International Journal of Engineering

Journal Homepage: www.ije.ir

A Hybrid Approach to Sentiment Analysis of Iranian Stock Market User's Opinions

M. Ahangari, A. Sebti*

Department of Computer Engineering, Faculty of Engineering, Golestan University, Gorgan, Iran

P A P E R I N F O

Paper history:
Received 14 December 2022
Received in revised form 22 January 2023
Accepted 24 Januray 2023

Keywords:
Sentiment Analysis
Opinion Mining
Lexicon Creation
Persian Lexicon
Support Vector Machine

A B S T R A C T

With the significant growth of social media, individuals and organizations are increasingly using public opinion in these media to make their own decisions. The purpose of sentiment analysis is to automatically extract people's sentiments from those social networks. Social networks related to financial markets, including stock markets, have recently attracted the attention of many individuals and organizations. people in these networks share their opinions and ideas about each share in the form of a post or tweet. In fact, sentiment analysis in this field is the assessment of people's attitude towards each share. There are different approaches in sentiment analysis, in this article, a hybrid approach is proposed for sentiment analysis. In this way the feature vector used in machine learning is obtained from a lexicon that is automatically extracted from user's tweets. This lexicon is made by using stock price information related to user's opinion. Also, by using the next day's price information of each share, amendments were suggested to this lexicon. Therefore, the lexicon generated for the feature vector was constructed in three ways, and all three methods reported about an 8% improvement over the baseline method in terms of F-score. The baseline method that is considered for this work, is the Persian version of SentiStrength lexicon which is designed for general purpose.

***doi**: 10.5829/ije.2023.36.03c.18*

NOMENCLATURE

BR	v	P_{Min}	Share minimum price in day
P_{Close}	Share closing price	P_{Max}	Share maximum price in day
$P_{PrevClose}$	Yesterday close price		

1. INTRODUCTION

What others think is always an important piece of information in the decision-making process for most people. The internet and the web now (among other things) make it possible for us to use the opinions and experiences of a wide range of people who are neither our personal acquaintances nor recognized professional critics in our everyday decisions [1]. With the dramatic growth of social media (e.g., reviews, forum discussions, blog and social networks) on the web, individuals and organizations increasingly use public opinion in these media for their decision-making process. However, finding and monitoring tweets on forums and reviewing the information contained in them, still remains a difficult task due to wide variety of sites. A mature reader will have difficulty in identifying relevant sites and accurately summarizing the information and opinions contained in them [2]. Today, customers and business owners use these opinions to identify the strengths and weaknesses of their products, but due to an increase in the volume of tweets, this no longer possible to study them case by case and draw a final conclusion and as a result, the need for a system to acquire knowledge automatically under the title of sentiment analysis (also known as opinion mining) emerged.

The sentiment analysis or opinion mining has many applications, in politics (elections, predicting political developments, the degree of unity of people or society in a case, etc.), social sciences and psychology (analysis of social and cultural issues, analysis of the impact of various events on people's behavior, etc.), management

*Corresponding Author Institutional Email: a.sebti@gu.ac.ir (A.Sebti)

Please cite this article as: M. Ahangari, A. Sebti, A Hybrid Approach to Sentiment Analysis of Iranian Stock Market User's Opinions, ***International Journal of Engineering, Transactions C: Aspects,*** Vol. 36, No. 03, (2023) 573-584

and leadership (assistance in decision-making, awareness of the level of satisfaction of customers or subscribers or a group of contacts, etc.).

So far, a lot of researches has been done in the field of opinion mining and sentiment analysis in English, Chinese and Russian languages. Despite the fact that Persian is the main language of Iran, Afghanistan and Tajikistan, and more than 110 million people around the world speak this language, very little researches have been done on the analysis of sentiments in Persian texts and there are still many complications and challenges in the analysis of sentiments and feelings in Persian language. Today, sentiment analysis can be used in various applications. Part of which includes the recognition of feelings towards specific issues in the field of product review, customer relationship management, financial markets and politics [3].

Financial markets, including stock markets, are markets that have always attracted the attention of many people. As a result, social networks that are related to these markets have been given much attention. In these social networks, users publish their opinions in the form of a post or tweet based on the state of shares or the market as a whole. User's opinions can be a reflection of the Functionality, news or price value of companies and based on the analysis that exists about a share or even they could be without any analysis and support, completely emotional and non-technical. Hence, it can be said that these comments contain positive or negative emotions.

Sentiment analysis in this field is the evaluation of people's attitude towards each share. which can be used for senior managers in various organizations, companies for whom the satisfaction or dissatisfaction of shareholders is important, or to predict the future trend of a stock.

A unique feature of these networks is that the price information of each share is available on the day of comment for that share. Since shares are growing or falling every day. Opinions or comments that shareholders or analysts publish every day are affected by this share growth and decline. In this research, we proved that the sentiment hidden in the comments or tweets posted on the days when the stock is falling or growing are correlated with the price information of that stock. In the proposed system, we used this topic to extract words automatically, to improve the quality of opinion mining in these networks.

In general, there are two main methods of sentiment analysis. Lexicon based method and machine learning based method [4]. Machine learning methods work to predict the polarity of sentiments based on training and testing data sets. And they can use supervised learning methods (they use labeled data for text classification) and unsupervised learning methods (they use raw data for text classification). While the Lexicon based approach does not require initial training to extract data. In these methods, a predefined list of words is used. So that each word is associated with a specific feeling. For our work, we first created a dictionary of the stock market corpus and then using this dictionary along with support vector machine (SVM), we evaluated our approach.

There are three general techniques for building sentimental lexicon. Manual technique, where each word is labeled by an expert according to the polarity of that word, which is time-consuming and costly. Creating emotional lexicon based on a dictionary, such as WordNet [5], which contains synonyms and antonyms of words and creating sentimental lexicon using corpus, this method is usually used to create sentimental lexicon for certain fields. The construction of sentimental lexicon in this way can be created by using the occurrence of words with each other, for example, if a word comes with words with positive polarity, it can be said that word also has positive polarity, or it is possible to match the polarity of the words of an already existing dictionary with that corpus by using a corpus in a certain area.

