term Frequency-Inverse Document Frequency) weighting equation is as follow

$$w_i = tf_i * \mathbf{log}\left(\frac{n}{n_i}\right)$$

where, tf_i is the frequency of term i in document **d**, n is the total number of documents and n_i is the number of documents that include term i [33].

3.3.2. Word Embedding

This is a text representation in which similar words have similar representations. In other words, in a coordinate system, corresponding words are placed close to each other [34,35]. Word2vec [36], GloVe [37] and fastText [38] are the most common word embedding models. Mikolov et al. used Artificial Neural Networks (ANN) in a Word2vec model. Word2vec is based on the prediction of a word from surrounding words (Continuous Bag of Words, CBOW) or the prediction of surrounding words from a given word (Skip gram). We used word embedding and pre-trained word embedding with fastText in our study. The feature vector size was 300.

3.4. Classifier Models

Deep learning algorithms have made impressive advances in research areas like pattern recognition, image recognition, etc., in recent years. Because of deep learning algorithms' results and developments in Neural Network-based word embedding [8,9] representations, recent Natural Language Processing (NLP) research has increasingly used deep learning algorithms and word embedding instead of SVM and logistic regression techniques.

3.4.1. Convolutional Neural Networks (CNN)

Convolutional Neural Networks have impressive results in computer vision and image processing areas [39–41]. It is a model that has come to be increasingly used in NLP research. The use of of CNNs for texts first started with Collobert and Weston's research [42]. They used a look-up table to transform words into a vector representation. Firstly, the word tokenization process takes place, whereby these words are transformed into a word embedding matrix with of selected or determined dimension. After this step, the convolution process is applied to the embedding matrix with selected kernels to create a feature map. The max-pooling operation follows the convolution step to reduce the dimension of output and obtain the fixed-length output [11,43]. Figure 4 shows CNN modeling for text.

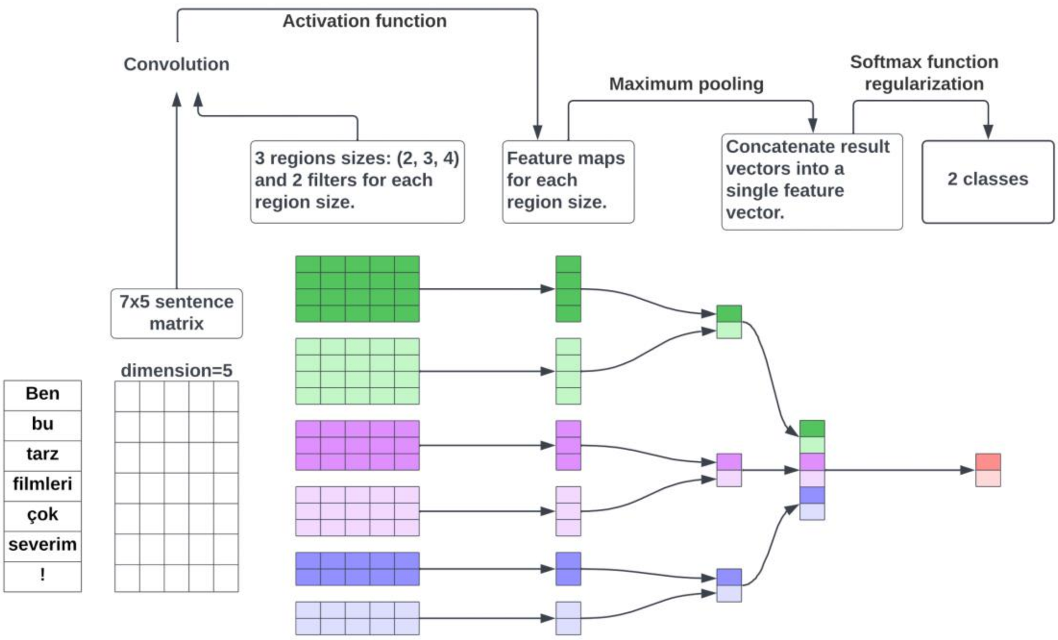

Figure 4. CNN modelling for text [11] (© 2018 IEEE. Reprinted, with permission, from IEEE Computational Intelligence Magazine).

3.4.2. Recurrent Neural Networks (RNNs)

Recurrent Neural Networks trust the principle that sequential information processing is primarily based on the Elman network [44]. An RNN recursively applies the previously computed results into a computation for every instance in an input sequence. Figure 5 shows a simple RNN structure [11,43]

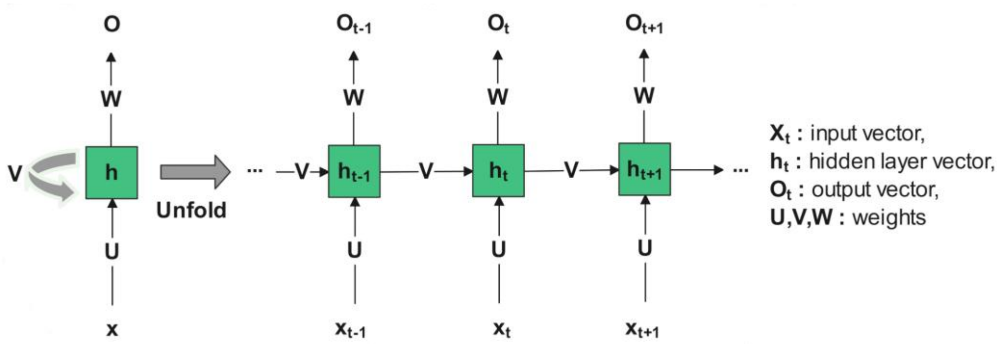

Figure 5. Simple RNN structure [11] (© 2018 IEEE. Reprinted, with permission, from IEEE Computational Intelligence Magazine).

