Tabriz Journal of Electrical Engineering (TJEE), vol. 55, no. 4, 2025 Serial no. 114
DOI: 10.22034/tjee.2025.64303.4917

Tehran Stock Exchange Price Movement Prediction using Daily News with Hierarchical Attention Network Plus BERT

L. Hafezi, M. R Pajoohan*, S.Zarifzadeh

Computer Engineering Department, Yazd University, Yazd, Iran
leilahafezi@stu.yazd.ac.ir, pajoohan@yazd.ac.ir, szarifzadeh@yazd.ac.ir
*Corresponding author

Received: 03/11/2024, Revised: 21/02/2025, Accepted; 21/04/2025.

Abstract
The stock market's significance in the global economy necessitates demands more accurate prediction methods. This paper introduces a novel hierarchical attention mechanism aimed at enhancing the performance of predicting stock price movements. Hierarchical attention networks assume that not all news segments hold equal relevance in forecasting stock market trends. Furthermore, we assert that not all daily news carries an equivalent significance in predicting market trends. To tackle this challenge, we suggest a hierarchical attention network plus BERT that emulates the news hierarchy and assigns weights to news items based on their significance, and also the most informative news articles in each trading day in stock market prediction. Our HAN+BERT method incorporates three levels of attention mechanisms, operating at the word, sentence, and news level. This allows the model to identify the most significant news stories of the day and select the most informative sentences and words within these articles. Using BERT as the word embedding approach has resulted in better performance for our stock trend prediction model. Empirical results on Persian financial news and three stock market indices reveal the effectiveness of our HAN+BERT model, with a peak accuracy of 65.49%, which is 3% higher than the best baseline model.

Keywords
Deep learning models, Stock market prediction, Attention mechanism, Hierarchical attention network, BERT embedding.

1. Introduction

Forecasting stock market trends has a rich history in financial economics, characterized by diverse methods and features. Historically, studies relied on historical data to predict market movements. However, unexpected events often disrupt the accuracy of these models. Recently, there has been an increased focus on using financial news to forecast price movements, as news articles can capture unforeseen events to some extent [1]. Daily news plays a crucial role for investors by providing information about organizations, corporations, and societal events, thus aiding in the formulation of trading strategies within the stock market. However, experts in Iran argue that the Tehran Stock Exchange (TSE) does not adhere to the standards of developed global stock exchanges due to government intervention and the opacity of state-owned companies [2], [3]. They claim that the TSE is less responsive to news events, for instance, the TSE maintained an upward trend despite the negative impact of the coronavirus on global markets (as illustrated in Fig. 1). This study investigates the potential for predicting stock trends in Iran by utilizing Persian text extracted from financial news articles.

In recent times, deep learning has achieved remarkable success across numerous applications and predictive tasks, notably excelling in both financial time series forecasting and natural language processing (NLP). Deep learning models, especially those integrating attention mechanisms, have set new benchmarks in NLP [4]. For instance, Yang et al. introduced a hierarchical attention network (HAN) for document classification, outperforming traditional models [5]. This hierarchical approach reflects the structure of documents, where sentences are consisting of words, and documents are consisting of sentences. The HAN approach incorporates dual attention mechanisms, applied both at the word level and the sentence level, allowing it to dynamically prioritize and extract the most relevant content within a document. It is crucial to recognize that within a news article, different words and sentences carry varying levels of informativeness, and their importance can change based on the specific context. Thus, the HAN model's dual-level attention mechanism is pivotal in assigning appropriate attention to the most significant words and sentences in a news article.

While several studies have employed this approach and noted its effectiveness on stock market prediction [6-9], a drawback of this method is that it treats all the news articles equally among the vast volume of daily publications, without effectively identifying and selecting the most vital and influential news items.

Given the vast volume of financial news published daily, it is evident that not all news articles are of equal importance and impact on stock values. As a result, we

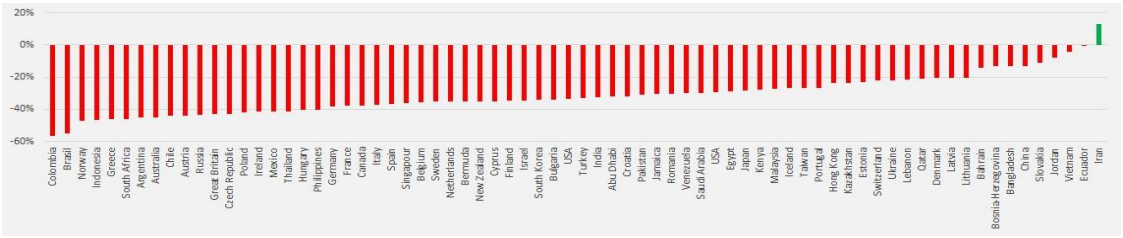

Fig. 1. Stock market index by countries after covid-19 epidemic, Iran is different.

derive inspiration from HAN model and put forth a novel model for stock market trend prediction. Consequently, we incorporate an additional level into this model to specifically identify more significant news articles, and we refer to this new model as "HAN+".

Another limitation of the HAN model is its reliance on Glove embeddings, which struggle to encode unknown or out-of-vocabulary words effectively. Additionally, Glove embeddings may fall short when handling large volumes of text data because the same word can convey various meanings depending on the context of each sentence. Thus, we employ BERT instead of Glove2Vec word embeddings in HAN+ model. Unlike the Glove2Vec model, which generates context-free word embeddings, BERT provides contextual understanding by employing bidirectional transformers. This allows BERT to thoroughly analyze entire sentences and accurately interpret their context [10].

The limitations of traditional methods for stock market prediction, led us to adopt HAN+BERT. Recurrent Neural Networks and single-layer LSTMs treat text as a flat sequence, disregarding the hierarchical structure of financial news. These models process all words with equal importance, despite the fact that some words and sentences carry more significance in predicting market movements. HAN+BERT overcomes this limitation by leveraging a hierarchical attention mechanism, first capturing word-level dependencies within sentences, then sentence-level relationships within a news article, and ultimately identifying the most critical financial news on a given day. Additionally, classical models struggle with long texts, whereas HAN+BERT utilizes attention layers and a hierarchical framework for better processing of lengthy news articles.