In our work, we use corpus-based method to build our proposed lexicon. The corpus used in this method is in the field of financial markets, which we extracted them from the , which includes 1,100,000 comments from the shareholders of the stock markets. In the proposed method, words and their polarity are extracted automatically.

In this article, a hybrid method is introduced that uses a lexicon and SVM algorithm to present a new model for sentiment analysis. In this way, the feature vector used in model learning is calculated from the constructed lexicon. In order to categorize the comments of each user, we used a 3-class classification, which includes three classes: positive, negative, and neutral. Also, using the generated dictionary in the feature vector has been done and compared in three ways. The conducted tests show that the use of the proposed method in all three modes has an improvement of about 8% in F-Score compared to the baseline method.

The rest of this article deals with the following topics: In section two, some of the most important researches related to the development of sentimental vocabulary for the Persian language are discussed. In the third section, the process of creating sentimental lexicon and the proposed model for sentiments analysis are described in detail. Evaluation and testing of the proposed model was done in the fourth section. And in the fifth section, we presented our conclusion.

2. RELATED WORKS

The most important and main step in the process of obtaining customer satisfaction is to identify the expectations, demands and possibly the requirements proposed by the consumer. Keeping in mind that not all customers are available in person or people who have

important comments about the product do not share it with the producers, it is possible to use social networks and taking into account the fact that people in this network share their opinions with others, get the information they need without the presence of the customer. Applications of sentiment analysis are examined at three levels: document level, sentence and aspect level [6]. Our work focuses on sentiment prediction at the document level in which we assign a positive, negative or neutral label to each document.

User comments are subjective documents that are rapidly being produced in virtual worlds. Therefore, they have a large volume. And it is not possible to review comments manually. Therefore, we need to use and choose appropriate techniques and methods to check opinions automatically. In general, the available techniques for sentiment analysis are divided into two categories; Techniques based on machine learning and techniques based on lexicon. Until now, most of the studies in the field of sentiment analysis have used machine learning techniques [7, 8]. Most of these works used SVM and Naïve Bayes algorithm for their work. They can use supervised or unsupervised learning methods. Supervised methods use labeled data for text classification, while unsupervised methods only use raw data [9]. Lexicon-based techniques use a dictionary to detect the sentiments of a text and determine the polarity of the given text with statistical calculations on positive and negative words. One of the most important advantages of this technique is that it is fast and does not require training data. And the main disadvantage of this method is its lack of scalability [10]. In the rest of this section, we will discuss some of the presented methods based on machine learning and based on lexicon.

2. 1. Machine Learning Based Method Turney [11] presented a simple unsupervised learning algorithm to classify reviews as recommended (thumbs up) or not recommended (thumbs down). in which the classification of a review is predicted by the average semantic tendency of the phrases in the review, which contain adjectives or adverbs. His proposed algorithm reaches an average accuracy of 74% when evaluated on 410 reviews from the site, opinions, sampling from four different domains (car reviews, banks, movies, and travel destinations). Rani and Kumar [12] conducted sentiment analysis on movie review data, which is in Hindi. They did this using different configurations of Convolutional Neural Network (CNN) configurations. They compared the results given by their Convolutional Neural Network model with the most advanced results of classic machine learning algorithms. The results of their work showed that their proposed model is able to achieve better performance than classical machine learning approaches and reached 95% accuracy. To overcome the shortcomings of current sentiment analysis methods, a sentiment analysis method based on Recurrent Neural Network (RNN) called Bidirectional Long Short-Term Memory (BiLSTM) [13]. After embedding the words using the Word2vec model and calculating the weight of the words using the TF_IDF algorithm, they converted their data, which includes 15,000 hotel review texts, into a weighted vector and then applied it to a BiLSTM network. Their experiments showed that their proposed model has higher accuracy and F1-Score than CNN, RNN, long short-term memory (LSTM) and Naïve Bayes algorithm. The accuracy and F1-Score obtained in their workwere about 91 and 92%, respectively. In a similar work conducted by Rhanoui et al. [14], they used Doc2vec model instead of Word2vec for word embedding, and also used CNN along with LSTM network to extract features as best as possible. They showed with experiments that their proposed method has higher accuracy compared to CNN, LSTM, BiLSTM and CNN-LSTM. The accuracy obtained for their proposed method, on their dataset, which is French articles obtained from national and international newspapers, is around 91%. So far, many works have been done using Support Vector Machine and Naïve Bayes algorithm for sentiment analysis [15-19]. The first work that has been done in the field of sentiment analysis in Persian language is the use of two standard methods of Support Vector Machine and Naïve Bayes in the field of movie review [20]. Also, the characteristics of the presence and Frequency of Unigrams, Bigrams and Trigrams for displaying documents were compared. According to the evaluations they made, they realized that in their work, the SVM algorithm performed better than Naïve Bayes. Also, using the Unigram feature improves the efficiency of the classifier compared to Bigram and Trigram. Also, considering only the presence of a feature has a better result than repeating it. In a similar work conducted by Saraee and Bagheri et al. [21], by examining four different information criteria, including Document Frequency (DF), Term Frequency Variance (TFV), Mutual Information (MI), and Modified Mutual Information (MMI), which were proposed by them, they found that the proposed method has a relatively better performance than the approaches of DF, TFV, and MI. The corrected mutual information can generally reach 85% of the F-Score criterion.