The capacity for memorization of the previous results is the main difference or advantage of an RNN [11]. So, it is convenient for various NLP tasks like sentiment analysis, speech recognition, etc. In practice, these simple RNNs suffer from a vanishing gradient problem, which complicates the learning and tuning parameters of the preceding layers in the network [11].

This problem has led to the development of various RNN derivative models like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU).

3.4.3. Long Short-Term Memory (LSTM)

LSTM has "forget gates" in addition to the simple RNN architecture to handle vanishing and exploding problems. Figure 6 shows the LSTM structure.

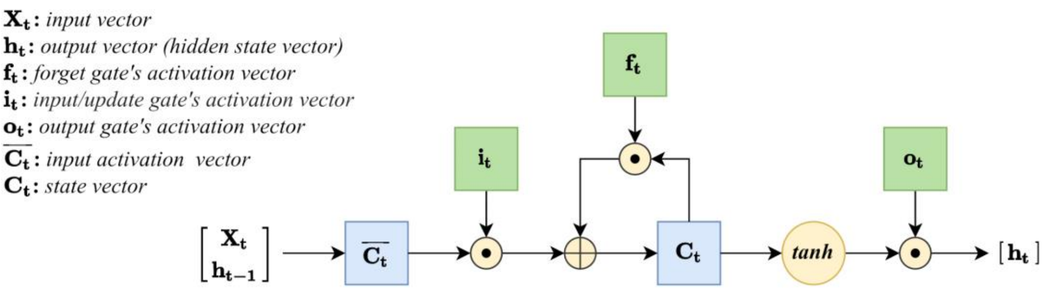

Figure 6. Long short-term memory [11] (© 2018 IEEE. Reprinted, with permission, from IEEE Computational Intelligence Magazine).

Unlike the simple RNN, LSTM back-propagates errors through a limitless number of time steps [11].

3.4.4. Gated Recurrent Unit

The GRU is another RNN derivative model. It has less complexity but a similar performance to LSTM. GRU adds reset and update gates to simple RNN. Figure 7 shows a Gated Recurrent Unit.

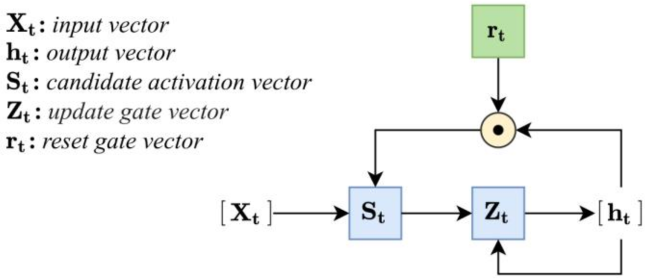

Figure 7. Gated Recurrent Unit [11] (© 2018 IEEE. Reprinted, with permission, from IEEE Computational Intelligence Magazine).

The high training accuracies (100% for some models) suggest overfitting, where the model memorizes the training data rather than learning generalizable patterns. This leads to poor performance on unseen data.

Regularization methods, such as L1, L2 and elastic net, impose penalties on excessive model complexity, serving as a deterrent against overfitting to particular data points in the training set [45].

Dropout Layers: Randomly dropping out neurons during training forces the model to rely on other features and prevents overfitting to individual neurons [46].

Balanced Dataset: An imbalanced dataset, wherein there is a prevalence of either positive or negative tweets, can result in the model exhibiting bias toward the majority class. This may lead to high training accuracy but might not ensure effective generalization [47,48].

Oversampling/Undersampling: Employing techniques such as oversampling, which involves replicating data points from the minority class, or under sampling, which entails removing data points from the majority class, aids in balancing the dataset. These approaches aim to alleviate bias, fostering a more equitable learning experience for the model from both classes [49].

4. Experimental Setup and Results

In our study, we used Neural Network, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU) and GRU-CNN algorithms together with word embedding and fastText's pre-trained word embedding. These models were used for binary and multi-class classifications. While the softmax function was used in the output layer for multi-class classifications, the sigmoid function was used in the output layer for binary classifications. For all models, five-fold cross validation was used for the training and testing processes.

4.1. Simple Neural Network Model

Figure 8 shows our simple Neural Network model for binary classification and Table 5 contains its maximum training and testing accuracies.

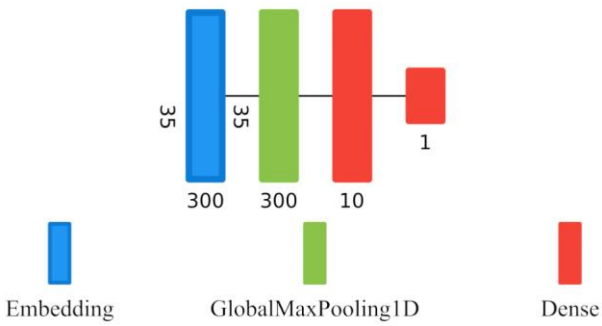

Figure 8. Binary Neural Network model [50].

Table 5. Maximum training and testing accuracies for the Neural Network model.

Classification	Embedding	Train/Test	Max Accuracy (%)
Binary	Word Embedding	Train Test	100.00 80.25
	Pre-trained word embedding	Train Test	100.00 79.32
Multi-class	Word Embedding	Train Test	99.57 63.85
	Pre-trained word embedding	Train Test	99.51 65.23

4.2. Convolutional Neural Network (CNN) Model

We designed a binary CNN model, as shown in Figure 9, and the model's maximum training and testing accuracies are presented in Table 6.

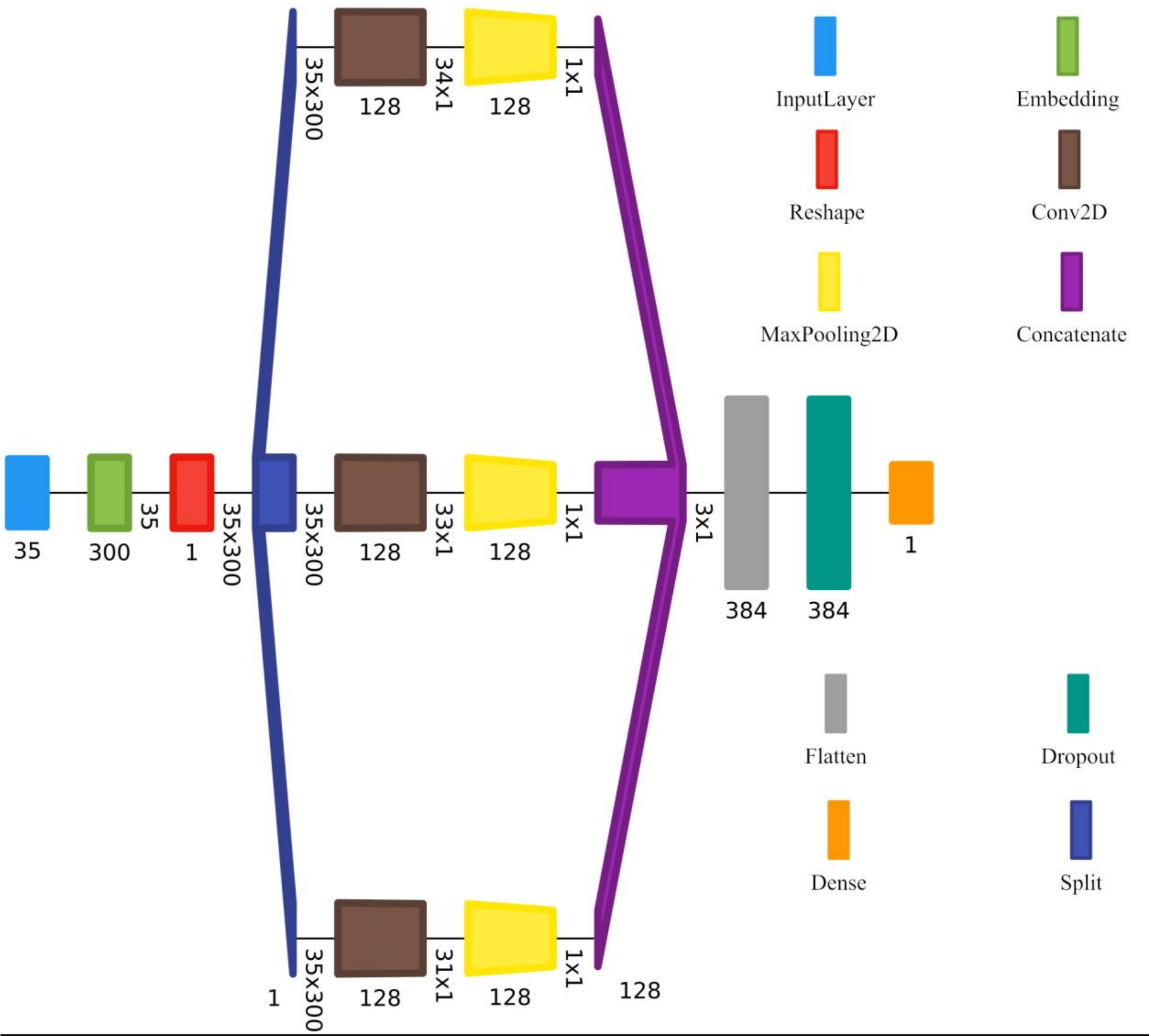

Figure 9. Multi-channel binary CNN model [50].

Table 6. Maximum training and testing accuracies for CNN model.

Classification	Embedding	Train/Test	Max Accuracy (%)
Binary	Word Embedding	Train Test	100.00 77.23
	Pre-trained word embedding	Train Test	100.00 83.02
Multi-class	Word Embedding	Train Test	99.73 63.71
	Pre-trained word embedding	Train Test	99.73 72.72

4.3. Long Short-Term Memory (LSTM) Model

Figure 10 shows the LSTM model designed by us, and its maximum training and testing accuracies are revealed in Table 7.

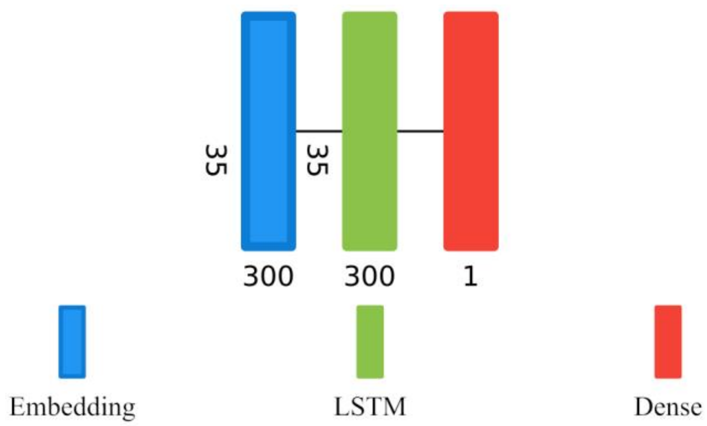

Figure 10. Binary LSTM model [50].