Modern Transformer-based models have shown strong performance in financial text processing by capturing long-range dependencies between words. However, these models struggle with sequence length limitations, requiring techniques like Sliding Window or specialized architectures to handle longer texts. Our method, on the other hand, can process long news articles naturally without additional modifications. Moreover, Transformer models do not explicitly account for the hierarchical structure of news articles, as they only model dependencies at the sentence level. In contrast, our method follows a hierarchical approach, identifying key words within sentences, crucial sentences within a news article, and ultimately the most influential financial news of the day. Another limitation of Transformer-based models is their computational complexity and resource-intensive nature. We evaluated several Transformer-based models on our datasets under the same conditions as HAN+BERT model, and they exhibited significantly higher complexity in terms of both time and memory consumption.

Thus, HAN+BERT provides a more interpretable, efficient, and context-aware approach for stock price prediction compared to traditional and modern deep learning models.

This paper presents several novel contributions to stock market prediction, including:

- Leveraging hierarchical attention not only to identify key words and sentences within a news article but also to determine the most impactful news pieces from daily financial reports. This multi-level selection process enhances stock price prediction performance by ensuring that the model focuses on the most market-moving information.
- Utilizing BERT for word embeddings, enabling the model to capture contextual meaning by analyzing entire sentences.
- Developing a publicly available Persian financial news dataset to facilitate future research.
- Outperforms baseline models, achieving an improvement in accuracy.

The following sections of this paper are organized as follows: Section 2 reviews related work, Section 3 presents the proposed methodology, and Section 4 discusses the application of the method on a Persian financial news dataset, including a comparison with several other methods. Finally, Section 5 provides the conclusion.

2. Related Works

Financial news holds crucial information for stock market prediction. While various techniques exist for financial forecasting, the use of attention mechanisms in this context remains underexplored. This section reviews prior research on stock prediction using financial news, with a focus on attention mechanisms. Notably, there is no published study in leading journals and conferences that has been carried out on Persian financial news using deep learning techniques. Existing research falls into two main categories: sentiment analysis (also known as opinion mining [11]) of financial news, which examines the impact of optimistic or pessimistic news on stock movements [12], and direct analysis of financial news content to assess its influence on stock prices. Overall, financial news features have shown promise in enhancing prediction accuracy.

Liu et al. [8] proposed a Multi-element Hierarchical Attention Capsule Network (MHACN) to forecast U.S.

stock movements by integrating financial news and tweets. They implemented a multi-element hierarchical attention mechanism to assess event impacts and utilized a capsule network for enhanced contextual understanding. The researchers gathered news data from Yahoo Finance and analyzed 47 stock price datasets from the S&P 500, achieving an accuracy of 60.56% with the MHACN model.

Chen et al. [13] introduced a hierarchical attention network for stock prediction, called NMNL, designed to integrate comprehensive news and market information for improved accuracy. NMNL consists of three modules: (1) News Encoder, which creates multi-view news representations using attention mechanisms across headlines, bodies, and sentiment; (2) Market Information Encoder, which captures daily news impacts on stocks through self-attention and combines this with technical indicators; and (3) Temporal Auxiliary Module, which processes multi-day market information via a BiLSTM with attention for temporal context. Using 1,035,269 news articles on A-share companies, the model achieved a Directional Accuracy (DA) of 0.613 and MCC of 0.1165 for CSI 100, and DA of 0.608 and MCC of 0.1072 for HS 300.

Zhang et al. [14] proposed a CNN-BiLSTM-Attention model to enhance the accuracy of stock price and index predictions. This approach combines CNN for capturing nonlinear local features, BiLSTM for extracting bidirectional temporal features, and an attention mechanism to prioritize relevant data while reducing redundancy. Tested on twelve stock indices (four Chinese, eight international), the model outperformed other methods, improving key metrics such as RMSE, MAPE, and R^2. For the FTSE index, RMSE and MAPE decreased by 10.7% and 0.9, with R^2 rising from 0.972 to 0.978. Similarly, JKSE index predictions saw RMSE and MAPE reductions of 16.3% and 0.7, with R^2 increasing from 0.980 to 0.986.

Zhang et al. [15] introduced the Transformer Encoder-based Attention Network (TEANet) to improve stock movement prediction by addressing temporal dependency issues in financial data and integrating diverse data sources like historical prices and tweets. TEANet employs a feature engineering framework tailored to small datasets, utilizing the transformer model and multiple attention mechanisms to capture essential dependencies and extract deep features. Experimental results show TEANet's strong performance, with accuracy and Matthews Correlation Coefficient (MCC) improvements of over 20% and 104%, respectively, compared to other models, even when trained on just five days of data.

Zhang et al. [16] proposed the FA-CNN model, a convolutional neural network integrating a deep factorization machine and attention mechanism, aimed at enhancing stock price movement prediction accuracy. Unlike models focusing mainly on temporal features, FA-CNN captures intraday interactions and incorporates sub-industry index data to account for co-movement between industry indices and individual stocks. Testing across three industries showed FA-CNN achieving 64.81% accuracy, surpassing traditional LSTM models by 7.38% and models without sub-industry data by 3.71%.

Bi et al. [9] introduced a Hierarchical Attention Network (HANet) for improving long-term multivariate time series forecasting by reducing the influence of irrelevant factors and aligning exogenous variables with target outcomes. The model incorporates a Factor-aware Attention Network (FAN) to assign minimal weight to extraneous factors and a multi-modal fusion network (MFN) to selectively integrate target and exogenous information. This combination enhances the model's adaptability and prediction accuracy. Tested on Beijing PM 2.5 and Chlorophyll datasets, HANet excelled in long-term multivariate time series forecasting.

Li and Xu [17] proposed an approach to improve stock price prediction by combining generative adversarial networks (GANs) and transformer-based attention mechanisms. GANs are used to generate synthetic stock price data while incorporating market sentiment and volatility. Attention mechanisms focus on key features and patterns in the data, enhancing the identification of critical market indicators. They have integrated data from Seeking Alpha, a reputable platform providing daily news and analysis, to capture market sentiment and news events impacting Apple Inc.'s stock price. Experimental evaluations on real-world stock market data reveal that standard RMSE values are 2.414 for the training set and 3.331 for the testing set.