Another version of support vector machine used with particle swarm optimization algorithm for movie review data on Twitter [22]. It shows that using the particle swarm optimization algorithm to determine the parameters of the support vector machine, as well as using the features of n-grams and especially Unigrams can improve the accuracy by 4%. This improvement can be increased to about 2% more by cleaning the data. The first dataset called Pars-ABSA, which is completely based on aspects in Persian language, was presented by Ataei et al. [23]. To test the dataset, 6 models that have recently been used for the sentiment analysis based on aspects, in different fields in English, and their focus is

on deep learning methods, were used. Among the results of all models on their proposed work, the TD-LSTM model has had surprising results, because this model had poorer results compared to other models in English datasets, but in their work [24], it has better results than other models. In another work, two deep learning models (automatic encoders and complex neural networks) were used in the Persian movie review data set. The results obtained from these two models were compared with multilayer perceptron, the results showed that automatic encoders have higher accuracy than multilayer perceptron, and the proposed complex neural network model also performs better than automatic encoders with an accuracy of 82.6%.

Also, in some works, sentiment analysis is used to make predictions for other topics. Derakhshan et al. [25], used sentiment analysis to predict share price movement in stock markets. They proposed a new method that incorporates part-of-speech tags into topic modeling methods and called their method "LDA-POS" method. The average accuracy of the results for this method on quite large datasets in both English and Persian languages reach promising results of 56.24% and 55.33%, respectively, and they outperformed better than the related work that used its English dataset. Also, they produced a dataset for Persian language including five stocks, user opinions and their price movements, which is a valuable resource and they claim that this dataset is the first dataset of Persian stocks which containing quite a protracted time. In a similar work by Li et al. [26], they build a stock prediction system and propose an approach that 1) converts historical prices into technical indicators that summarize aspects of the price information, and models news sentiments by using different sentiment dictionaries and represents textual news articles by sentiment vectors, 2) constructs a two-layer LSTM neural network to learn the sequential information within market snapshots series, 3) constructs a fully connected neural network to make stock predictions. Experiments have been conducted on more than five years of real Hong Kong stock market data using four different sentiment dictionaries. Two baseline models, i.e. MKL and SVM, are employed as benchmarks to compare the performances of their proposed approach. They found from the results that, 1. Based on both information sources, the LSTM outperforms the MKL and the SVM in both prediction accuracy and F1 score. 2. The LSTM incorporating both information sources outperform the models that only use either technical indicators or news sentiments. 3. Among the four sentiment dictionaries, finance domain-specific sentiment dictionary models the new sentiments better, which brings at most 120% prediction performance improvement, compared with the other three dictionaries (at most 50%).

2. 2. Lexicon-Based Methods Most people working in the field of sentiment analysis have focused on machine learning-based methods, and few of them have turned their attention to vocabulary-based methods. Until now, a lot of good lexicon has been produced for English-language works [27-30], but the production of lexicon for the sentiment analysis in Persian language has not been given much attention.

One of the known polarity lexicon for English is SentiWordNet [31], In SentiWordNet, three points are assigned to each of the sets of synonyms in WordNet, which shows how positive, negative and neutral these sets are. This resource contains more than 117,000 sets of synonyms. The main idea of building SentiWordNet is to classify WordNet synonym sets using vocabularies of synonym sets. The important thing to consider about the scores assigned to each set of synonyms is that these scores do not indicate the strength of the polarity. They only indicate how positive, negative or neutral a set of synonyms is. As a result, the sum of these three scores for each set is equal to one. Another vocabulary of general polarity known for the English language is NRC lexicon [32]. This lexicon assigns tags to each word such as emotional, anger, fear, expectation, disgust, . as well as positive and negative tags, which is about 15,000 words and was manually created through Amazon's Mechanical Turk. The first approach of sentiment analysis in Persian [33], which is based on lexicon, was done by providing a framework for sentiment analysis. They introduced 2 sources for sentiment analysis in Persian language: 1. A Persian vocabulary that is related to Persian sentimental words along with its polarity. 2. A dataset that is manually collected and labeled by an expert. They used the Dempster-Shafer theory to determine the polarity of each document. The results of their work show that their proposed method gets a higher F-Score rating of about 90% compared to the methods based on machine learning. In another work by Basiri and Kabiri [34], help to solve the problem of lack of resources for sentiment analysis in Persian language, they present two new resources named SPerSent and CNRC for sentiment analysis in Persian language. SPerSent is a sentence-level dataset where each sentence is associated with two labels, a binary label for determining polarity and a 5-star rating label. CNRC is a Persian lexicon, which was created using the NRC [35] lexicon along with three steps of processing. To evaluate the CNRC lexicon, they compared it with the Persian version of the NRC and Senti_Str [36] lexicons on the SPerSent dataset by the Naïve Bayes machine learning algorithm. The results showed that the CNRC vocabulary has higher efficiency and accuracy than the other two lexicons. In a work similar to the method discussed by Basiri and Kabiri [37], they compared four vocabularies to prove that the direct translation of an English vocabulary into Farsi does not have the right quality in sentiment analysis, which includes the Persian version of Adjectives, CNRC, SentiStrength, NRC. The results showed that the direct translation used in NRC has the weakest performance,

while the pre-processing and lexicon refinement used in SentiStrength and CNRC improved the performance. Also, the results showed that using only adjectives leads to better results compared to using NRC.

Sabeti et al. [38] have proposed a new graph-based method for selecting and expanding seeds to generate general polarity vocabulary, called LexiPers, which includes more than 6000 words.

Amiri et al. [39] presented another method for lexicon-based sentiment analysis. They collected a Persian vocabulary that consists of adjectives, words and expressions considered in two formal and informal categories. Also, the collected vocabulary corresponds to standard Persian or obsolete Persian used by a certain number of native speakers. They also created a web interface that enables native speakers to manually assign a score to vocabulary words. Following the creation of an annotated Persian emotion vocabulary, they designed and developed a 3-language pipeline based on the GATE framework. Its components included a Persian tokenizer, Sentence Splitter, Part of Speech Tags and Gazetteer. As a result, they reported an accuracy of about 65%, which was considered an improvement compared to similar vocabulary-based approaches.

Dehkharghani [40] proposed a new translation-based method for creating polarity vocabulary in languages where there are few lexical resources for sentiment analysis, and applies it to Persian language. Their proposed method is done in four steps, first the words are translated into Farsi, then the translated words are manually labeled as positive, negative and neutral. The next step is feature extraction by English polarity vocabulary, and then, classification is done by Logistic Classifier. This is done by learning the mapping between the inputs described by the extracted features, and the three class labels (positive, negative and neutral). Finally, after the experiments, it was able to reach 92.95% accuracy by considering all four mentioned English vocabularies to extract features.

3. THE PROPOSED SYSTEM

Our work is divided into two phases, the first phase shows the lexicon building process and the second phase shows the process of classification tweets using the generated lexicon.

3. 1. Lexicon Creation As mentioned earlier, due to the importance of financial markets, we automatically created a lexicon for sentiment analysis in the stock market. This work was done in several steps, which are shown in Figure 1. The data or tweets intended for this purpose, and also used for evaluation and testing, were extracted from the site sahamyab.com and stored in our database. Also, we extracted the price value of each share from the official website of Iran Stock Exchange[1] and stored it in a separate table from the database. Table 1 summarized the information of shares stored in the database.

According to the proposed system for lexicon creation in Figure 1, after data collection, pre-processing is done on them. In the following, we will review and explain each of these pre-processing.

3. 1. 1. Remove Unnecessary Tweets Unfortunately, the daily price information of shares is not

TABLE 1. Information of selected shares

Share	Frequency	year
Dey	85461	1392-1399
Fmeli	43054	1392-1399
Folad	35978	1392-1399
Haffari	8917	1392-1399
Hkashti	30907	1392-1399
Khodro	162371	1392-1399
Khsapa	157308	1392-1399
Satran	31383	1392-1399
Shabendar	147451	1392-1399
Shapna	88118	1391-1399
Shatran	22605	1392-1399
Tapico	51276	1392-1399
VTejarat	50341	1392-1399
VBMellat	77265	1392-1399
VBSader	107565	1392-1399

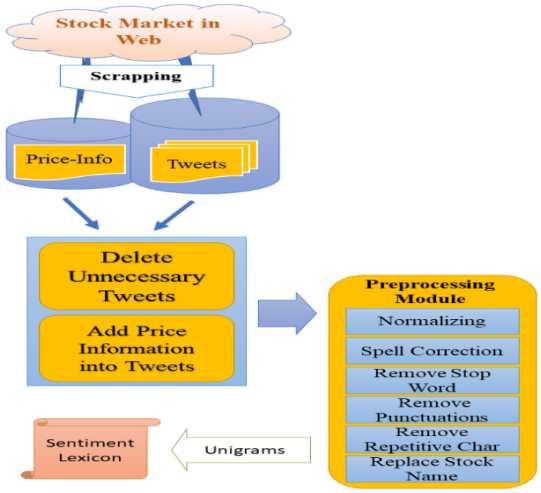

Figure 1. Proposed method for lexicon creation

[1] http://www.tsetmc.com/

available for all the days of posting comments. Since we needed this price information to automatically create the desired lexicon, we had to remove this category of tweets. The number of these tweets decreased to 900,000 tweets after this stage. Some examples of this price information along with time and stock name are given in Table 2.

3. 1. 2. Adding Share Price Information To Tweets

After removing unnecessary tweets, price information was added for each tweet. This information, which part of it is given in Table 2, has been used to determine the weight and polarity of the words of the final lexicon. Table 3 shows some of the tweets along with their price information.

3. 1. 3. Preprocessing One of the important steps in sentiment analysis is data pre-processing, choosing appropriate pre-processing methods can improve the correct classification of data [41]. Therefore, we perform data preprocessing in six steps:

TABLE 2. Some information for each share

Share	Date	Maximum Price	Close price	Yseterday Price
Dey	1399/12/25	18813	19486	19941
Khodro	1399/02/27	587	1174	1174
Haffari	1397/11/29	4121	4072	4116
Hkashti	1393/07/30	5607	5605	5687
Folad	1388/03/04	1884	1865	1900

TABLE 3. Example of Price information for each share

Price Information	Tweets
‘Share symbol’: ‘خودرو’, ‘date’: 1392/08/26, ‘maximum price’: 3150, ‘minimum price’: 3070, ‘close price’: 3090, ‘yesterday price’: 3075, …	#خودرو #حکشتی امروز هر دو صف می‌شوند. خوشبین
‘Share symbol’: ‘شبندر’, ‘date’: 1393/01/26, ‘maximum price’: 2900, ‘minimum price’: 2854, ‘close price’: 2871, ‘yesterday price’: 11566, …	#شبندر سال ۹۴ تا سال ۹۶ شبندر بازم بهتر می‌شود.
‘Share symbol’: ‘خودرو’, ‘date’: 1396/09/27, ‘maximum price’: 3150, ‘minimum price’: 3070, ‘close price’: 3090, ‘yesterday price’: 2850, …	#فولاد سلام خدمت دوستان فولادی با توجه به رشد خوب فولاد که نوش همه سهامدارن آن باشد، متاسفانه واگرائی منفی را نشان می‌دهد. احتیاط بیشتر پیشنهاد می‌شود. الهی همه پر سود باشند.

[1] https://www.sobhe.ir/hazm/
[2] is negative
[3] prices

Step 1: Normalization

For data normalization, we used the HAZM tool[1], which is a library for the Python language. This tool consists of different modules. that we used its normalization module for our work. Some words in Farsi consist of several parts. These words are usually separated from each other by a semi-space. But usually, people do not follow this point in the tweets they post on social networks. For example, the word "آن‌ها" is also incorrectly written as "آن ها". One of the tasks that this module does for us is to convert these types of spaces into thin spaces. Also, some letters in Farsi have different Unicode. In particular, the letters "ی" and "ک" which are sometimes written as "ي" and "ك" respectively. These letters are also converted into their standard form by this module. In fact, in this section, we converted the non-standard words into standard form.

Step 2: Correcting Words

In most informal texts of Persian websites, users write the words as they use them in daily conversations. For this reason, these texts contain a large number of words with non-standard spelling. Therefore, checking the spelling of words in Persian is more challenging than in English [33]. To solve this problem, a list of words along with the number of their occurrences was extracted from the tweets stored in the database. Then the size of this list was reduced in two steps. In the first step, we used the known corpus of Hamshahri [42] in order to match each word from the obtained list with it. In the second step, with the assumption that the words whose frequency is less than ten times in the whole data will have a small effect on the estimation of the sentimental polarity of the sentences, these words were removed from the list, That left 2782 words from the list. Table 4 shows some of the words in this list along with their correct equivalents. Finally, the correct equivalent of each of these words was manually entered and applied to the entire database.

Step 3: Removing Stop Words

To reduce the size of the data and improve the accuracy of sentiment analysis, we removed the stop words that do

TABLE 4. Part of the list of informal Persian words with its formal equivalent

Incorrectly Spelled Words	The Correction Equivalent
منفیه	منفی است[2]
قیمتا	قیمت‌ها[3]
میریزه	می‌ریزد[4]
میگرده	می‌گردد[5]
نشن	نشوند[6]

[4] decant
[5] turn
[6] don’t be

not have a significant impact on sentiment recognition and are frequently repeated in our dataset. For this purpose, unlike other works that use the list of known stop words for recognition, we created a Unigrams of words in the dataset, along with the frequency of each of them, to identify stop words. Assuming that the frequency of stop words in the data set will have a significant difference with other words. With this method, effective words in the dataset, which may be recognized as a stop word by the existing stop word lists, remain in the dataset.

Step 4: Removing Punctuation Marks
All punctuation marks except "#" and "." were removed from the dataset.

Step 5: Removing Duplicate Letters
Sometimes, among tweets, people write some words by repeating one letter to emphasize a topic, for example, they use "نههههههه" to emphasize the word "نه". In this module, the repetition of these letters was removed.

Step 6: Changing The Share Name
Since the name of the share does not affect the process of recognizing sentiments and causes redundancy in our lexicon, in this section we changed the name of shares to a special noun (#سهم), in order to prevent redundancy in the dataset, and to improve the word matching process.

3. 1. 4. Building Lexicon As mentioned in section 1, there are 3 general ways to build lexicon. that our job is to build lexicon based on corpus. The approach of the proposed system to build lexicon is to use n-grams features (especially Unigram and Bigram) in tweet data. Also, price information corresponding to each tweet is calculated to calculate the growth rate of each share at the time of sending tweet. The reason for calculating the growth rate is that we assume that when the growth rate of a stock is positive, people who tweet about that stock in these social networks will use more positive words in their tweets due to their sense of satisfaction. Also, if the share growth rate is negative, more negative words are used in the tweets that these people publish, due to their feeling of discomfort. Based on this, for each tweet, the growth rate was calculated according to the time of posting and which share this tweet is related to. To calculate the growth rate, we used Equation (1).

$$\text{Growth Rate} = \begin{cases} \frac{P_{Close}-P_{Min}}{P_{Min}} & , P_{Close} - P_{PrevClose} > 0 \\ \frac{P_{Close}-P_{Max}}{P_{Max}} & , P_{Close} - P_{PrevClose} < 0 \\ 0 & , P_{Close} - P_{PrevClose} = 0 \end{cases} \quad (1)$$

In our data, the "closing price" is available for each stock on each day. The reason for this is that if a share starts with a high value at the beginning of the trading hours and faces a significant decrease at the end of the trading hours, the comments or tweets that users publish for that share are very important and significant in terms of sentiments .The opposite is also true, i.e. if a share starts trading with a low value at the beginning of the market trading hours, and this value increases during the trading hours, the comments published by users are often associated with positive feelings. Therefore, using the lowest price and the highest price makes a better distinction for calculating the growth rate. Better distinction means that if the mentioned conditions happen, the growth rate for those tweets will be more positive or negative. For this purpose, to calculate the growth rate, we first check whether that share has increased or decreased compared to the previous day, if there is an increase, we calculate the growth rate from the first rule of formula, and if there is a decrease, we use the second rule of the formula.

As mentioned, the extraction of lexicon words was done by the feature of n-grams, and we entered the growth rate and frequency of these words along with it. The calculation of the growth rate of each of these lexicon words is that first, when these words are selected from tweets by the n-gram feature, the growth rate of each of the tweets in which that word is located is considered in a list for that word and after scanning all the tweets, the average of these growth rates for each word is considered as the growth rate of each word. For example, if the word "فروش" is present in 1000 tweets, the average growth rate of these tweets is considered as the growth rate of this word. As mentioned, we used unigram feature as n-gram feature. The details of which are mentioned below.

- **Building ParsStock-v1 Lexicon**

A part of the lexicon created using the Unigram feature is given in Table 5.

As mentioned before, our lexicon is automatically extracted and its size is 11422. After extracting words, these words were sorted and stored based on the absolute value of frequency multiplication in growth. Our purpose of doing this sorting is to show the impact of the growth rate on the feelings of shareholders or people who are active in social media related to the stock market and

TABLE 5. Part of the Lexicon obtained with Unigram feature

Words	Frequency	Mean of Grow Rate (%)
فروش[1]	89179	-0.533
خرید[2]	128991	0.051
خوشبین[3]	161192	0.266
حمایت[4]	30432	-0.604

[1] sale
[2] buy
[3] optimist
[4] support

share their opinions in these networks. Also, this sorting helps to reduce the feature vector dimension and its more effective conclusion, so that the words that have a greater impact on the sentiment analysis process are placed at the top of the lexicon.

After sorting, these words were used to form a feature vector for the comments, and for this we considered 11 different cases, which actually differ in the length of the vector for each comment. These 11 cases include vectors with dimensions of 11000, 10000, 9000, 8000, 7000, 6000, ..., 1000, the length of each vector is based on the number of the first N words of each list, which means that a vector with a length of 11000 is the same as the first 11000 words of the arranged lexicon. Since we used SVM for classification in our work, these feature vectors were used for SVM inputs. To build these feature vectors, we tried 3 different methods and compared the results of each of them. In the following, we will consider each of these methods.

Method 1: Constructing the Feature Vector using the Presence and Absence of Words

In this method, the numbers 1 and 0 are respectively assigned to the presence and absence of tweet's words in the feature vector and then given as input to SVM.

Mehtod 2: Feature Vector Construction using Growth Rate

In this method, for the presence of tweet words in the feature vector, the growth rate of that word is used, and if it is not present, the value is considered zero.

Method 3: Feature Vector Construction using Modified Growth Rate

This method is the same as the previous method, but the amount of growth rate has been modified. The reason for these adjustments is that although some lexicon words express a positive feeling in nature, the growth rate for them has become negative. The reason why the value of the growth rate in these words has become negative is that despite the promising tweets in which these words were seen, they were tweeted on the days in which the growth rate was negative. In order to recognize and correct some of these words in the lexicon and change their growth rate from negative to positive, we used the following method.

By looking carefully at the posted tweets, we will find that some of the comments posted by analysts are actually analysis or predictions of the future state of the share. In other words, the analyst observes the trading status of the stock on the current day, which is a negative day, and sees signs of the stock's return, and as a result, posts a positive tweet about the stock. It should be noted that the mentioned symptoms do not lead to a positive growth rate and are only discovered by a sharp analyst. With this argument, examining the tweets that were posted on negative days with a positive tomorrow can lead us to those words. Therefore, we selected tweets whose growth rate was negative on that day and positive on the next day, and among these selected tweets, we finally selected those whose growth rate changes from negative to positive were significant. In such a way that the growth rate on the day of tweet publication is less than -2 and on the day after the tweet publication, this growth rate is more than 2. After selecting tweets, Bigram words were extracted from these tweets. The reason for this is that we assumed that if the words that are extracted in this condition express a positive feeling, their frequency number will be higher compared to the case where only the growth rate on the day of the tweet is negative and the growth rate on the next day is not important. Therefore, we calculated the frequency of each bigram in the case where the growth rate becomes positive the next day, compared to the case where the growth rate is negative only at the time the tweet was sent and among these bigrams, we selected those whose ratio was greater than 1 and we saved it in a list called GrowRateChangeList.

After this, according to the previous method, we scored each of the feature vector values for each tweet. With the difference that, in addition to calculating the unigram for each tweet and checking its presence or absence in the feature vector, the Bigram values of that tweet are also calculated, and if these Bigrams exist in the GrowRateChangeList, the feature vector values for those words in the tweet, which is the growth rate here, are changed to positive if they are negative. And then they are considered as the input of the SVM.

3. 2. Classification After the comments were placed in the vector space, we used the SVM for classification. Then, all the states obtained by the three mentioned methods were given as input to the SVM to build the model and were evaluated and compared.

Figure 2 shows the proposed model for building the final model.

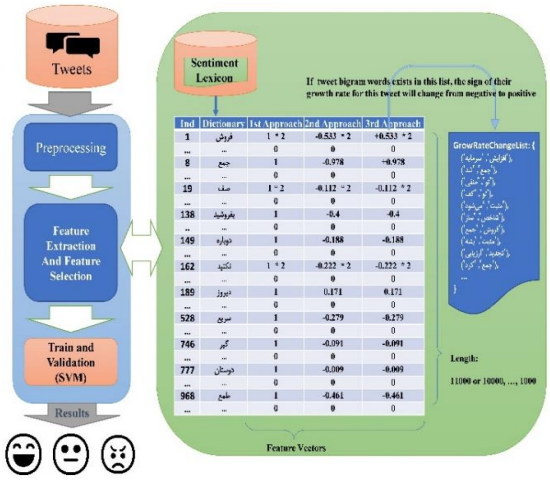

Figure 2. Proposed model for building the final model

4. EVALUATION

As mentioned, using the lexicons we generated, we performed three different methods to construct the feature vector for each tweet, and for each of these methods, 11 different cases are considered, which differ in the length of their feature vector. In order to compare, in our work we calculated all these states for each method and compared their results. Also, in order to show the efficiency of this method and compare it with other lexicons, we used a general lexicon that was used in the work of Basiri et al. [33], to calculate sentiments in our work. And we compared the obtained results with the results obtained from the proposed lexicon**.** For comparison, we randomly selected 1000 tweets from our database, and manually determined the polarity of each of them. In order to select the polarity, we used three classes: positive, negative and neutral. Table 6 shows the characteristics of our dataset.

4. 1. SentiStrength Lexicon An accessible library that is used to detect the polarity and strength of short and informal social texts [43]. Basiri et al. [33] used this library to analyze sentiments on their work. Since this software was designed and created for the English language, Basiri et al. [33] first manually translated the main list of words into Persian. And after removing the repeated words, they used them in their work.

4. 2. Assessment Steps In order to determine the polarity of each tweet in the dataset, we used machine learning methods in our work, and for this purpose, the SVM algorithm was used for classification. In the training and validation phase, 10-fold cross validation was used to make the process independent of the training data. In order to evaluate the classification model, since our data is unbalanced, we used three common evaluation criteria namely Precision, recall, and F1-score. Also we used Python to implement the proposed method. To use SVM, we used the Scikit-Learn library of Python. Scikit-Learn contains the svm library, which contains built-in classes for different SVM algorithms. Since our task is a classification task, we used the support vector classifier class written as SVC in Scikit-Learn's SVM library. This class has one parameter which is the kernel type. In the case of a simple SVM, we simply set this parameter as "linear" because simple SVMs can only classify linearly separable data.

TABLE 6. Dataset specifications

Label	Frequency
Positive	500
Negative	166
Neutral	334

4. 3. Final Evaluation In Tables 7, 8 and 9, the evaluation results for all cases obtained by different methods are summarized. Also, the best result obtained by the feature vector construction method using the modified growth rate is highlighted. As can be seen in Tables 7, 8 and 9, in general, the results obtained from the feature vector construction method using the modified growth rate are better than the results obtained by other methods in terms of the F1-score measure.

This improvement is tangible in two aspects, first in terms of the F1-score measure, in 9 out of 11 different cases, the mentioned method is better than other methods. As well as, the highest F1-score value is when the length of the vector is equal to 6000, which also belongs to the feature vector construction method using the modified growth rate. In Figure 3, a graph is shown

TABLE 7. Results of the evaluation of the proposed lexicons based on feature vector constructed using the presence and absence of words

Cases	Precision	Recall	F-Score
Feature vector length 11000	57.6	56.0	56.5
Feature vector length 10000	57.7	55.9	56.5
Feature vector length 9000	57.6	55.8	56.4
Feature vector length 8000	57.3	55.4	56.1
Feature vector length 7000	57.6	55.8	56.4
Feature vector length 6000	57.3	55.5	56.1
Feature vector length 5000	57.9	56.4	56.8
Feature vector length 4000	57.6	55.7	56.3
Feature vector length 3000	57.3	55.7	56.2
Feature vector length 2000	56.9	55.0	55.6
Feature vector length 1000	56.9	54.0	55.1

TABLE 8. Results of the evaluation of the proposed lexicons based on feature vector constructed using growth rate

Cases	Precision	Recall	F-Score
Feature vector length 11000	56.9	57.7	56.9
Feature vector length 10000	57.1	57.8	57.0
Feature vector length 9000	57.2	57.8	57.1
Feature vector length 8000	56.9	57.5	56.8
Feature vector length 7000	57.4	57.9	57.3
Feature vector length 6000	56.8	57.5	56.8
Feature vector length 5000	57.1	57.7	57.1
Feature vector length 4000	56.5	57.0	56.4
Feature vector length 3000	56.1	56.5	56.0
Feature vector length 2000	56.8	56.9	56.6
Feature vector length 1000	55.4	55.7	55.3

TABLE 9. Results of the evaluation of the proposed lexicons and SentiStrength using modified growth rate

Cases	Precision	Recall	F-Score
Feature vector length 11000	57.5	58.4	57.5
Feature vector length 10000	57.5	58.4	57.6
Feature vector length 9000	57.8	58.6	57.8
Feature vector length 8000	57.6	58.3	57.5
Feature vector length 7000	57.4	58.1	57.4
Feature vector length 6000	**57.8**	**58.6**	**57.9**
Feature vector length 5000	57.8	58.6	57.8
Feature vector length 4000	57.4	58.0	57.4
Feature vector length 3000	56.4	56.8	56.3
Feature vector length 2000	55.0	55.7	55.1
Feature vector length 1000	54.8	54.8	54.6

that draws the F1-score measure for each of the proposed methods.

In Figure 3, the effect of the growth rate in the calculation of emotions as well as the polarity of the sentences was well shown. As can be seen in the figure, correcting the growth rate made the results better in 9 out of 11 cases. Also, in order to compare the lexicon obtained in our work with other works, we used SentiStrength lexicon which is created for general usage and compared the results of this lexicon with our work in Table 10.

In Table 10, the best result of the proposed method is compared with the result obtained from SentiStrength lexicon. According to Table 10, the task of polarity detection using SentiStrength lexicon, which was created for general purpose, has the least accuracy in stock market data. The reason for this low accuracy is that the words in the data of the stock market are specific to the stock market and are not widely used in other fields. Therefore, SentiStrength lexicon among other lexicons obtained the lowest accuracy in all states.

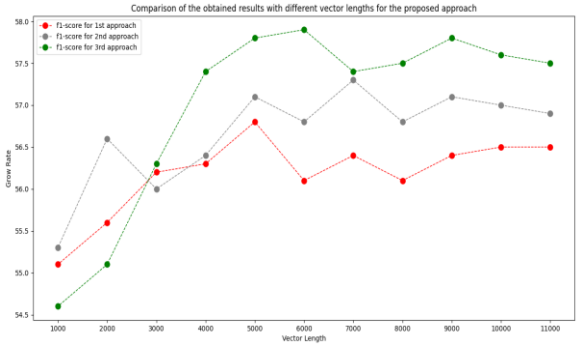

Figure 3. Comparing the results obtained from different cases that exist for the proposed lexicons

TABLE 10. Best result of the evaluation of the proposed lexicons and SentiStrength

Approach	Precision	Recall	F-Score
Feature vector construction using modified growth rate with vector length 6000	**57.8**	**58.6**	**57.9**
Feature vector construction using SentiStrength lexicon	51.2	49	49.8

5. CONCLUSION AND SUGGESTION

The purpose of sentiment analysis is to automatically extract people's opinions on various topics on the web**.** Social networks are an environment where people discuss and exchange opinions every day. One of these networks, which is always the focus of society, is the social networks related to the financial market. The stock market is one of these financial markets where people buy and sell shares. In these networks, people express their opinions about stocks. In our work, we performed opinion mining on tweets that are published daily in the stock market. The unique feature of these networks is the presence of price information for each share every day. We used this to automatically extract lexicon from 900,000 tweets available in these networks. In fact, these features were used to calculate the growth rate of each share at the time of tweeting, and the scores of obtained words for lexicon were determined with the help of these growth rates. To evaluate the proposed method for lexicon generation and feature vector construction based on it, the generated lexicon was compared with the Persian version of SentiStrength lexicon, which is designed for general use, and the results were evaluated. The results showed that their lexicon is less efficient than the lexicon produced by the proposed method in the field of stock exchange. Also, on our dataset, the best F-Score obtained by the proposed method is equal to 57.9%, which is 8% more than the value obtained for SentiStrength lexicon. In addition, after the investigations carried out in the feature selection, we were able to slightly improve the F-Score in our work by creating a new solution to modify the growth rate values that are used to score the feature vectors.

In the price analysis that was performed, the share price analysis of the day of posting tweets and the days before and after that was used. And it was shown that the state of the share in the coming days can help us to identify sentiments. Because many of the comments posted by users, are actually a reflection of the share's status in the upcoming days.

6. REFERENCES

1. Pang, B. and Lee, L., "Opinion mining and sentiment analysis", ***Foundations and Trends in Information Retrieval***, Vol. 2, No. 1-2, (2008), 1-135. doi: 10.1561/9781601981516

2. Liu, B. and Zhang, L., A survey of opinion mining and sentiment analysis, ***Mining Text Data***. 2012, Springer.415-463. doi: 10.1007/978-1-4614-3223-4_13
3. Chaturvedi, I., Cambria, E., Welsch, R.E. and Herrera, F., "Distinguishing between facts and opinions for sentiment analysis: Survey and challenges", ***Information Fusion***, Vol. 44, (2018), 65-77. doi: 10.1016/j.inffus.2017.12. 006
4. Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M., "Lexicon-based methods for sentiment analysis", ***Computational Linguistics***, Vol. 37, No. 2, (2011), 267-307. doi: 10.1162/COLI_a_00049
5. Miller, G.A., "Wordnet: An electronic lexical database", ***MIT press***, (1998). doi: 10.7551/mitpress/7287.001.0001
6. Liu, B., "Sentiment analysis and opinion mining", ***Synthesis Lectures on Human Language Technologies***, Vol. 5, No. 1, (2012), 1-167. doi: 10.1007/978-3-031-02145-9
7. Li, S., "Sentiment classification using subjective and objective views", ***International Journal of Computer Applications***, Vol. 80, No. 7, (2013). doi: 10.5120/13875-1749
8. Jo, Y. and Oh, A.H., "Aspect and sentiment unification model for online review analysis", in Proceedings of the fourth ACM International Conference On Web search and Data Mining. (2011), 815-824. doi: 10.1145/1935826 .1935932
9. Maynard, D. and Funk, A., "Automatic detection of political opinions in tweets", in Extended semantic web conference, Springer. (2011), 88-99. doi: 10.1007/978-3-642-25953-1_8
10. Dashtipour, K., Hussain, A., Zhou, Q., Gelbukh, A., Hawalah, A.Y. and Cambria, E., "Persent: A freely available persian sentiment lexicon", in International conference on brain inspired cognitive systems, Springer. (2016), 310-320. doi: 10.1007/978-3-319-49685-6_28
11. Turney, P.D., "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews", arXiv preprint cs/0212032, (2002).
12. Rani, S. and Kumar, P., "Deep learning based sentiment analysis using convolution neural network", ***Arabian Journal for Science and Engineering***, Vol. 44, No. 4, (2019), 3305-3314. doi: 10.1007/s13369-018-3500-z
13. Xu, G., Meng, Y., Qiu, X., Yu, Z. and Wu, X., "Sentiment analysis of comment texts based on bilstm", ***Ieee Access***, Vol. 7, No., (2019), 51522-51532. doi: 10.1109/ACCESS.2019.2909919
14. Rhanoui, M., Mikram, M., Yousfi, S. and Barzali, S., "A cnn-bilstm model for document-level sentiment analysis", ***Machine Learning and Knowledge Extraction***, Vol. 1, No. 3, (2019), 832-847. doi: 10.3390/make1030048
15. Ahmad, M., Aftab, S. and Ali, I., "Sentiment analysis of tweets using svm", ***International Journal of Computer Applications***, Vol. 177, No. 5, (2017), 25-29. doi: 10.5120/ijca2017915758
16. Ahmad, M., Aftab, S., Bashir, M.S., Hameed, N., Ali, I. and Nawaz, Z., "Svm optimization for sentiment analysis", ***International Journal of Advanced Computer Science and Applications***, Vol. 9, No. 4, (2018). doi: 10.14569/IJACSA.2018.090455
17. Korovkinas, K., Danėnas, P. and Garšva, G., "Svm and k-means hybrid method for textual data sentiment analysis", ***Baltic Journal of Modern Computing***, Vol. 7, No. 1, (2019), 47-60. doi: 10.22364/bjmc.2019.7.1.04
18. Dey, L., Chakraborty, S., Biswas, A., Bose, B. and Tiwari, S., "Sentiment analysis of review datasets using naive bayes and k-nn classifier", arXiv preprint arXiv:1610.09982, (2016).
19. Narayanan, V., Arora, I. and Bhatia, A., "Fast and accurate sentiment classification using an enhanced naive bayes model", in International Conference on Intelligent Data Engineering and Automated Learning, Springer. (2013), 194-201. doi: 10.1007/978-3-642-41278-3_24
20. Hajmohammadi, M.S. and Ibrahim, R., "A svm-based method for sentiment analysis in persian language", in International conference on graphic and image processing (ICGIP 2012), SPIE. Vol. 8768, (2013), 697-701. doi: 10.1117/12.2010940
21. Saraee, M. and Bagheri, A., "Feature selection methods in persian sentiment analysis", in International conference on application of natural language to information systems, Springer. (2013), 303-308. doi: 10.1007/978-3-642-38824-8_29
22. Basari, A.S.H., Hussin, B., Ananta, I.G.P. and Zeniarja, J., "Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization", ***Procedia Engineering***, Vol. 53, (2013), 453-462. doi: 10.1016/j.proeng.2013.02.059
23. Ataei, T.S., Darvishi, K., Javdan, S., Minaei-Bidgoli, B. and Eetemadi, S., "Pars-absa: An aspect-based sentiment analysis dataset for persian", arXiv preprint arXiv:1908.01815, (2019).
24. Dashtipour, K., Gogate, M., Adeel, A., Ieracitano, C., Larijani, H. and Hussain, A., "Exploiting deep learning for persian sentiment analysis", in International conference on brain inspired cognitive systems, Springer. (2018), 597-604. doi: 10.1007/978-3-030-00563-4_58
25. Derakhshan, A. and Beigy, H., "Sentiment analysis on stock social media for stock price movement prediction", ***Engineering Applications of Artificial Intelligence***, