Journal of Positive School Psychology http://journalppw.com
2022, Vol. 6, No. 5, 9797-9811

Developing a Text Mining Model in Persian News Websites on the Iranian Capital Market

Ramin Kazemi Kalajahi[1], Abolfazl Danaei[2] (Ph.D), Farshad Faezi[3](Ph.D), Maryam Khoshnevis[4] (Ph.D)

[1] . PhD student in Media Management, Semnan Branch, Islamic Azad University, Semnan, Iran.

[2] . Associate Professor, Department of Media Management, Semnan Branch, Islamic Azad University, Semnan, Iran. (Author)

[3] . Associate Professor, Department of Industrial Management, Semnan Branch, Islamic Azad University, Semnan, Iran.

[4] . Assistant Professor, Department of Economic Sciences, Semnan Branch, Islamic Azad University, Semnan, Iran..

Abstract

The critical role of financial markets in the dynamics of economics has made the need to study market trends a necessity and fascinates every investor and economist; Given that the capital market is influenced by multiple factors, it is very difficult to accurately predict the course of its changes. On the other hand, the prevailing uncertainty in the capital market enforces the role of media, especially cyber media, in directing investors. By publishing news about the market's movement and stocks on the web on a large scale, these media, directly or indirectly, encourage their audiences to buy/sell a specific share. In this article, we have tried to analyze the feeling of news published on the Persian websites through text mining patterns. In this regard, after extracting the news through the web crawler, the related news texts were separated through a query. The present study was conducted with the aim of analyzing the feeling of news published on Persian websites through text mining patterns and models. In this regard, after extracting the published news through the web crawler, the related news texts were separated through queries. Then, the number of stock exchange symbols and their corresponding industrial groups were counted mechanically. 9985 news text was manually tagged, and specialized datasets were created. Data preprocessing was performed through the BeautifulSoup and Hazm Python libraries. Vectorization was made by the WordPiece algorithm. Finally, emotion analysis was performed by the parsBERT algorithm and three-class emotion analysis (e.g., positive, negative, and neutral). The accuracy of the model, assessed by the F1-score, precision, accuracy, and recall criteria in the Google Kolab and Jupyter Notebook platforms, was 83.78. Comparing the suggested model with those introduced in former studies, clarified that our model could analyze the feeling of capital market news published in cyber media with acceptable accuracy.

Keywords: text mining, media, stock exchange, capital market

INTRODUCTION

The capital market (stock exchange) is a very attractive and profitable market in any country, which is influenced by multiple factors. For this reason, it is very difficult to accurately predict this market. However, its sensitivity and attractiveness to investors and scholars have led to many studies .

Given the intense uncertainty in the capital market, the media plays a vital role in directing investors. In fact, the media directly/indirectly persuades audiences to buy certain stocks, or enter a certain market by publishing news, texts, analysis, and regular reports on the movement of markets/stocks on various cyber platforms. In addition, intended rumors, as well as ordinary users' opinions in these interactive media are other factors influencing market orientation. A strong correlation between published news and further market performance is discovered (Im et al., 2014). Therefore, detecting patterns in news published in the media about the Iranian online capital market seems useful for predicting the behavior of audiences and the future of the stock market. However, the news week structure limits our understanding. The lack of comprehensive research on the effects of media reports, in particular, cyber media, on the capital market makes this gap quite noticeable. Therefore, this research is performed with the aim of designing a specific model for analyzing news published in cyber media about the Iranian capital market.

Definitions

Text mining: It refers to the automatic/semi-automatic extraction of hidden and potentially valuable information or implicit patterns from a large unstructured textual dataset, including natural language texts (Hassani et al., 2020). Text mining is recently exercised as a practical method in this field.

Object-oriented paradigm (OOP): OOP focuses primarily on how objects interact to communicate and share information, and pursues at least three goals: reusability, scalability, and flexibility (Triaji, Pratomo, and DP, 2021). In this programming approach, the program tends to an object. In other words, the data and probable operations are gathered together as much as possible in a form called an "object" to make a "unite" (object). Then, they will be isolated from the external environment, so that the alien operations outside the object can no longer change the inside data (Kendal, 2009).

Python Language: It is a powerful, high-level, object-oriented, and professional programming language (Srinath, 2017) first published in 1991. Python language has a rich and agile ecosystem, and contains several useful libraries (Babuji et al., 2019).

Python Library: A collection of modules written in C or Python languages (Nagpal and Gabrani, 2019).

Web Crawler: It is a program/software or script that scans the World Wide Web in a systematic and automated manner. The web structure is a directional graph in which the web page and link acts as node, and edge, respectively. Thus, the search operation can be summarized as a directional graph navigation process. By following the web link structure, the web crawler may scroll through several new web pages starting from the same page. This crawler moves from one page to another through the graphical structure of web pages. These programs are also known as robots and spiders. Web crawlers are designed to retrieve web pages and insert them into local datasets (Abu Kausar, Dhaka, and Kumar Singh, 2013).

ETL: A regular sequence of operations – i.e., extraction, conversion, and loading - performed with the aim of systematically processing the source data in order to make it available in a more convenient format for the intended use (Galici et al., 2020).

Scrapy: A distributed Python-based crawler framework for large-scale web exploration that quickly and efficiently reaches web pages and extracts relevant information. This framework completes the secondary development according to user needs with good scalability (Ma and Yan, 2021).

Scrapyd: A program to deploy and execute code sources of scrapyd crawlers. This program has made it possible to deploy and load projects (Eyzenakh, Rameykov, and Nikiforov, 2021).

Machine learning: It enables computers to learn without explicit programming. This type of learning focuses on setting up and exploring the methods and algorithms by which computers and systems are able to learn and learn, and enable efficient data management. Machine learning is applied when interpreting data from the extracted information is not possible after viewing the data (Mahesh, 2018).

parsBERT: BERT is a pre-trained natural language processing model with special popularity due to its good performance. However, this model focuses on the English language, while other languages with limited resources use multilingual models. The parsBERT is a monolingual BERT model in the Persian language which performs better than other multilingual architectures, like the BERT multilingual model. Despite the very limited dataset in Persian regarding the natural language processing, a large linguistic dataset, collected from various sources, has been used to train this model (Farahani et al, 2020).

Research background

In recent years, numerous studies have investigated text media analysis in cyber media and its application in various domains. Many of these studies have examined the stock market, various methods for predicting market trends, and the impact of news media on further stock market performance in different countries. In the following, several studies whose subjects are close to the present study are summarized.

In "Understanding the Pandemic through Mining COVID News Using Natural Language Processing" in 2021, natural language processing techniques to extract knowledge about the Coronavirus, including the number of patients with COVID-19 and popular topics per month were used to analyze emotions. In this study, a new dataset called "NNK Dataset" was introduced, which included 1,050 newspaper articles. It showed that news reports could guess the epidemic situation based on different news sources. Also, during various experiments, the importance of NLP in the analysis of newspaper reports was proved (Sadman et al., 2021).

In "Applying Text Mining Methods to Extract Information from News Articles" in 2021 in Bulgaria, the applications of text mining methods were examined to extract information from news articles. It was reported that the extracted information could be useful for researching topics covered in news articles (Georgieva and Dechev, 2021).

For assessing environmental performance in tourism areas and examining its effective factors using web news text mining in China, Fang et al. (2020) performed "What Can The News Tell Us About The Environmental Performance Of Tourist Areas? A text mining approach to China’s National 5A Tourist Areas". This study, which used more than 1,300,000 words from online news sources, confirmed the effectiveness of text mining regarding environmental news (Wang et al., 2020).

In "The Role of Text Mining in Mitigating the Threats from Fake News and Misinformation in Times of Corona" in 2021, a text prototype mechanism was implemented to detect fake news and misinformation using the text mining technique. This system provided patterns for presenting facts in texts, and by abstractly representing facts, searched for similar facts to detect fake news and false information. The data sources used for this article were COVID-19 articles published in various newspapers and on Twitter (Englmeier, 2021).

In a study entitled, "The Impact of Persian News on Stock Returns through Text Mining Techniques", after collecting information about the stock index and news published during a certain period about the Tehran Stock Exchange, several Iranian scholars used techniques of text mining and emotion analysis to determine the semantic load of news sentences. Using machine learning algorithms, they divided news into positive and negative categories, and finally, evaluated the relationship between news and stock index. Based on the results, the published news had a positive or negative semantic load and influenced the value of the index. The accuracy of the model presented in this study was 52.2% (Azizi et al., 2021).

In "COVID-19 Public Sentiment Insights: A Text Mining Approach to the Gulf Countries" in 2020, semantic analysis at three levels (negative, neutral, and positive) was used to measure emotions towards epidemic and quarantine. The 2-month Twitter dataset was analyzed using the Natural Language Processing (NLP) technique. The results showed that the feelings of the Persian Gulf countries towards the COVID-19 epidemic were approximate as follows: 50.5% were neutral, 31.2% were positive and 18.3% were negative. This study yielded successful results in regard to social media textual analysis (Albahli et al., 2021).

An algorithm that used the text mining methods to categorize the automotive industry according to its competitive actions was introduced in "Developing an Advanced Algorithm Capable of Classifying News, Articles and Other Textual Documents Using Text Mining Techniques". Preliminary experiments led to the identification of two logistic regression (LR) and artificial neural network (ANN) algorithms. After testing several parameters in each algorithm, the best outcome was appeared by ANN. Accuracy, recall, and F1 scores of the final model were, 0.80, 0.78, and 0.76, respectively. After removing three noise-making classes, the accuracy of the final algorithm increased to 0.94 (Knudsen, Rasmussen, and Alphinas, 2020).

Zhao and Zeng (2019) in "Analysis of Timeliness of Oil Price News Information Based on SVM", implemented various analysis techniques to predict international oil prices through collecting and extracting news published on a network. They suggested a new method based on SVM (Support Vector Machine) to check the timeliness of oil price news. This was a multi-scale trend detection method to extract more flexibly the trend of oil price fluctuations in various scales and dimensions. Experimental results supported the high reliability of this method. It was also found that the news provided information about relatively long-term trends, and the information described in the news affected the oil future more strongly than fluctuations in past prices (Zhao and Zeng, 2019).

Sun et al., (2019) conducted "How Mood Affects the Stock Market: Empirical Evidence from Microblogs". They examined 22,504 tweets extracted from a microblog site, Sina Weibo, and identified two clusters of microblog users. Then, by performing text mining, the pathways of tweets' effects on the stock market were clarified. An inverted U-shaped curve between stock returns, and the news focus and investors. It was also confirmed that news focus had a positive moderating effect on the relationship between investors' attention and stock returns. Finally, it was discovered that social interaction can moderate the impact of news media and investors' inclinations on stock returns (Sun et al., 2019).

In "Can Economic News Predict Taiwan Stock Market Returns?" a number of Taiwanese researchers (2019) examined the usefulness of news to predict stock returns in the Taiwan stock market through text mining. After extracting the text of economic news, the text documents were turned into a matrix of keywords, and the outcomes were used as news variables. Along with other quantitative variables, a stable model was presented to examine stock market return behaviors in 20 shares of the studied set (January 2008 to December 2014). Experimental analysis showed that news variables provided useful information for predicting Taiwan stock market returns. Additionally, economic news had an asymmetric effect on the prediction of stock market returns, meaning that the forecast accuracy of a stock market boom was greater than that of a recession (Wu, Hou, and Lin, 2019).

In 2017, " Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis" was performed with the aim of creating an effective model for predicting future stock market trends with a low error ratio and improving forecasting accuracy. This model provided a forecasting opportunity based on the analysis of financial news sentiment and historical stock market prices. A dataset containing stock prices of three companies was used. In the first phase, news emotion analysis was performed to identify the polarity of the text using the Naive Bayes algorithm. The results of the prediction accuracy ranged from 72.73% to 86.21%. In the second phase, a combination of news poles and historical stock prices was used to predict future stock prices; the forecast accuracy improved to 89.80 (Khedr, Salama, and Yaseen, 2017).

In his doctoral dissertation, Nassirtoussi (2015) performed "Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment" At the University of Kuala Lumpur, Malaysia. Recent studies in behavioral economics confirm that the behavioral responses of all investors to the information published in the news are effective in increasing / decreasing prices. This theoretical basis formed the economic foundation of this research. This researcher proposed a new approach to predicting the overnight movement of a currency pair in the foreign exchange market based on the texts about financial news. A predictive relationship between this type of market (foreign exchange market) and news text data was found. The proposed techniques implemented in this system led to a significant increase in model accuracy (up to 83.33) (Nassirtoussi, 2015).

Bhardwaj et al. (2015) conducted "Sentiment Analysis for Indian Stock Market Prediction Using Sensex and Nifty". Natural language analysis (NLP) was used to analyze emotions or comments, while user analysis, evaluation, inclinations, attitudes, and emotion analysis were performed to identify and extract mental content. Additionally, important indicators including Sensex and Nifty were used in stock market sentiment analysis to predict stock prices. For this purpose, the Python programming language, which has a fast executable environment, was used. The researchers acknowledged that the Python script could be more advanced in the future (Bhardwaj et al., 2015).

In 2014, Rechenthin conducted his doctoral dissertation on "Machine-learning classification techniques for the analysis and prediction of high-frequency stock direction" at the University of Iowa in the United States and forecasting the stock market. She designed a decision support framework that helped traders to suggest signs of future stock prices, as well as the possibilities associated with them. It selected a different approach to improve machine learning with outstanding concepts. In other words, the concepts were assumed to be floating and turned into a model. The framework built thousands of conventional classifiers (SVMs, decision trees, and neural networks) through random subsets of past data, and by covering similar stocks, as well as an exploratory combination of the best basic classifiers, succeeded in adapting to changes in the stock market, and achieved better outcomes for forecasting the future market's direction (Rechenthin, 2014).

In "Text Mining Approaches for Stock Market Prediction", Nikfarjam, Emadzadeh, and Muthaiyah (2010) presented different methods to examine the impact of financial news on the stock market forecast. For determining the news tags, prices were generally controlled in accordance with the time of news release. However, the technical analysis of the stock market through the news led to more accurate tags. Seven technical indexes were used, and after analyzing the index and news values separately, the results were combined. The accuracy of the studied methods was 50.2 to 78.4. The researchers acknowledged that more accurate results could be obtained if the classifier input included simultaneously both the news and market conditions. They reported that using new semantic methods in text classification may lead to a promising outcome for the issue under discussion (Nikfarjam, Emadzadeh, and Muthaiyah, 2010).

Niknam and Niknafs (2016) performed "Improving Text Mining Methods for Market Forecasting through Prototype Selection Algorithms" in Iran. They presented a feature selection method based on target features that reduced the size of the feature space. To prevent the increase in the volume of training samples, which is the result of using the initial sample selection methods, the training set was modified. The method was implemented in three phases; each phase was an improved version of the former phase. Each phase's output was satisfactory, while the highest efficiency was achieved at the end of the third phase. This method was compared with a successful algorithm in market forecasting; despite the reduction of training samples by the prototype selection algorithms, the suggested method presented much better outcomes. It was claimed that by considering the basic data in the proposed method, it can be used to predict the stock market index in Iran (Niknam and Niknafs, 2016).

In 2016, a research entitled "Predicting Economic Indicators by Text Mining of Persian News" in Iran, investigated changes in the price of the total index of the Iranian Stock Exchange by analyzing the news of the last 10 years published in Hamshahri Newspaper as well as historical data of the stock exchange. After pre-processing the textual data, text mining methods were performed. In the first phase, keywords affecting economic indicators were used as vector characteristics of machine learning algorithms. In the next phase, relying on text mining algorithms, the future of the stock market was predicted through text mining news. The accuracy of forecasting the future of the stock market using text mining was reported in this study as 68.9% (Farazandehnia, Majidi, and Movaghar, 2016).

Methodology

First, specific patterns for analyzing news published in cyber media about the Iranian capital market were identified, and then, a new model for analyzing these news contents was presented. For this purpose, stock market news and analysis were collected from two platforms: Persian websites and Telegram channels. In the next phase, the data were analyzed by artificial intelligence methods and the feelings of each news (positive, negative, neutral) were extracted. Python coding in Google Kolab and Jupyter Notebook platforms was run to implement algorithms related to metadata processing, artificial intelligence, and text mining.

Project's Architecture Details: The following figure shows the project's architecture in full detail, along with the used libraries, testing process, and data generation method.

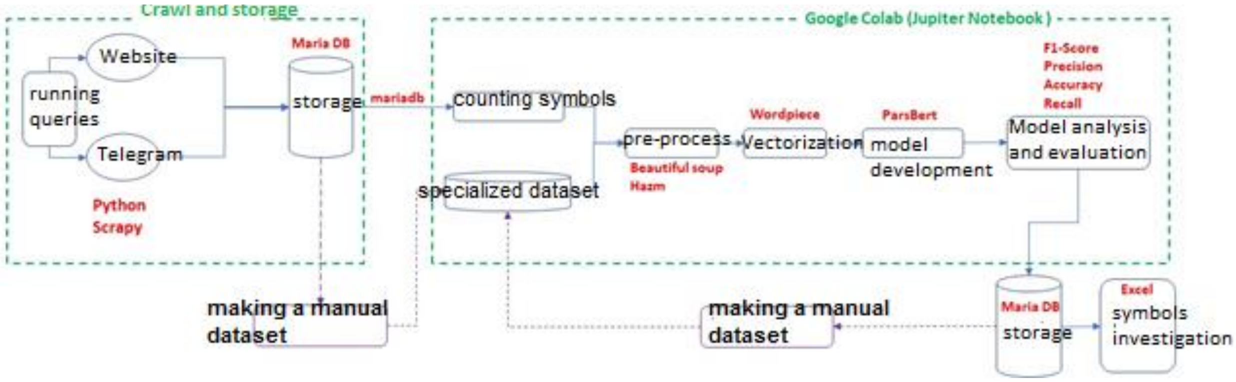

Figure 1- Details of the project architecture

Data Collection:

In this study, news texts were collected from two platforms – Persian websites and Telegram channels - using web crawlers in the Python language, and through nine queries. Web crawlers are managed by Scrapyd, a management system for data collection (Eyzenakh, Rameykov, and Nikiforov, 2021). These data were put in the system datasets, and prepared for the text mining process after the initial processing.

Q1	=	Capital Market / Stock Exchange / OTC / Hafez Hall / Basic Market
Q2	=	Technical analysis / Fundamental analysis
Q3	=	Codal / Supreme Council of the Stock Exchange
Q4	=	First Market / Second Market
Q5	=	Symbol / Capital increase
Q6	=	Symbol & (share / stock / shareholder / stock exchange / capital market)
Q7	=	Revaluations and (share / stock / symbol)
Q8	=	Total Index & (Capital Market / Stock Exchange / OTC / Base Market)
Q9	=	Initial offering & (stock / symbol / stock / stock exchange / OTC)

Sample size: News texts are among the most basic contents that users try to understand their specific meanings. News, articles, and news statements contain meaning and knowledge. News texts are made up of several words, some of which are more important, and called "keywords". Text mining techniques help to find the connection between the keywords in each news text and facilitate its meaning comprehension.

The unit of analysis in this research is "news". News related to the capital market available on the Persian websites and official Telegram channels in a one-year period were collected mechanically through the web crawler (Python and Scrapy) and nine queries. After running ETL, pre-processing, and other tasks (e.g., cleaning the data and clearing the html tags), these news were stored in the system dataset. A total of 9985 news with a symbol regarding a stock exchange or industrial group were entered into the text analysis process.

Data Analysis Tools

In order to increase the efficiency, optimization, speed and ease of implementation, Google Kolab and Jupyter Notebook were used. These tools provide an experimental environment for initial development and implementation when limited server resources and hardware are available.

Google Kolab: It allows all users to write and execute custom Python code through a browser (Kuroki, 2021), and is especially useful for machine learning, data analysis, and training machine learning models, which are discussed below. More technically, Google Kolab is a hosted Jupyter Notebook service that requires no special settings and is run entirely in the cloud. Performing Google Kolab provides an opportunity to write and execute codes, store and share analysis, and access to powerful computing resources, including GPUs, for free (Ray, Alshouiliy, and Agrawal, 2020). Google Kolab stores project files in Google Drive (Gunawan et al., 2020). In the current study, this tool was used in the experimental implementation stage.

Jupyter Notebook: As open-source software for coding and training, Juoyter Notebook is increasingly used for coding and training. Also, it is one of the most widely used software due to its simplicity and active support for graphical and numerical libraries. Jupyter Notebook is a popular programming platform that easily processes programming languages including Python (Tang, 2021).

Data Analysis

The text mining process was first performed in the lab environment of Google Kolab and Jupyter Notebook platform. Since emotion analysis should be run based on symbols and industrial groups, at this stage, the number of each symbol in each of the news, the number of each symbol in the total news, and the number of news for each industrial group were counted. Of the total collected news, 9985 news contained a symbol about an exchange or an industrial group. Due to project implementation considerations, only these news were entered the text mining process.

Counting the number of each symbol in each of the news: Two separate tables were formed in the dataset. One included the collected news intended to be searched in, and the other included the stock market symbols and their correspondent industrial groups. By referring to these tables, each stock symbol was counted in each of the news.

In this phase, words similar to a symbol but with different meanings needed paying special attention. However, since most of the symbols have unusual names, this did not affect the output significantly, while artificial intelligence techniques could solve this problem.

```
for i in range (0, len(dfSymbol)):
   initialDataframe [dfSymbol.at [i, 'symbol']] =
dfNews.transpose(). iloc [0]. str. count (' ' +
dfSymbol.at [i, 'symbol'] + ' ')
```

Number of news for each industrial group: In the symbols-containing table, the corresponding industrial groups of each symbol were also listed. This column was used in addition to the previous section, and the total news available for each industry group was displayed. An example of the table is manifested in the figure below.

Agriculture and related services	Computer and related services	Electricity, gas, steam and hot water supply
1	644	1

Figure 2 - An example of the number of industrial groups in the total news

```
for i in range (0, len (dfSymbol)):
   if countSymbolInNews[i]! = 0:
      industrialCountArr.append
({symbolBesideIndustryGroup.at [i, 'Industry
groups']: countSymbolInNews[i]})

# Merge same keys of industrialCountArr
countIndustrialGroupsInNews = dict (functools.
reduce (operator.add, map (collections. Counter,
industrialCountArr)))
```

Pre-processing the News

The input data were pre-processed. The pre-processing phase, which was facilitated through the BeautifulSoup and Hazm libraries, focused on cleaning data, normalizing news texts, and eliminating stop words. The WordPiece algorithm was implemented to vectorize the data. At the end, the news' feelings were analyzed with the parsBERT algorithm. In the present study, a three-class emotion analysis (positive, negative and neutral) was applied.

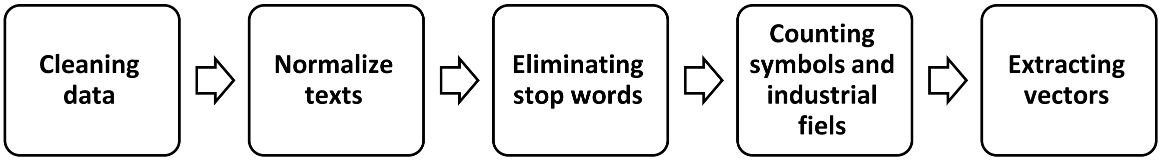

Figure 3 - The process stages of the data, from data storage to enter them into the artificial intelligence algorithm

Cleaning the data: This process is run to identify inaccurate, incomplete, or irrational data, and to improve the data quality by correcting identified errors and omissions. In general, data cleaning helps to foster the quality by reducing errors and improves data quality (Kumar Singh and Kumar Dwivedi, 2020).

Normalizing the data: This phase was dedicated to standardize the news texts. Sometimes texts are very similar but the machine considers them as different due to simple differences in appearance. Therefore, it was attempted to eliminate these simple apparent differences. For this, the texts were pre-processed before comparison. Clearly, the stronger these pre-process, the more reliable the results of comparing texts would be. However, much more problems arose due to the non-structured nature of the Persian language.

Unstructured texts, those with no specific format presuppositions, are considered a regular set of sentences. Processing the Persian language processing is different from the English language. In the English, all letters and words are written separately with a specific rule, but in the Persian, some letters are stuck together while, others are written separately; some words are integrated, and some words are divided into two or more parts by a space or a half distance. All aspects of natural language processing deal with real texts in some way. Non-standard forms of letters and words are abundant in this type of written text (Ghafouri et al., 2009).

Eliminating the stop words: Stop words are those that, despite their great repetition in most texts, lack meaningful information (Ousirimaneechai and Sinthupinyo, 2018), like "if", "and", "but", and "that". Deleting these words increases the algorithms' efficiency and reduces the processing volume.

Extracting the vectors: Most machine learning methods are applicable to numerical data, while applying them to textual data needs texts to be converted to a set of numbers. Therefore, all approaches to converting text to numerical vectors try to extract a set of appropriate features from natural language texts (Kumari Singh and Shashi, 2019). In this phase, texts' features were extracted and made ready to be delivered to the artificial intelligence algorithm.

Analyzing The News' Feelings With parsBERT

9985 news texts were read manually, and the emotion analysis process was performed by a human. Then, 9985 data were labeled and parsBERT, an artificial intelligence algorithm for emotion analysis, was run.

The most important symbols in each of the news: The files in the first section were analyzed, and a maximum of three symbols and three industrial groups of symbols with the most repetition in the news text were extracted.

	Content	Cleaned	Industry group 1	Industry group 2	Industry group 3	Symbol 1	Symbol 2	Symbol 3
0	Aria Bazaar - People who <p> intend to buy and sell directly….	Aria Bazaar - People who intend to buy and	Retail except motor vehicles	Automobiles and auto parts manufacturing	None	Ofogh	Khodro	None

		sell directly….						
1	According to the online economy <p> quoted by the Senat; Yasser Fallah…	According to the online economy quoted by the Senat; Yasser Fallah…	Retail except motor vehicles	Automobiles and auto parts manufacturing	None	Ofogh	Khodro	None
2	<ul> <li role="presentation" Id="menu-item-3…	SMS system support for prices, frequently asked questions	chemical products	Automobiles and auto parts manufacturing	Insurance and pension fund except Social Security Insurance	Parsian	Hamrah	Pars
3	<ul> <li role="presentation" Id="menu-item-3…	SMS system support for prices, frequently asked questions	chemical products	Automobiles and auto parts manufacturing	Insurance and pension fund except Social Security Insurance	Parsian	Hamrah	Pars
4	<ul> <li role="presentation" Id="menu-item-3…	SMS system support for prices, frequently asked questions	chemical products	Automobiles and auto parts manufacturing	Insurance and pension fund except Social Security Insurance	Parsian	Hamrah	Pars

Figure 4- Initial news, cleaned news, extracting in maximum of three industrial groups and three symbols from that news

Analyzing the News' Feelings

Pre-processing - At this stage, the news were divided into two categories: educational data and experimental data. For modeling, the educational data were used, while the experimental data were deployed for testing the suggested model. 9985 news were tagged and used for the further process. Each tag represented the perceived feeling of news that could be positive, negative or neutral. As mentioned earlier, the BeautifulSoup and Hazm libraries in the Python environment were used for pre-processing. In the cleaning step, keywords (including "from", "and", "we", etc.) were removed. Other tasks were also performed during the cleaning phase, including eliminating the stop words. Additionally, according to the used dataset and machine learning algorithm, the words described below were added to this list.

```
# clean text

def cleanhtml(raw_html):
   cleanr = re.compile('<.*?>')
   cleantext = re.sub(cleanr, '', raw_html)
   return cleantext

def cleaning(text):
   text = text.strip()

   # regular cleaning
   text = clean(text,
             fix_unicode=True,
             to_ascii=False,
             lower=True,
             no_line_breaks=True,
             no_urls=True,
             no_emails=True,
             no_phone_numbers=True,
             no_numbers=False,
             no_digits=False,
             no_currency_symbols=True,
             no_punct=False,
             replace_with_url="",
             replace_with_email="",
             replace_with_phone_number="",
             replace_with_number="",
             replace_with_digit="0",
             replace_with_currency_symbol="",
             )

   # cleaning htmls
   text = cleanhtml(text)

   # normalizing
   normalizer = hazm.Normalizer()
   text = normalizer.normalize(text)
```

Extracting the vector and adding it to the model: The WordPiece vectorization algorithm was used to turn the news texts into the model's inputs. Then a machine learning algorithm was selected, and after adjusting its configuration, the training vectors were added to the model to train its structure. The parsBERT algorithm was performed in this phase.

```
#class for text vectorize/embedding

class InputExample:
   """ A single example for simple sequence
classification. """

   def __init__ (self, guid, text_a, text_b=None,
label=None):
      """ Constructs a InputExample. """
      self.guid = guid
      self.text_a = text_a
      self.text_b = text_b
      self.label = label

#tokenize data

train_dataset_base, train_examples =
make_examples (tokenizer, x_train, y_train,
maxlen=128)
valid_dataset_base, valid_examples =
make_examples(tokenizer, x_valid, y_valid,
maxlen=128)

test_dataset_base, test_examples =
make_examples(tokenizer, x_test, y_test,
maxlen=128)
[xtest, ytest], test_examples =
make_examples(tokenizer, x_test, y_test,
maxlen=128, is_tf_dataset=False)
```

Testing the model: available experimental data were tested on the trained model. As mentioned, the output of this step for each of the news was 0, 1 or 2 for a negative, neutral or positive feeling, respectively.

```
def build_model (model_name, config,
learning_rate=3e-5):
   model =
TFBertForSequenceClassification.from_pretraine
d (model_name, config=config)

   optimizer =
tf.keras.optimizers.Adam(learning_rate=learnin
g_rate)
   loss =
tf.keras.losses.SparseCategoricalCrossentropy(
from_logits=True)
   metric =
tf.keras.metrics.SparseCategoricalAccuracy('ac
curacy')
   model.compile(optimizer=optimizer,
loss=loss, metrics=[metric])

   return model

#train model with our data

r = model.fit(
   train_dataset,
   validation_data=valid_dataset,
   steps_per_epoch=train_steps,
   validation_steps=valid_steps,
   epochs=EPOCHS,
   verbose=1)
```

```
# print('FINAL ACCURACY MEAN: ',
np.mean(final_accuracy))

predictions = model.predict(xtest)
ypred = predictions[0].argmax(axis=-1).tolist()
```

Assessing the output accuracy: To measure the accuracy of the model, the output difference of the model was measured with the correct labels in the experimental data. Among various criteria for this purpose, the "accuracy" criterion was chosen.

At this stage, 9985 labeled data about the capital market (specialized dataset) were prepared. About 70% of them were returned to the model, and the model was trained. Finally, 3,000 data (about 30%) about the capital market were entered into the model as experimental input. The outcomes are presented below:

	Precision	Recall	F1-score	Support
Negative	1.00	0.57	0.73	170
Neutral	0.74	1.00	0.85	1350
Positive	1.00	0.73	0.84	1480
Accuracy			0.84	3000
Macro avg	0.91	0.77	0.81	3000
Weighted avg	0.88	0.84	0.84	3000
F1: 0.837803839715638 = **83.78**				

The top three rows of the table contain negative, neutral, and positive items. Each of the labels was examined separately. F1-score is the harmonic mean of accuracy and the most accurate measure of accuracy. **Precision** is the ratio of correctly predicted positive observations to the total positive observations. High precision refers to a low false positive rate. **Recall** is the ratio of positive observed predictions to the total number of observations. In general, a recall value above 0.5 is reasonably good (Chandrika and Srinivasan, 2021).

In other words, data accuracy refers to the difference between actual data and predicted data. The first three lines represent the negative, neutral, and positive data, respectively. The row "Macro avg" represents the arithmetic mean (the top three parts), while the last row shows the weighted average (the top three parts). The first three columns represent the three different indexes of data accuracy described in the previous paragraph. The last column ("Support") indicates the number of data in that section. Finally, by calculating the F1 score for the total data, the total accuracy was determined. The output of the model was satisfactory, and its accuracy reached 83.78%.

Discussion and Conclusion

Given the vital importance of the capital market in the economic development of any country, and the necessity of gaining knowledge about market trends, the impact of cyber media such as websites and social networks on fluctuations in the capital market due to increasing news and textual information, has made the exploration of textual data and published news and their analysis in this type of media especially important. In the current study, a text analysis model was presented for analyzing news and texts published in cyber media about the Iranian capital market. Text mining approaches usually refer to the process of extracting valuable information from an unstructured text. In summary, identifying hidden information in news articles is a challenge that could be supported by implementing appropriate text mining algorithms (Georgieva-Trifonova and Dechev, 2021). In this study, Persian news texts published on the Persian websites and Telegram channels were extracted through Python and Scrapyd, and put in the system datasets. In the next phase, the capital market-related news was extracted by nine queries and prepared for the text analysis process.

This research was performed in a laboratory environment. Therefore, to test the text mining algorithms and prepare the final project dataset, increase efficiency, optimization, speed and ease of implementation, as well as to prevent rising infrastructure costs, the Jupyter Notebook was used in the Google Kolab. Since emotions were analyzed based on symbols and industrial groups, the number of each symbol in each of the news and in the total news, and the number of news for each industrial group were counted mechanically. Among the total collected news, 9985 news had the symbol of an exchange or an industrial group which entered into the text mining process. Input data were also pre-processed. The BeautifulSoup and Hazm libraries were used at this stage. The preprocessing phase focused on clearing data, normalizing news texts, and eliminating the stop words. The news was then divided into two categories: educational and experimental data. The educational data were used for modeling, while the experimental data were deployed for testing the suggested model. 9985 news items were manually tagged and advanced to the next stage. Each tag was the news-related sentiment which could be positive, negative, or neutral. The WordPiece algorithm was used for vectorization, and emotion analysis was performed by the parsBERT algorithm. Satisfactory results were also reported for using this algorithm by Azizi et al., (2021). About 70% of the labeled data related to the capital market (specialized dataset) was deployed to train the model. Finally, 3,000 capital market-related data entered into the model as the experimental input. The accuracy of the model at this stage was 83.7%, which is very desirable. F1-score, accuracy, precision and recall criteria were calculated to measure the model's accuracy.

As mentioned before, emotion analysis in the present study was performed based on the three categories of positive, negative and neutral emotions. It was manifested that our models could performe better than other previous models, including the one suggested by Azizi et al., 2021) with an accuracy of 52.5% (Azizi et al., 2021).

The results of text mining in the Google Kolab laboratory environment confirmed the acceptable accuracy of the suggested model in predicting the feeling of incoming news. This research focused only on identifying the title of the symbol and the related industrial group in the extracted news. Therefore, by focusing on other news parameters and enriching the specialized dataset, better results could be obtained from predicting the feeling of news texts compared to using symbols alone. In other words, by optimizing existing queries and using more accurate queries, the text mining process will be performed with fewer resources. Undoubtedly, this suggestion will be of special importance in the commercialization of this model.

Finally, given the lack of specialized datasets and the shortcomings of available general datasets in Persian, particularly specialized datasets for capital market news, as well as the acceptable accuracy of existing machine learning algorithms, it is suggested that future researchers create more accurate and richer specialized datasets in order to achieve more accurate results in the field of specialized text mining of news in the Persian websites.

REFERENCE

Abu Kausar, Md. Dhaka, V. S. Kumar Singh, Sanjeev. (2013). Web Crawler: A Review. International Journal of Computer Applications (0975 – 8887) Volume 63– No.2, February 2013

Albahli, Saleh. Et al. (2021). COVID-19 Public Sentiment Insights: A Text Mining Approach to the Gulf Countries. Computers, Materials & Continua. Vol.67, No.2: 1613-1627

Azizi, Zahra. Abdolvand, Neda. Ghalibaf Asl, Hassan. Rajaee Harandi, Saeedeh. (2021). The Impact of Persian News on Stock Returns Through Text Mining Techniques. Iranian Journal of Management Studies (IJMS) 2021, 14(4): 799-816

Babuji, Yadu Et al. (2019). Parsl: Pervasive Parallel Programming in Python. High Performance Distributed Systems (Best Paper Nominees). HPDC '19, June 22–29

Bhardwaj, Aditya. et al. (2015). Sentiment Analysis for Indian Stock Market Prediction Using Sensex and Nifty. Procedia Computer Science, 70: 85 – 91

Chandrika, P. V. Sakthi Srinivasan, K. (2021). Predicting Stock Market Movements Using Artificial Neural Networks. Universal Journal of Accounting and Finance 9(3): 405-410, 2021

Englmeier, Kurt. (2021). The Role of Text Mining in Mitigating the Threats from Fake News and Misinformation in Times of Corona. Procedia Computer Science 181 (2021) 149–156

Eyzenakh, D.S. Rameykov, A.S. Nikiforov, I.V. (2021). High Performance Distributed Web-Scraper. Trudy ISP RAN/Proc. ISP RAS, vol. 33, issue 3, 87-99

Farahani, Mehrdad. Et al. (2020). ParsBERT: Transformer-based Model for Persian Language Understanding. Preprint, compiled June 2, 2020

Farazandehnia, Mansoureh. Majidi, Babak. Movaghar, Ali. (2016). Predicting Economic Indicators By Text Mining of Persian News. First International Conference on Computer Science and Engineering. 22-21 February 2017.

Galici, Roberta. Et al. (2020). Applying the ETL Process to Blockchain Data. Prospect and Findings. Information 2020, 11, 204

Georgieva-Trifonova, Tsvetanka. Dechev, Miroslav. (2021). Applying Text Mining Methods to Extract

Information From News Articles. IOP Conf. Series: Materials Science and Engineering 1031 (2021) 012054

Ghafouri, Seyyed Majid. Rahati, Saeed. Pahlavan Nezhad, Mohammad Reza. Azimi Zadeh, Ali. (2009). Normalization of Persian Texts, 15th Annual Computer Conference of Iranian Computer Association, Tehran, https://civilica.com/doc/79184.

Gunawan, Teddy Surya. Et al. (2020). Development of video-based emotion recognition using deep learning with Google Kolab. TELKOMNIKA Telecommunication, Computing, Electronics and Control Vol. 18, No. 5, October 2020, pp. 2463-2471

Hada, Riwa Rambu. Hariadi, Enda Fajar. (2020). Implementasi Text Mining Dengan Fitur Tf ldf Pada Pencarian Berita bErbahasa Indoneisa. JOINCS (Journal of Informatics, Network, and Computer Science) | Vol. 3, No.1:1-10

Hassani, Hossein. Et al. (2020). Text Mining in Big Data Analytics. Big Data Cogn. Comput. 2020, 4, 1

Hussein Ali1, Ahmed. Zaki Abdullah, Mahmood. (2019). A Survey on Vertical and Horizontal Scaling Platforms for Big Data Analytics. INTERNATIONAL JOURNAL OF INTEGRATED ENGINEERING VOL. 11 NO. 6 (2019) 138-150

Im, T. L., San, P. W., On, C. K., Alfred, R., & Anthony, P. (2014). Impact of Financial News Headline and Content to Market Sentiment. International Journal of Machine Learning and Computing, 4(3): 237-242.

Kendal, simon. (2009). Object oriented programming using java.

Khedr, Ayman E. Salama, S.E. Yaseen, Nagwa. (2017). Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis. I.J. Intelligent Systems and Applications, 7: 22-30

Knudsen, R. B. Rasmussen, O. T. Alphinas, R. A. (2020). Developing an Advanced Algorithm Capable of Classifying News, Articles and Other Textual Documents Using Text Mining Techniques. International Journal of Electrical and Information Engineering Vol:14, No:11, 362-370

Kumari Singh, Anita. Shashi, Mogalla. (2019). Vectorization of Text Documents for Identifying Unifiable News Articles. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 7, 2019, 305-310

Kuroki, Masanori. (2021). Using Python and Google Kolab to teach undergraduate microeconomic theory. International Review of Economics Education 38 (2021) 100225

Ma, Xiaoju. Yan, Min. (2021). Design and Implementation of Craweper Based on Scrapy. Journal of Physics: Conference Series. Ser. 2033 012204

Mahesh, Batta. (2018). Machine Learning Algorithms - A Review. International Journal of Science and Research (IJSR). Volume 9 Issue 1, January 2020, 381-386

Nagpal, Abhinav. Gabrani, Goldie. (2019). Python for Data Analytics, Scientific and Technical Applications. 978-1-5386-9346-9/19. 2019 IEEE

Nassirtoussi, Arman Khadjeh. (2015). Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment. PhD (Doctor of Philosophy) thesis, University of MALAYA

Nikfarjam, Azadeh. Emadzadeh, Ehsan. Muthaiyah, Saravanan. (2010). Text Mining Approaches For Stock Market Prediction. Computer and Automation Engineering (ICCAE), 2010 the 2nd International Conference on, Volume 4: 256:260

Niknam, Farzad. Niknafs, Ali Akbar. (2016) Improving Text Mining Methods For Market Forecasting Through Prototype Selection Algorithms. Journal of Information

Technology Management, Faculty of Management, University of Tehran, Volume 8, Number 2, pp. 434-415.

Ousirimaneechai, Nattapong and Sinthupinyo, Sukree. (2018). Extraction of Trend Keywords and Stop Words from Thai Facebook Pages Using Character n-Grams. International Journal of Machine Learning and Computing, Vol. 8, No. 6, December 2018. 589-594

Ray, Sujan. Alshouiliy, Khaldoon and Agrawal, Dharma P. (2020). Dimensionality Reduction for Human Activity Recognition Using Google Kolab. Information 2021, 12, 6.

Rechenthin, Michael David. (2014). Machine-learning classification t echniques for the analysis and prediction of high-frequency stock direction. PhD (Doctor of Philosophy) thesis, University of Iowa. https://doi.org/10.17077/etd.f vui75i2

Sadman, Nafiz. Et al. (2021). Understanding the Pandemic Through Mining Covid News Using Natural Language Processing. 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC). 0362-0367

Srinath, k.r. (2017). Python –The Fastest Growing Programming Language. International Research Journal of Engineering and Technology (IRJET). Vol: 04 Issue: 12. 354-357

Sun, Yuan. Liu, Xuan. Chen, Guangyue. Hao, Yunhong. Zhang, Zuopeng (Justin). (2019). How Mood Affects the Stock Market: Empirical Evidence from Microblogs. Information and Management. Available online 7 August, 103181

Tang, Chang. (2021). Computer-aided Linear Algebra Course on Jupyter-Python Notebook for Engineering Undergraduates. Journal of Physics: Conference Series 1815 (2021) 012004

Triaji, Bagas. Pratomo, Cucut Hariz. DP, Bambang Purnomosidi. (2021). PROGRAMMER'S PERSPECTIVE OF OBJECT-ORIENTED PROGRAMMING (OOP) IN SOFTWARE DEVELOPMENT USING CORRELATION ANALYSIS. SINTECH JOURNAL. Vol. 4, No 1: 79-87

Wang, Fang. Et al. (2020). What Can The News Tell Us About The Environmental Performance Of Tourist Areas. Sustainable Cities and Society. Volume 52, January 2020, 101818

Wu, George Guan-Ru. Hou, Tony Chieh-Tse. Lin, Jin-Lung. (2019). Can Economic News Predict Taiwan Stock Market Returns? Asia Pacific Management Review 24: 54e59

Zeebaree, Subhi R. M. Et al. (2020). Characteristics and Analysis of Hadoop Distributed Systems. TRKU. Volume 62, Issue 04, April, 2020, 1555-1564

Zhao, Lu-Tao & Zeng, Guan-rong. (2019). Analysis of Timeliness of Oil Price News Information Based on SVM, Energy Procedia 158: 4123–4128