Table 7. Maximum training and testing accuracies for LSTM Model.

Classification	Embedding	Train/Test	Max Accuracy (%)
Binary	Word Embedding	Train Test	98.30 74.69
	Pre-trained word embedding	Train Test	100.00 79.32
Multi-class	Word Embedding	Train Test	99.24 59.52
	Pre-trained word embedding	Train Test	99.24 61.69

4.4. Gated Recurrent Units (GRU) Model

Figure 11 shows the bidirectional GRU model designed by us, and its maximum training and testing accuracies are indicated in Table 8.

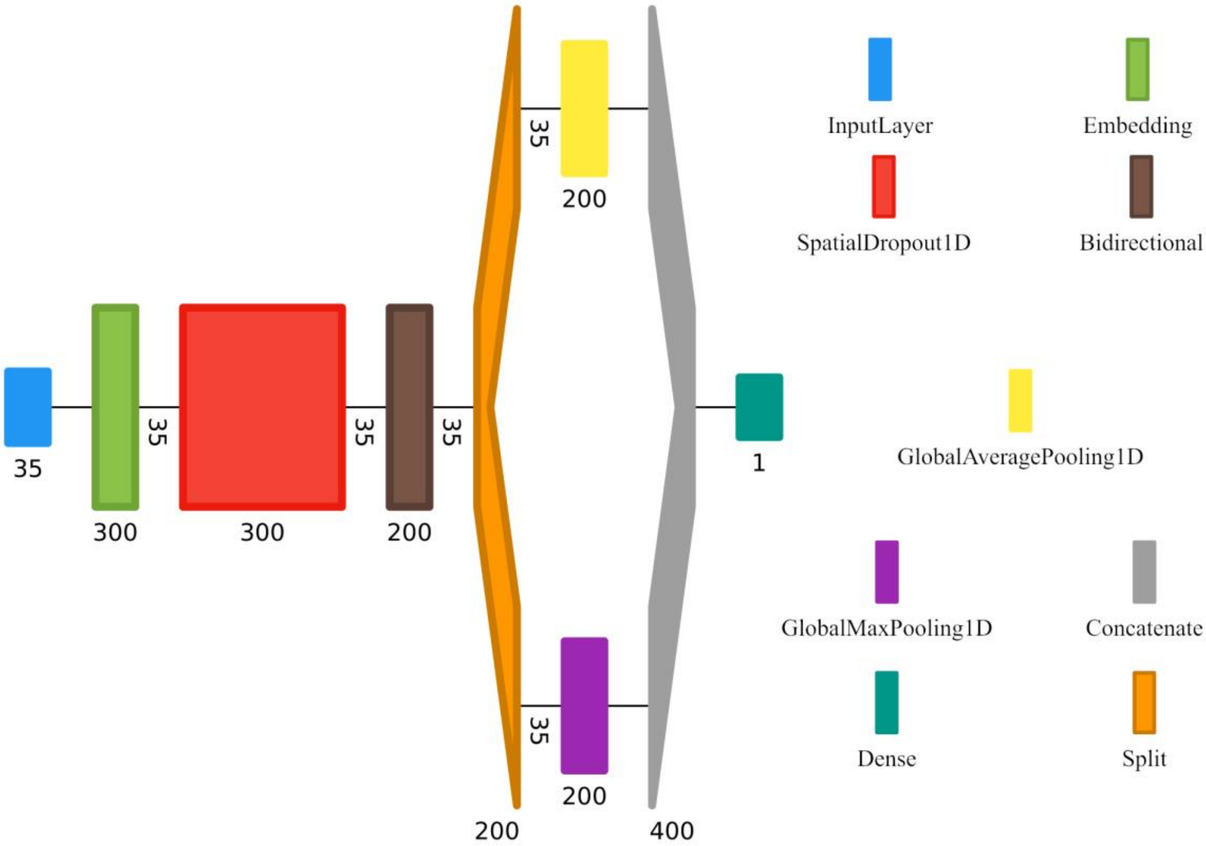

Figure 11. Bidirectional binary GRU model [50].

Table 8. Maximum training and testing accuracies for GRU model.

Classification	Embedding	Train/Test	Max Accuracy (%)
Binary	Word Embedding	Train Test	100.00 80.25
	Pre-trained word embedding	Train Test	100.00 80.31
Multi-class	Word Embedding	Train Test	99.62 62.99
	Pre-trained word embedding	Train Test	99.68 64.07

4.5. GRU-CNN Model

Lastly, we designed a model that had bidirectional GRU and CNN modules, as shown in Figure 12, and the model's maximum training and testing accuracies are presented in Table 9.

4.6. Supporting Findings and Conclusions

This article presents a thorough examination of sentiment analysis applied to Turkish financial tweets, utilizing diverse machine learning algorithms, namely Neural Network, CNN, LSTM, GRU and GRU-CNN, with a specific emphasis on word embedding and pre-trained word embedding techniques. The primary objective of this study is to categorize tweets into positive, negative and neutral sentiments, and to explore the potential applications of sentiment analysis within the domain of stock market decision-making.