Liu et al. [18] introduced SA-TrellisNet, a news-driven stock market index prediction model that integrates TrellisNet with an attention mechanism. The model leverages CNN and LSTM for sentiment analysis, extracting sentiment indices from large-scale financial news. An attention mechanism fuses stock data with sentiment indices for prediction using TrellisNet. Evaluated on seven major stock indices, SA-TrellisNet outperforms competing models, significantly reducing MSE, MAPE, and MAE for the SSE index and improving RMSE and R^2 for others. However, it underperforms on the NYSE index due to insufficient sentiment data. Table I presents a summary of the methods reviewed in previous studies.

Table I. Comparison of Review Papers.

Paper	Metric	Datasets	Value
[8]	Accuracy	S&P 500	60.56
[9]	RMSE	PM 2.5	45.93
		Chlorophy11	78.13
[13]	Accuracy	CSI 100	61.30
[14]	R^2	FTSE	0.978
		JKSE	0.986
[15]	Accuracy	stocknet	65.16
[16]	Accuracy	HCSI	61.90
		FBSI	64.81
		RESI	59.38
[17]	RMSE	Apple	3.331
[18]	R^2	S&P500	0.979
		NYSE	0.963
		DJI	0.995
		NASDAQ	0.993
		FTSE	0.998
		N225	0.981
		SSE	0.998

In light of the existing literature, our work is motivated by two key factors. First, there is a notable gap in research specifically focused on Persian datasets, an area that remains underexplored. Second, the method we propose is designed to offer superior accuracy and enhanced performance, addressing limitations in current approaches and promising more reliable results.

3. Our Methodology

In this section, we initially introduce the concepts and terminology employed in this study, followed by a detailed description of the specific problem that our investigation aims to address. Then, we present the details of HAN+ model for stock market prediction.

3.1. Terminology and Problem Statement

Accurately predicting stock market trends can result in substantial revenue generation. This research aims to assess the predictability of the Tehran Stock Exchange (TSE) and introduce an innovative model that outperforms conventional models. To accomplish this goal, we will utilize HAN+ model, which employs an attention mechanism, to enhance the accuracy of predicting the direction of stock trends, whether they are upward (bull market) or downward (bear market). For this purpose, we have incorporated daily financial news as input for this problem.

Assuming we have u trading days, and on each day, we choose d daily news articles, we can define a set of news as follows: $E_{ti} = \{e_{ti} | t \in \{1..u\}, i \in \{1..d\}\}$. Each news articles e_{ti} comprises m sentences, $e_{ti} = (s_1, s_2, \dots, s_m)$, and each sentence comprises n words, $s_j = (w_1, w_2, \dots, w_n)$. We frame stock movement prediction as a binary classification task. Given the stock s on the trading day t, determining whether it rises or falls from $t-1$ to t is defined as follows:

$$Y_t = \begin{cases} 0 \; if \; close_t < close_{t-1} \\ 1 \; if \; close_t \geq close_{t-1} \end{cases} \quad (1)$$

where $close_t$ represents the closing price of stock s on trading day t. When the closing price of Y_t on trading day t exceeds or equals that of the previous trading day $t-1$, $Y_t = 1$, signifying an increase in the stock price s; on the other hand, if the closing price is lower than the previous trading day, then $Y_t = 0$, denoting a decline in the stock price.

The Tehran Stock Exchange has 65 industries; we selected three major ones and a leading company within each. We collected news on these companies, their industries, and the broader market to predict daily price movements for individual stocks, industry indexes, and the overall market index.

3.2. Hierarchical Attention Network Plus (*HAN+*)

We design Hierarchical Attention Network Plus (HAN+) to sift through the vast amount of financial and stock market news to select the most influential ones. It follows a hierarchical structure by first selecting the most influential news articles, then the most influential sentences within each news article, and finally the most influential words within each sentence. This process helps determine whether the overall market sentiment is positive or negative, aiding next-day stock predictions.

The model's architecture comprises three layers. At each layer, the inputs are encoded into one-dimensional vectors through a sequence encoder. Following this, an attention mechanism is utilized to give greater emphasis to more informative inputs, which are then forwarded to the subsequent layer. Ultimately, a Sigmoid function is applied to make a final determination about the output, predicting whether the stock market will be positive or negative the next trading day.

The architectural diagram of HAN+ model is depicted in the Fig. 2, and the subsequent sections will offer a more comprehensive elucidation of this approach.

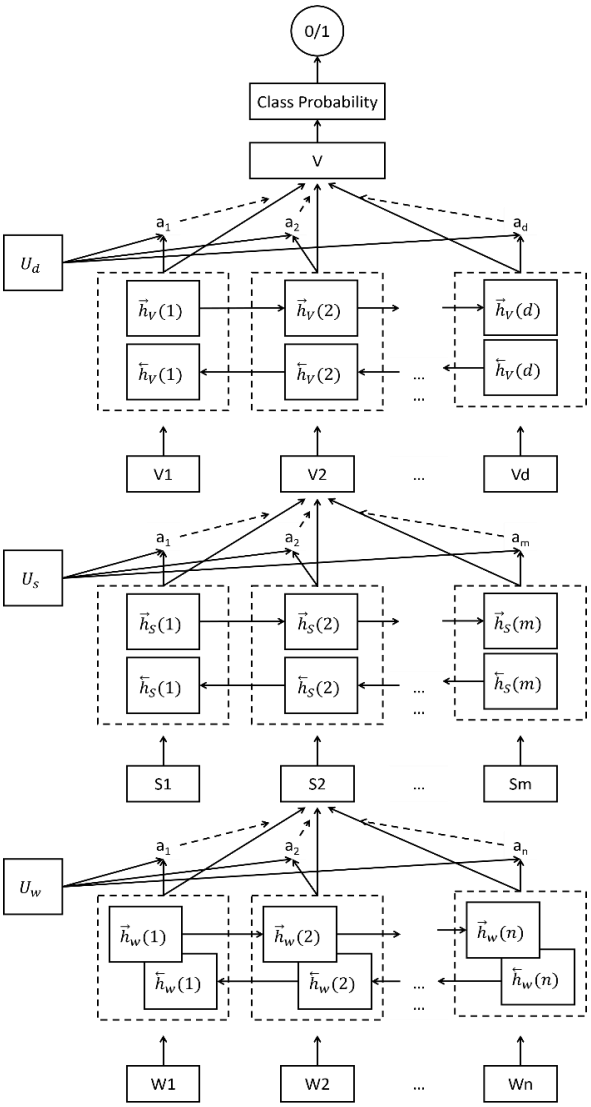

Fig. 2. Hierarchical Attention Network Plus (HAN+) Architecture.

HAN+ model is composed of multiple components: a word sequence encoder, a word-level attention layer, a sentence encoder, a sentence-level attention layer, a news sequence encoder and a news-level attention layer. Subsequently, we provide a detailed description of these components. The flowchart depicted in Fig. 3 illustrates the step-by-step process of the HAN+BERT model for

stock index prediction. Each element in the flowchart signifies a fundamental step in the model's functioning.

3.2.1 Data Preprocessing

The collected financial news articles underwent several preprocessing steps to ensure the data's quality and suitability for analysis. The preprocessing steps were as follows:

1. Data Cleaning: First, we removed any articles that were purely images or those with a length of fewer than 500 characters, as they were deemed irrelevant for meaningful analysis.
2. Removing Special Characters and Numbers: Next, all special characters, punctuation marks, and numbers were removed from the text, as they did not contribute to the semantic analysis of the news content.
3. Tokenization: After cleaning the text, we performed tokenization, which involved splitting the text into individual words (tokens) to prepare it for further processing. we employ Hazm, a Python NLP toolkit, for tokenizing our Persian texts.

These preprocessing steps helped to refine the dataset, eliminating noise and ensuring that the remaining text was ready for feature extraction and modeling.

3.2.2 BERT Word Embedding

Word embedding transforms words into *n*-dimensional vectors, placing semantically similar or related words closer together based on patterns found in the training data. GloVe is commonly used for word embedding, which rely on unsupervised learning methods.

The GloVe model creates denser and more expressive vector representations by training on a collective global word-to-word co-occurrence matrix derived from a collection of text documents. However, a primary drawback of GloVe embeddings is their inability to encode unknown or out-of-vocabulary words. To address this issue, advanced models like BERT, which stands for Bidirectional Encoder Representations from Transformers have been developed [19], introduced by Google AI researchers in 2018, BERT marked a significant advancement in NLP. BERT captures contextual information from vast amounts of text data, as a pre-trained language model. Unlike the context-free embeddings created by models like Glove, BERT's bidirectional transformers provide contextualized word embeddings by analyzing entire sentences and understanding their context. Thus, in our HAN+BERT method, we employ BERT as an alternative to Glove2Vec word embedding.

We utilize ParsBERT, a Persian language understanding model based on Google's BERT architecture. ParsBERT is pre-trained on extensive Persian corpora encompassing various subjects and writing styles, including scientific texts, novels, and news articles. This model includes over 3.9 million documents, 73 million sentences, and 1.3 billion words [20].

3.2.2 GRU-based Sequence Encoder

The Gated Recurrent Unit (GRU), introduced by Cho et al. in 2014 [21], and further refined by Chung et al. [22], is a gating mechanism designed for efficient processing of sequential data. It stands out for its faster computation compared to many traditional recurrent neural network (RNN) models. The GRU uses update and reset gates to regulate information flow, improving performance and learning. The update gate z_t^j controls how much past information is retained by linearly combining the current input x_t with the previous hidden state h_{t-1} and then applying the Sigmoid function σ to this combination, which insures that z_t^j is between 0 and 1:

$$z_t^j = \sigma(W_z x_t + U_z h_{t-1})^j \quad (2)$$

where x_t is the sequence vector at time t, and W_z and U_z are the weight matrix. The reset gate r_t^j determines how much of the previous hidden state should be disregarded or reset by combining the previous hidden state and the current input linearly and then applying an activation function. The reset gate is computed as follows:

$$r_t^j = \sigma(W_r x_t + U_r h_{t-1})^j \quad (3)$$

The new state h_t^j at time t is a linear interpolation between the previous state h_{t-1}^j and the current new state $\hat{\mathrm{h}}_t^j$:

$$h_t^j = \left(1 - z_t^j\right) h_{t-1}^j + z_t^j \hat{\mathrm{h}}_t^j \quad (4)$$

The candidate state is computed as follows:

$$\hat{\mathrm{h}}_t^j = tanh\left(Wx_t + U(r_t \odot h_{t-1})\right)^j \quad (5)$$

where $\odot$ is an element-wise multiplication. If r_t is zero, the previous state is completely forgotten.

3.2.3 Word Sequence Encoder

Assume that a news article has m sentences and n represents the number of words in each sentence. The j^{th} word in the i^{th} sentence from the k^{th} news article is represented by w_{kij} with $j\epsilon[1,n]$ and $k\epsilon[1,d]$. At the first step, the words are transformed into vectors using an embedding matrix W_e, $x_{kij} = W_e w_{kij}$.

Next, we use a bidirectional GRU, which processes input sequences in both forward and backward directions. In the forward direction, the input sequence is processed from the first word to the last word, with each word's representation influenced by the preceding words in the sequence. Conversely, in the backward direction, the sequence is processed from the last word to the first word, with each word's representation influenced by the subsequent words in the sequence.

$$\begin{aligned} & x_{kij} = W_e w_{kij}, j\epsilon[1,n], i\epsilon[1,m] \; and \; k\epsilon[1,d] \quad (6) \\ & \vec{h}_{kij} \\ & = \overrightarrow{GRU}\left(x_{kij}\right), j\epsilon[1,n], i\epsilon[1,m] \; and \; k\epsilon[1,d] \\ & \overleftarrow{h}_{kij} \\ & = \overleftarrow{GRU}\left(x_{kij}\right), j\epsilon[n,1], i\epsilon[m,1] \; and \; k\epsilon[d,1] \end{aligned}$$

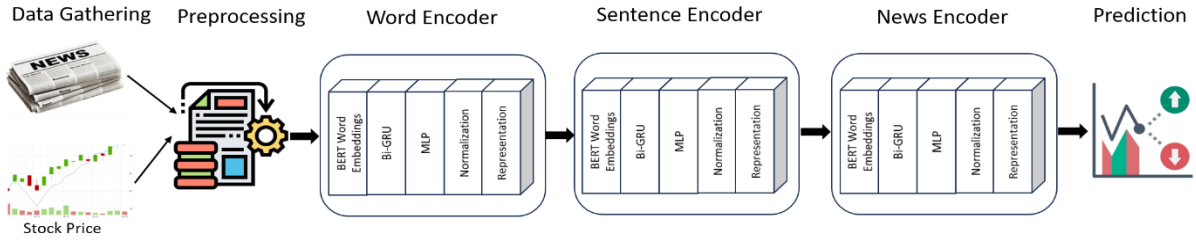

Fig. 3. The step-by-step flowchart of HAN+BERT.

3.2.4 Word Attention

Some words hold greater significance in conveying the meaning of a sentence. To effectively combine the representations of these crucial words into a sentence vector, our model employs an attention mechanism. The vector from the bidirectional GRU, h_{kij} is forwarded through a one-layer MLP to acquire u_{kij}. The importance of each word is then calculated, yielding a normalized importance weight a_{kij} through a Softmax function. The representations of these words are then aggregated to compute a sentence vector s_{ki}, which essentially involves calculating the weighted sum of each vector based on their respective weights. Here is a breakdown of the steps:

One-Layer MLP: The hidden states obtained from the Bi-GRU are then passed through a single-layer Multilayer Perceptron (MLP). This neural network layer performs a linear transformation on each hidden state, followed by a non-linear activation function. The purpose of this MLP is to learn and calculate importance scores for each word j in i^{th} sentence of k^{th} news.

$$u_{kij} = \tanh\left(W_w h_{kij} + b_w\right) \tag{7}$$

where W_w and b_w are weight matrix and bias, respectively. h_{kij} is the hidden state or feature vector of word j in the i^{th} sentence of k^{th} news article.

Importance Measurement: The output of the MLP represents the importance scores for each word within the sentence. These scores indicate how important each word is in the context of the entire sentence.

Normalization (Softmax): To ensure that the importance scores form a probability distribution, they are passed through the Softmax function. This function converts the importance scores into normalized importance weights, where higher scores correspond to higher weights.

$$\alpha_{kij} = \frac{\exp\left(u_{kij}^T u_w\right)}{\sum_p \exp\left(u_{kip}^T u_w\right)} \tag{8}$$

Sentence Representation: The sentence representation is computed as a weighted sum of word vectors using normalized importance weights. The result is a sentence vector that captures the most important information from the words in the sentence, with different words contributing differently based on their importance.

$$s_{ki} = \sum_p \alpha_{kip} h_{kip} \tag{9}$$

This approach enables the model to concentrate on the most relevant words in the sentence when creating a representation for the entire sentence.

3.2.5 Sentence Sequence Encoder

The procedure described for encoding words is similarly applied to the derived sentence vectors to generate the overall document vector. A bidirectional GRU is utilized to encode the sentences:

$$\begin{aligned} \vec{h}_{ki} &= \overrightarrow{GRU}(s_{ki}), i\epsilon[1,m] \; and \; k\epsilon[1,d] \\ \overleftarrow{h}_{ki} &= \overleftarrow{GRU}(s_{ki}), i\epsilon[m,1] \; and \; k\epsilon[d,1] \end{aligned} \tag{10}$$

For getting an annotation of sentence i, $\vec{h}_{ki}$ and $\overleftarrow{h}_{ki}$ must be concatenated, i.e., $h_{ki} = [\vec{h}_{ki}, \overleftarrow{h}_{ki}]$. The vector h_{ki} encapsulates the surrounding sentences' context while maintaining a primary focus on sentence i. In this part of HAN+ model, the GRU-based sequence encoder employed here mirrors that utilized in the word encoder.

3.2.6 Sentence Attention

Each sentence within a news article conveys a distinct semantic meaning; Hence, it is essential to calculate the attention weights for each sentence separately and reward sentences that are more important in stock movement prediction. For computing document vector v_k that outlines all the information of sentences in a news, following formula is used:

$$\begin{aligned} u_{ki} &= \tanh(W_s h_{ki} + b_s) \\ \alpha_{ki} &= \frac{\exp(u_{ki}^T u_s)}{\sum_p \exp\left(u_{kp}^T u_s\right)} \\ v_k &= \sum_p \alpha_{kp} h_p \end{aligned} \tag{11}$$

3.2.7 News Sequence Encoder

Given the news vector v_k, we can get news vectors in each day in a similar way. We assume that each trading day we have d news. We employ a bidirectional GRU for encoding the news of each day:

$$\begin{aligned} \vec{h}_k &= \overrightarrow{GRU}(v_k), k \in [1,d] \\ \overleftarrow{h}_k &= \overleftarrow{GRU}(v_k), k \in [d,1] \end{aligned} \tag{12}$$

We have $h_k = [\vec{h}_k, \overleftarrow{h}_k]$, that h_k summarizes the information of news around k^{th} news article.

3.2.8 News Attention

The attention mechanism aims to implement a weighted combination of all the news in trading day t, assigning the highest weights to the most relevant news. Specially, the attention mechanism in this part aggregates all the news representations into a comprehensive overall representation. This mechanism computed as:

$$\begin{aligned} u_k &= \tanh\left(W_n h_k + b_n\right) \\ \alpha_k &= \frac{\exp(u_k^T u_n)}{\sum_p \exp\left(u_p^T u_n\right)} \end{aligned} \tag{13}$$

$$d_k = \sum_i \alpha_p h_p$$

where d_k is the k^{th} news vector that outlines all the information in the corresponding news article.

3.2.9 Prediction

We use Sigmoid function to predict the probability of binary classification (we only have 2 classes, rise and fall of stock values).

$$\rho = sigmoid(W_p d + b_p) \quad (14)$$

We utilize the cross-entropy loss function for model optimization during training. The cross-entropy takes the predicted probabilities $q(x)$ and measures the distance from the true labels $p(x)$. We can represent cross-entropy as:

$$CE = -\sum_x p(x) \log_2 q(x) \quad (15)$$

4. Experiments

We crawl public financial news and Tehran stock exchange news to create required datasets over the period of March 21, 2017, to September 22, 2022. We conduct our experiments on predicting Tehran market index, industry index and stock index. The historical data we collect allows for an in-depth analysis and modeling of various aspects of financial markets over the specified time period. In this study, we utilize daily news articles from the previous day to predict stock price fluctuations for the next trading day. The model is designed to forecast daily price changes based on historical data from the prior trading day. Below, we provide details of our datasets and models, followed by a comparison of the predicted results.

4.1. Datasets

In recent years, one of the most pivotal developments in Iranian politics and the economy has been the imposition of U.S. sanctions on Iran. We examine the impact of financial news on the Tehran Stock Exchange within three key Iranian industries—metal, petrochemical, and oil—that have been significantly affected by U.S. sanctions. We analyze four years of news data, spanning both pre- and post-sanction periods, capturing market fluctuations during times of prosperity and recession. By focusing on one representative company from each industry, we aim to assess how financial news influences stock performance in the context of economic restrictions imposed on Iran's major sectors. We select one company in each mentioned industry. Fameli (The stock of National Iranian Copper Industries Company) in metal category, Shebandar (The stock of Bandar Abbas oil refinery) in oil products category and Shekabir (The stock of Amir Kabir Petrochemical) in petrochemical category are selected. Fameli and Shekabir are directly sanctioned, but Shebandar is affected by sanctions. We collected the news from March 21, 2017 to September 22, 2022. We extracted financial news from three websites that publish specialized stock market news [1] using Scrapy and Selenium for web scraping. These tools were employed to efficiently crawl and collect relevant articles, enabling us to gather large volumes of data for further analysis. We also collected specialized news about each company from their respective websites or by searching for the names of the companies and industries in economic and financial news. The data set statistics are summarized in Table II, with 80% allocated for training, 10% for validation, and the remaining 10% for testing purposes.

Table II. Datasets Statistics.

Dataset	# News	#Unique Tokens	#Sentences
Fameli	65324	121889	253114
Shebandar	69820	199208	272702
Shekabir	56318	109567	201386

4.2. Settings

Testing various values for model parameters and selecting the optimal settings is an influential step in the model development and fine-tuning process. We experimented with different parameter configurations to determine the settings that yield the best performance for our model. The hyper-parameters of our HAN+BERT model are shown in table III.

Table III. Hyper-parameters Setting.

Name	**Value**
max sentence length	100
max sentence number in a news	15
Embedding dimension	100
Validation split	20%
Epochs	10
GRU dimension	50
Word dimension	100
Sentence dimension	100
Number of news in a day	15
batch size	256

Our primary evaluation metric is accuracy, which assesses the ratio of correctly predicted instances to the total number of instances in our dataset. Mathematically, accuracy is defined as:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} * 100 \quad (16)$$

4.3. Baselines

In order to investigate HAN+ model performance, we compare it with several baseline methods, including LSTM, LSTM-Attention and HAN models. Below, we provide a brief overview of these methods.

4.3.1 LSTM

Long Short-Term Memory (LSTM) networks, a type of recurrent neural network, are widely used in stock market prediction and their efficacy has been well-documented [23]. LSTM outperforms traditional RNNs by capturing long-term dependencies and regulating information flow through its cell state. Each LSTM unit consists of three gates that control the flow of information, determining

[1] [www.sena.ir, www.boursenews.ir, www.tejaratnews.com]

whether to retain or pass it to the next time step. LSTM is defined as the following equation:

$$
\begin{aligned}
f_t &= g\big(W_f \cdot [x_t . h_{t-1}] + b_f\big) \\
i_t &= g(W_i \cdot [x_t . h_{t-1}] + b_i) \\
o_t &= g(W_o \cdot [x_t . h_{t-1}] + b_o) \\
c_t &= f_t \otimes + i_t \otimes \tanh(W_c \cdot [x_t . h_{t-1}] + b_c) \\
h_t &= o_t \otimes \tanh\,(c_t)
\end{aligned} \tag{17}
$$

where g represents the element-wise Sigmoid function, tanh denotes the hyperbolic tangent function, $\otimes$ signifies the element-wise product, x_t stands for the input vector (such as the word embedding) at time t, h_t represents the output vector (hidden state) and c_t is the cell state vector. Moreover, f_t, i_t and o_t are called the gate of units. Although, the LSTM hidden state h_t captures information only from previous steps, it lacks knowledge of forward sequence information. Consequently, by effectively utilizing Bidirectional-LSTM (BiLSTM), we can retain information from both past and future at each time step, aiding right-to-left languages like Persian by considering two-time directions of input information simultaneously.

4.3.2 LSTM-Attention

Since LSTM models store information for the entire input sequence in a single vector, this can result in information loss as it compresses all necessary details into a fixed-length vector. Additionally, not all words in a text equally contribute to its representation. Leveraging an attention mechanism helps capture the most relevant words. Integrating an attention layer into any LSTM or BiLSTM model can significantly enhance performance, especially in NLP tasks, by improving the accuracy of sequential predictions. Specifically, we can represent attention mechanism as [24]:

$$u_i = \tanh\,(W_n n_i + b_n) \tag{18}$$

$$\alpha_i = \frac{\exp(u_i^T u_n)}{\sum_p \exp\,(u_p^T u_{n)}}$$

$$d_i = \sum_p \alpha_p n_p$$

4.3.3 HAN model

The Hierarchical Attention Network (HAN) captures document structure by first encoding sentences and then forming a document representation while emphasizing words and sentences of varying importance. To account for this, HAN incorporates two levels of attention mechanisms—one at the word level and another at the sentence level. These mechanisms allow the model to appropriately allocate attention to each word and sentence, thereby constructing a more nuanced document representation. The HAN model comprises several components: a word sequence encoder, a word-level attention layer, a sentence encoder, and a sentence-level attention layer. These parts are the same as HAN+ model.

4.4. Recent Approach

To provide a comparative benchmark, we also explore a recent model in the field, which has demonstrated improved performance over traditional approaches.

4.4.1 DistilBERT

DistilBERT [25] is a Transformer-based model designed as a lighter and faster alternative to BERT. It uses knowledge distillation to retain most of BERT's performance while significantly reducing model size and computational cost. Knowledge distillation is a process in which a smaller model (the "student") is trained to mimic the behavior of a larger, pre-trained model (the "teacher"). The distillation process focuses on transferring the most important features learned by the teacher to the student, thereby enabling DistilBERT to maintain high performance despite having fewer parameters. With approximately 60% fewer parameters and 60% faster inference time, DistilBERT is especially well-suited for real-time applications and scenarios where computational efficiency is crucial, while still maintaining 97% of BERT's language understanding capabilities. By maintaining the self-attention mechanism of Transformers, DistilBERT efficiently processes text, making it suitable for NLP tasks with limited resources.

4.5. Experimental Results

This section presents the outcomes of our experiments, which aimed to assess the effectiveness of the proposed hierarchical attention mechanism plus BERT in predicting stock market trends. Our experiments were conducted on a comprehensive dataset of Persian financial news articles, in three different stocks and industries, covering the period of March 21, 2017 to September 22, 2022. Table IV presents the experimental results on all data sets in our study.

Table IV presents the accuracy of various methods across three datasets, with the best results in bold. Our model is evaluated against three baselines. Nguyen and Shirari [26] noted that 56% accuracy in binary stock prediction is considered satisfactory, as even small improvements can yield significant financial gains.

Table IV. Performance of five different models on three datasets.

Model	Dataset	Accuracy of index	Accuracy of industry index	Accuracy of stock index
LSTM	Fameli	58.50	54.98	52.50
	Shebandar	60.01	56.14	53.32
	Shekabir	60.22	57.03	54.67
LSTM+ ATTENTION	Fameli	58.61	55.97	52.66
	Shebandar	60.25	56.8	53.50
	Shekabir	60.70	57.57	55.01
HAN	Fameli	61.65	57.71	54.96
	Shebandar	63.28	58.03	55.27
	Shekabir	62.67	59.15	56.85
HAN+ (This work)	Fameli	63.30	59.44	56.72
	Shebandar	64.56	60.02	57.11
	Shekabir	64.10	61.24	58.66
HAN+BERT (This work)	Fameli	**64.38**	**59.78**	**57.03**
	Shebandar	**65.13**	**61.43**	**58.30**
	Shekabir	**65.49**	**62.25**	**59.25**

In general, the HAN+BERT model outperforms the other four models, which indicates that selecting more important news and understanding the context could achieve better performance in stock market prediction.

The attention mechanism could assign more grant to a specific news which has more impacts on the predictions. LSTM model has the worst performance among all models. The LSTM-Attention model outperforms the LSTM model but the improvement is insignificant. In addition, HAN model also performs better than LSTM-Attention model (by about 5%), but its result is weaker than HAN+ and HAN+BERT model (by 5%-7%).

We compared our model not only with baseline models but also with more advanced transformer-based architectures. Table V presents a comparison between HAN+BERT and one of the successfully executed transformer models. Both models were evaluated under the same conditions. As evident, while DistilBert model achieves a comparable accuracy to HAN+BERT, it exhibits significantly higher computational complexity.

Table V. Comparison of state-of-art model with HAN+BERT.

Model	Accuracy of index	Time Complexity (1 epoch)
DistilBER	66.75	07:33:18
HAN+BERT	65.49	02:40:46

The superiority of HAN+BERT shows that weighting and selecting the most important information from financial and stock market news is quite an effective idea for improving the accuracy of stock price predictions. By applying this hierarchical method, we are effectively filtering out noise and focusing on the most relevant information, which is essential for accurate forecasting. The most important news of a day, the most influential sentences of each news, and the most influential words of each sentence are selected in three levels.

Additionally, using BERT for word embedding is a wise choice, as it captures contextual information and semantics, leading to higher-quality feature extraction. By combining these two strategies (hierarchical information selection and advanced word embeddings) our method is well-equipped to make more precise predictions for stock price fluctuations. The design of this model has been carefully considered, highlighting its importance for success in financial prediction and NLP tasks. Let's summarize the key advantages of Method HAN+BERT over Method HAN:

A) Importance Weighting of News:

- Method HAN+BERT recognizes the inherent variability in the importance of different news articles. By assigning weights to the news based on their significance, the model prioritizes the most influential news events.
- This strategy enables the model to concentrate on critical information, such as major geopolitical events (e.g., war news or JCPOA negotiations in Iran), which can significantly impact financial markets.

B) Optimized Word Embedding:

- Leveraging the effective word embedding methods, such as BERT, demonstrates the importance of capturing contextual information and semantic relationships between words.
- By using advanced word embeddings, HAN+BERT improves the quality of feature extraction and representation, leading to more accurate predictions.

C) Improved Accuracy:

- The combination of importance weighting and advanced word embedding contributes to a 3% increase in accuracy for HAN+BERT compared to HAN and 5% and 7% increase in accuracy compared to LSTM+Attention and LSTM, respectively.
- Although DistilBERT achieved slightly higher accuracy compared to HAN+BERT model, its computational complexity is significantly higher. Given its marginal accuracy improvement alongside substantial temporal and spatial complexity, our model presents a more practical and efficient choice.

In summary, HAN+BERT's ability to weigh the importance of news articles and leverage advanced word embedding techniques has proven to be advantageous, resulting in a more accurate model for predicting stock movements compared to the other methods. The lack of daily news for certain stocks is a common challenge in financial prediction. This situation can pose difficulties for models that rely heavily on news data for making predictions and is resulted the low accuracy of individual stock index prediction.

5. Conclusion

The integration of deep learning with the abundant availability of fine-grained stock price data and financial news has revolutionized the empirical study of news impact on stock prices. Traditional methods struggled with analyzing vast textual data, but neural networks and NLP techniques now enable efficient data analysis. These advancements facilitate (i) large-scale text analysis of financial news and reports, (ii) empirical assessment of stock price reactions to news announcements, and (iii) predictability analysis to uncover patterns in market responses. This progress enhances understanding of how financial information is incorporated into stock prices.

In this study, we collected financial news related to the Tehran Stock Exchange from renowned financial news websites in Iran. We then proposed hierarchical attention network plus BERT (HAN+BERT) model, specifically designed to identify the most important news articles of a given day, extracting the most informative sentences from these articles, and further narrowing down to the essential words within those sentences to enhance the accuracy of stock price forecasts.

The experimental findings demonstrate the efficacy of our proposed HAN+BERT model, reaffirming the belief that news significantly impacts the forecasting of Tehran stock prices. This underscores the potential for further research by incorporating additional factors, such as sentiment analysis of news, and combining these insights with the current work could enhance model accuracy in future studies, advancing the field of stock market prediction.

6. Data Availability

The dataset generated for this study is openly available for public access (https://github.com/LeilaHaf/Data).

7. References

[1] P. Y. Hao, C. F. Kung, C. Y. Chang, J. Ou, "Predicting stock price trends based on financial news articles and using a novel twin support vector machine with fuzzy hyperplane," *Applied Soft Computing*, vol. 98, 2021.

[2] M. F. Shams, M. Mohammadi, "A model for price manipulation prediction case study: Tehran Stock Exchange", *Economic Research*, vol. 11, pp. 30-41, 2011.

[3] M. H. Ziya, A. Vatanka, "The Iranian government's risky stock market bet", Online Content, 2020.

[4] J. Qiu, B. Wang, C. Zhou, "Forecasting stock prices with long-short term memory neural network based on attention mechanism", *PLoS One*, vol. 15, no. 1, 2020.

[5] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, "Hierarchical attention networks for document classification", In the association for computational linguistics: human language technologies, June 2016, San Diego, California, pp. 1480-1489.

[6] L. Huang, H. Yan, S. Ying, Y. Li, R. Miao, C. Chen, Q. Su, "Hierarchical attention network in stock prediction", In the Information Retrieval (Springer), August 2020, China, pp. 124-136.

[7] Q. Liu, X. Cheng, S. Su, S. Zhu, "Hierarchical complementary attention network for predicting stock price movements with news", In the ACM International Conference on Information and Knowledge Management, October 2018, Italy, pp. 1603-1606.

[8] J. Liu, H. Lin, L. Yng, B. Xu, D. Wen, "Multi-element hierarchical attention capsule network for stock prediction" *IEEE Access*, vol. 8, pp. 143114-143123, 2020.

[9] H. Bi, L. Lu, Y. Meng, "Hierarchical attention network for multivariate time series long-term forecasting" *Applied Intelligence*, vol. 53, pp. 5060-5071, 2023.

[10] C. Wang, P. Nulty, D. Lillis, "A comparative study on word embeddings in deep learning for text classification", In the 4th International Conference on Natural Language Processing and Information Retrieval, February 2020, New York, USA, pp. 37–46.

[11] A. Mohammadi, M. Yazdian-Dehkordi, M. A. Nematbakhsh, "Identification of Implicit Features using Persian Language Rules and Sentiments Clustering", *Tabriz Journal of Electrical Engineering*, Vol. 50, no. 3, pp. 1395-1404, 2020 (In Persian).

[12] F. Moodi, A. Jahangard-Rafsanjani, S. Zarifzadeh, ''Improving stock price prediction using technical indicators and sentiment analysis,'' *Tabriz Journal of Electrical Engineering*, 2024 (In Persian).

[13] X. Chen, X. Ma, H. Wang, X. Li, C. Zhang, "A hierarchical attention network for stock prediction based on attentive multi-view news learning", *Neurocomputing*, vol. 504, pp. 1-15, 2022.

[14] J. Zhang, L. Ye, Y. Lai, "Stock price prediction using CNN-BiLSTM-Attention model", *Mathematics*, vol. 11, no. 9, 2023.

[15] Q. Zhang, C. Qin, Y. Zhang, F. Bao, C. Zhang, P. Liu, "Transformer-based attention network for stock movement prediction", *Expert Systems with Applications*, vol. 202, 2022.

[16] X. Zhang, S. Liu, X. Zheng, "Stock price movement prediction based on a deep factorization machine and the attention mechanism", *Mathematics*, vol. 9, no. 8, 2021.

[17] S. Li, S. Xu, "Enhancing stock price prediction using GANs and transformer-based attention mechanisms", *Empirical Economics,* vol. 68, pp. 373–403, 2025.

[18] W. Liu, Y. Ge, Y. Gu, "News-driven stock market index prediction based on trellis network and sentiment attention mechanism", *Expert Systems with Applications*, Vol. 250, 2024.

[19] J. Devlin, M. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", In the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, May 2019, Minneapolis, Minnesota. Association for Computational Linguistics, pp. 4171–4186.

[20] M. Farahani, M. Gharachorloo, M. Farahani, "ParsBERT: Transformer-based model for Persian language understanding", *Neural Process Letter*, vol. 53, pp. 3831–3847, 2021.

[21] K. Cho, B. V. Merrienboer, D. Bahdanau, Y. Bengio, "On the properties of neural machine translation: encoder-decoder approaches", In SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, October 2014, Qatar, pp. 103-111.

[22] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling", Workshop on Deep Learning, 2014.

[23] Y. Liu, Q. Zeng, H. Yang, A. Carrio, "stock price movement prediction from financial news with deep learning and knowledge graph embedding", Knowledge Management and Acquisition for Intelligent Systems, July 2018, China, pp. 102-113.

[24] G. Liu, J. Guo, "Bidirectional LSTM with attention mechanism & convolutional layer for text classification", *Neurocomputing*, vol. 337, pp. 325-338, 2019.

[25] V. Sanh, L. Debut, J. Chaumond, Th. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter", arXiv *preprint arXiv:1910.01108,* 2019.

[26] T. H. Nguyen, K. Shirai, "Topic modeling-based sentiment analysis on social media for stock market prediction", In the association for computational linguistics and the 7th international joint conference on

natural language processing, July 2015, China, pp. 1354–1364.