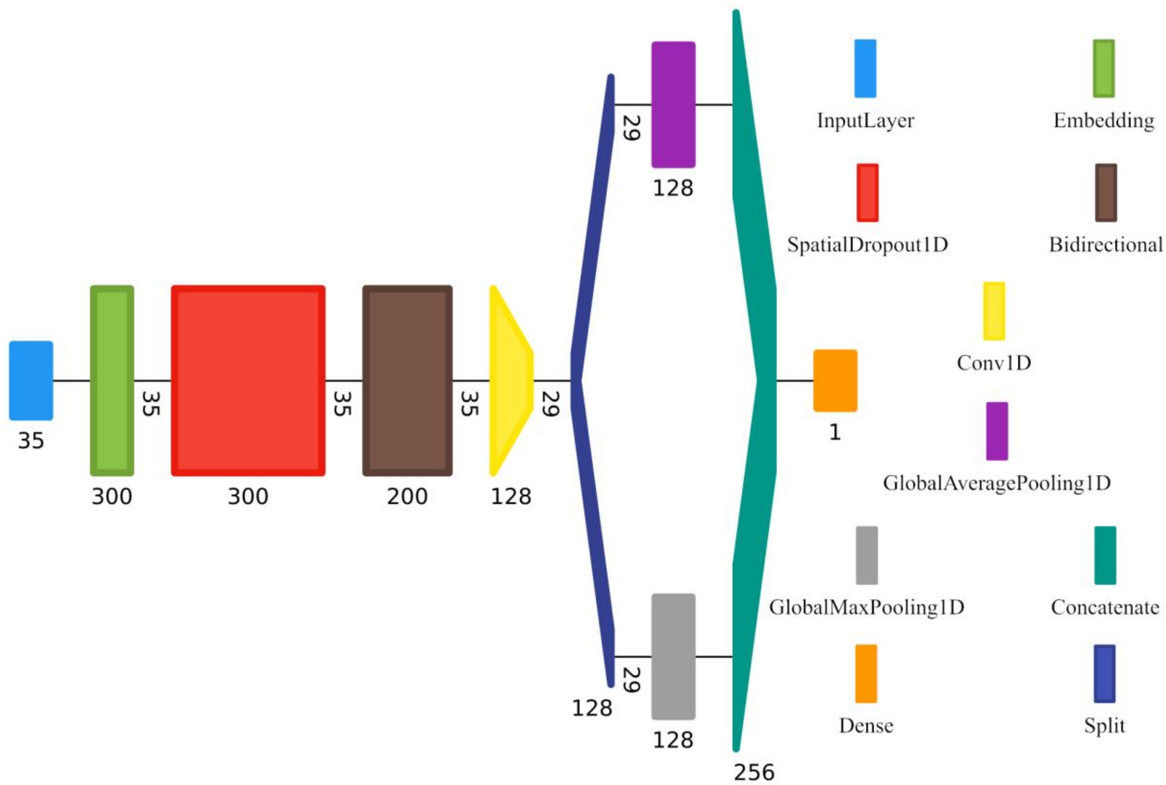

Figure 12. GRU-CNN model [50].

Table 9. Maximum training and testing accuracies for GRU-CNN model.

Classification	Embedding	Train/Test	Max Accuracy (%)
Binary	Word Embedding	Train Test	100.00 80.56
	Pre-trained word embedding	Train Test	100.00 80.25
Multi-class	Word Embedding	Train Test	99.68 61.47
	Pre-trained word embedding	Train Test	99.68 64.29

4.6.1. Creation and Processing of Datasets

This article provides a clear delineation of the methodology involved in gathering Turkish financial tweets, the manual tagging of sentiments and the creation of both binary and multi-class datasets. The pre-processing phase of tweets is meticulously detailed, encompassing actions such as removing superfluous elements, transforming text to lowercase and rectifying spelling errors through the utilization of the ITU Turkish NLP Web Service API.

4.6.2. Feature Extraction Techniques and Algorithm Deployment

This article delves into the utilization of word embedding and pre-trained word embedding, specifically fastText, as mechanisms for extracting features to represent tweets. This study employs a suite of five diverse machine learning algorithms for sentiment classification, with explicit configurations outlined for each algorithm.

4.6.3. Experiment Design and Outcome

This study furnishes an elaborate description of the experimental setup, incorporating the application of five-fold cross-validation during both training and testing phases. Outcomes for both binary and multi-class classifications are presented, accompanied by a comparative analysis of model performance based on different embedding techniques.

4.6.4. Acknowledgment of Limitations and Biases

This article conscientiously acknowledges various limitations and potential biases inherent in the research. These encompass the relatively modest dataset size, challenges related to the nuances of Turkish tweets, biases introduced during data collection and labeling, and potential biases associated with pre-trained word embeddings.

4.7. Pros and Cons of Employed Methods

4.7.1. Advantages

(1) This study embraces a diverse array of machine learning algorithms, providing a comprehensive evaluation of their efficacy in sentiment classification.
(2) Emphasis is placed on the advantageous use of pre-trained word embeddings, particularly fastText, for enhancing model performance.
(3) Valuable insights are offered into the practical applications of sentiment analysis within the Turkish stock market, holding potential significance for decision-makers.

4.7.2. Disadvantages

(1) This article candidly acknowledges drawbacks related to dataset size, biases in data collection and the subjective nature of manual sentiment labeling.
(2) This study falls short in providing an in-depth exploration of model interpretability and the underlying reasons for sentiments observed in financial tweets.

5. Constraints

Dataset scale: The size of the Turkish financial tweets' dataset is comparatively modest, which may constrain the applicability and reliability of the developed models.

Pre-processing complexity: Recognizing the intricacies in handling Turkish tweets, the authors concede the challenges arising from ambiguity and informal language during pre-processing. This may result in potential inaccuracies or biases in sentiment classification.

Binary versus multi-class classification: The discernible performance difference between binary and multi-class classifications underlines the complexities in effectively capturing more refined sentiment categories.

Domain specificity: Given that the models are specifically trained on financial tweets, there is a possibility that their effectiveness might not extend seamlessly to other domains or diverse sentiment analysis tasks.

Potential Biases

Data collection bias: Employing specific keywords for tweet collection may introduce selection bias, potentially skewing the representation of certain sentiment groups by either overemphasizing or underemphasizing them.

Labeling bias: The subjective nature of manual sentiment labeling makes it susceptible to individual biases, influencing the accuracy and reliability of sentiment categorization.

Model bias: The selection of algorithms and hyperparameters holds the potential to impact model performance, introducing biases that may affect the interpretation of sentiment analysis results.

Pre-trained word embedding bias: The biases inherent in the training data of pre-trained embeddings could be mirrored in sentiment analysis outcomes, potentially amplifying and perpetuating biases present in the initial word embedding data.

Although this research offers valuable perspectives on the sentiment analysis of Turkish financial tweets, both researchers and readers must remain cognizant of these limitations and biases. This awareness is crucial for the accurate interpretation and contextualization of this study's findings.

6. Conclusions

Sentiment analysis research has been conducted extensively on social media data in the English language. However, a limited amount of sentiment analysis research has been conducted on social media data in the Turkish language. We created our datasets using Turkish financial tweets, and we tried five different machine learning algorithms (Neural Network, CNN, LSTM, GRU and GRU-CNN) to find sentiments on those datasets together with word embedding and pre-trained word embedding. The binary classification results were better than the multi-class classification results, as shown in Table 10.

Table 10. Comparisons of model accuracies.

Model	Max. Training Accuracy (%)	Max. Testing Accuracy (%)	Average of Max Testing Accuracies of All Folds (%)
Binary NN model with word embedding	100.00	80.25	76.68
Binary NN model with pre-trained word embedding	100.00	79.32	76.19
Multiclass NN model with word embedding	99.57	63.85	61.59
Multiclass NN model with pre-trained word embedding	99.51	65.23	61.59
Binary CNN model with word embedding	100.00	77.23	75.82
Binary CNN model with pre-trained word embedding	100.00	83.02	78.35
Multiclass CNN model with word embedding	99.73	63.71	61.98
Multiclass CNN model with pre-trained word embedding	99.73	72.73	65.05
Binary LSTM model with word embedding	98.30	74.69	72.36
Binary LSTM model with pre-trained word embedding	100.00	79.32	75.14
Multiclass LSTM model with word embedding	99.24	59.52	58.34
Multiclass LSTM model with pre-trained word embedding	99.24	61.69	58.69
Binary GRU model with word embedding	100.00	80.25	76.93
Binary GRU model with pre-trained word embedding	100.00	80.31	77.60
Multiclass GRU model with word embedding	99.73	62.99	60.94
Multiclass GRU model with pre-trained word embedding	99.68	64.07	62.67
Binary GRU-CNN model with word embedding	100.00	80.56	76.44
Binary GRU-CNN model with pre-trained word embedding	100.00	80.24	78.47
Multiclass GRU-CNN model with word embedding	99.68	61.47	60.08
Multiclass GRU-CNN model with pre-trained word embedding	99.68	64.29	62.33

Our results reveal that, generally, all models perform better when they are run with pre-trained fastText word vectors. Also, binary classification results are better than multi-class classification results, as expected. Surprisingly, the results are close to each other. With pre-trained word embedding, CNN models had the best results of all. When we used word embedding, the GRU-CNN model gave better results for the binary classification and the Neural Network model gave better results for the multi-class classification.

We propose a CNN model with pre-trained word embedding for binary and multi-class classifications. Its maximum testing accuracy was 83.02% and the average of its maximum testing accuracies for all folds was 78.35% for binary classifications. For multi-class classifications, its maximum testing accuracy was 72.73% and the average of its maximum testing accuracies for all folds was 65.05%.

In future works, using additional layers in these models may improve their performances. The use of more specific pre-processing techniques could also improve model performances, as the collected Turkish tweets about the Turkish financial market contain many ambiguous words and phrases that make the pre-processing step difficult. In addition, enlarging the datasets could lead to better results.

Author Contributions: Conceptualization, E.M., H.A. and M.Y.; methodology, E.M., H.A. and M.Y; software, E.M.; validation, E.M., H.A. and M.Y.; formal analysis, E.M.; investigation, E.M.; data curation, E.M.; writing—original draft preparation, E.M.; writing—review and editing, E.M., H.A. and M.Y.; visualization, E.M.; supervision, H.A. and M.Y.; project administration, E.M., H.A. and M.Y.; J.R. and J.M.L.-G. revised and review. All authors have read and agreed to the published version of the manuscript.

Funding: The authors were supported by the Mobility Lab Foundation, a governmental organization of the Provincial Council of Araba and the local council of Vitoria-Gasteiz under the following project grant: "Utilización de drones en la movilidad de mercancías".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All of the Twitter data used were collected with Twitter API. Created data are not accessible but data creation methodology is included in the context of this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Twitter. It's What's Happening. Available online: https://twitter.com/ (accessed on 10 June 2019).
2. Twitter Statistics—Statistic Brain. Available online: https://twitter.com/StatisticBrain (accessed on 20 July 2020).
3. Cambria, E.; Schuller, B.B.; Xia, Y.; Havasi, C. New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intell. Syst.* **2013**, *28*, 15–21. [CrossRef]
4. Liu, B. Sentiment Analysis and Opinion Mining. In *Synthesis Lectures on Human Language Technologies*; Springer: Cham, Switzerland, 2012; Volume 5, pp. 1–167. [CrossRef]
5. Medhat, W.; Hassan, A.; Korashy, H. Sentiment Analysis Algorithms and Applications: A Survey. *Ain Shams Eng. J.* **2014**, *5*, 1093–1113. [CrossRef]
6. Memis, E.; Kaya, H. Association Rule Mining on the BIST100 Stock Exchange. In Proceedings of the 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies, ISMSIT 2019—Proceedings, Ankara, Turkey, 11–13 October 2019; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2019.
7. Eryiğit, G. ITU Turkish NLP Web Service. In Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Gothenburg, Sweden, 26–30 April 2014; Association for Computational Linguistics: Gothenburg, Sweden, 2014.
8. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed Representations of Words and Phrases and Their Compositionality. In *Advances in Neural Information Processing Systems*; 2013. Available online: https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf (accessed on 10 June 2019).
9. Mikolov, T.; Karafiat, M.; Burget, L.; Cernocky, J.; Khudanpur, S. Recurrent Neural Network Based Language Model. In Proceedings of the Interspeech 2010, Chiba, Japan, 26–30 September 2010; pp. 1045–1048.
10. Socher, R.; Perelygin, A.; Wu, J.Y.; Chuang, J.; Manning, C.D.; Ng, A.Y.; Potts, C. Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank. In Proceedings of the EMNLP 2013—2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, Seattle, WA, USA, 18–21 October 2013; pp. 1631–1642.
11. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent Trends in Deep Learning Based Natural Language Processing [Review Article]. *IEEE Comput. Intell. Mag.* **2018**, *13*, 55–75. [CrossRef]
12. Grave, E.; Bojanowski, P.; Gupta, P.; Joulin, A.; Mikolov, T. Learning Word Vectors for 157 Languages. In Proceedings of the LREC 2018, 11th International Conference on Language Resources and Evaluation, Miyazaki, Japan, 7–12 May 2018; pp. 3483–3487.
13. Nasukawa, T. Sentiment Analysis: Capturing Favorability Using Natural Language Processing Definition of Sentiment Expressions. In Proceedings of the 2nd International Conference on Knowledge Capture, Sanibel Island, FL, USA, 23–25 October 2003; pp. 70–77. [CrossRef]
14. Martínez-Cámara, E.; Martín-Valdivia, M.T.; Urena-López, L.A.; Montejo-Ráez, A.R. Sentiment analysis in Twitter. *Nat. Lang. Eng.* **2014**, *20*, 1–28. [CrossRef]
15. Almohaimeed, A.S. Using Tweets Sentiment Analysis to Predict Stock Market Movement. Master's Thesis, Auburn University, Auburn, Alabama, 2017.
16. Şimşek, M.U.; Özdemir, S. Analysis of the Relation between Turkish Twitter Messages and Stock Market Index. In Proceedings of the 2012 6th International Conference on Application of Information and Communication Technologies (AICT), Tbilisi, Georgia, 17–19 October 2012. [CrossRef]
17. Yıldırım, M.; Yüksel, C.A. Sosyal Medya Ile Hisse Senedi Fiyatinin Günlük Hareket Yönü Arasindaki Ilişkinin Incelenmesi: Duygu Analizi Uygulamasi. *Uluslararası İktisadi ve İdari İncelemeler Derg.* **2017**, 33–44. [CrossRef]
18. Ozturk, S.S.; Ciftci, K. A Sentiment Analysis of Twitter Content as a Predictor of Exchange Rate Movements. *Rev. Econ. Anal.* **2014**, *6*, 132–140. [CrossRef]

19. Eliaçik, A.B.; Erdogan, N. Mikro Bloglardaki Finans Toplulukları Için Kullanıcı Ağırlıklandırılmış Duygu Analizi Yöntemi. *Ulus. Yazılım Mühendisliği Sempozyumu* **2015**, 782–793.
20. Akgül, E.S.; Ertano, C.; Diri, B. Twitter Verileri Ile Duygu Analizi Sentiment Analysis with Twitter. *Pamukkale Univ. Muh. Bilim. Derg.* **2016**, *22*, 106–110. [CrossRef]
21. Bollen, J.; Mao, H.; Zeng, X. Twitter Mood Predicts the Stock Market. *J. Comput. Sci.* **2011**, *2*, 1–8. [CrossRef]
22. Velioglu, R.; Yildiz, T.; Yildirim, S. Sentiment Analysis Using Learning Approaches over Emojis for Turkish Tweets. In Proceedings of the UBMK 2018—3rd International Conference on Computer Science and Engineering, Sarajevo, Bosnia and Herzegovina, 20–23 September 2018; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2018; pp. 303–307.
23. Smailović, J.; Grčar, M.; Lavrač, N.; Žnidaršič, M. Stream-Based Active Learning for Sentiment Analysis in the Financial Domain. *Inf. Sci.* **2014**, *285*, 181–203. [CrossRef]
24. Bilgin, M.; Sentürk, I.F. Danışmanlı ve Yarı Danışmanlı Öğrenme Kullanarak Doküman Vektörleri Tabanlı Tweetlerin Duygu Analizi. *J. BAUN Inst. Sci. Technol.* **2019**, *21*, 822–839. [CrossRef]
25. Ayata, D.; Saraclar, M.; Ozgur, A. Makine Öğrenmesi ve Kelime Vektör Temsili Ile Türke Tweet Sentiment Analizi. In Proceedings of the 2017 25th Signal Processing and Communications Applications Conference, SIU 2017, Antalya, Turkey, 15–18 May 2017; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2017.
26. Park, H.W.; Lee, Y. How Are Twitter Activities Related to Top Cryptocurrencies' Performance? Evidence from Social Media Network and Sentiment Analysis. *Drus. Istraz.* **2019**, *28*, 435–460. [CrossRef]
27. Nyakurukwa, K.; Seetharam, Y. Does Online Investor Sentiment Explain Analyst Recommendation Changes? Evidence from an Emerging Market. *Manag. Financ.* **2023**, *49*, 187–204. [CrossRef]
28. Choi, J.J. Popular Personal Financial Advice versus the Professors. *J. Econ. Perspect.* **2022**, *36*, 167–192. [CrossRef]
29. Kraaijeveld, O.; De Smedt, J. The Predictive Power of Public Twitter Sentiment for Forecasting Cryptocurrency Prices. *J. Int. Financ. Mark. Institutions Money* **2020**, *65*, 101188. [CrossRef]
30. Duffy, D.; Durkan, J.; Timoney, K.; Casey, E. *Quarterly Economic Commentary, Winter 2012*; The Economic and Social Research Institute: Dublin, Ireland, 2012.
31. Yang, S.Y.; Mo, S.Y.K.; Liu, A. Twitter Financial Community Sentiment and Its Predictive Relationship to Stock Market Movement. *Quant. Financ.* **2015**, *15*, 1637–1656. [CrossRef]
32. Varanasi, R.A.; Hanrahan, B.V.; Wahid, S.; Carroll, J.M. TweetSight: Enhancing Financial Analysts' Social Media Use. In Proceedings of the 8th International Conference on Social Media & Society, Toronto, ON, Canada, 28–30 July 2017; pp. 1–10.
33. Özgür, A.; Özgür, L.; Güngör, T. Text Categorization with Class-Based and Corpus-Based Keyword Selection. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin/Heidelberg, Germany, 2005; Volume 3733, pp. 606–615.
34. Nabi, J. *Machine Learning—Text Processing*; Towards Data Science: Toronto, ON, Canada, 2018.
35. Harris, Z.S. Distributional Structure. *Distrib. Struct. WORD* **1954**, *10*, 146–162. [CrossRef]
36. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. In Proceedings of the 1st International Conference on Learning Representations, ICLR 2013—Workshop Track Proceedings, Scottsdale, AZ, USA, 2–4 May 2013.
37. Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global Vectors for Word Representation. In Proceedings of the EMNLP 2014—2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
38. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching Word Vectors with Subword Information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
39. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL HLT 2019—Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 4171–4186.
40. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional Architecture for Fast Feature Embedding. In Proceedings of the MM 2014—Proceedings of the 2014 ACM Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; Association for Computing Machinery, Inc.: New York, NY, USA, 2014; pp. 675–678.
41. Razavian, A.S.; Azizpour, H.; Sullivan, J.; Carlsson, S. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 23–28 June 2014; IEEE Computer Society: Washington, DC, USA, 2014; pp. 512–519.
42. Collobert, R.; Weston, J. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In Proceedings of the Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 5–9 July 2008; pp. 160–167.
43. Elvis Deep Learning for NLP: An Overview of Recent Trends. Available online: https://medium.com/dair-ai/deep-learning-for-nlp-an-overview-of-recent-trends-d0d8f40a776d (accessed on 24 August 2018).
44. Elman, J.L. Finding Structure in Time. *Cogn. Sci.* **1990**, *14*, 179–211. [CrossRef]
45. Girosi, F.; Jones, M.; Poggio, T. Regularization Theory and Neural Networks Architectures. *Neural Comput.* **1995**, *7*, 219–269. [CrossRef]

46. Park, S.; Kwak, N. Analysis on the Dropout Effect in Convolutional Neural Networks. In Proceedings of the Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; Revised Selected Papers, Part II 13;. Springer: Cham, Switzerland, 2017; pp. 189–204.
47. Wang, L.; Han, M.; Li, X.; Zhang, N.; Cheng, H. Review of Classification Methods on Unbalanced Data Sets. *IEEE Access* **2021**, *9*, 64606–64628. [CrossRef]
48. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
49. Yap, B.W.; Rani, K.A.; Rahman, H.A.A.; Fong, S.; Khairudin, Z.; Abdullah, N.N. An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets. In Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013), Kuala Lumpur, Malaysia, 16–18 December 2013; Springer: Singapore, 2014; pp. 13–22.
50. Bäuerle, A.; Van Onzenoodt, C.; Ropinski, T. Net2Vis: Transforming Deep Convolutional Networks into Publication-Ready Visualizations. *arXiv* **2019**, arXiv:1902.04394.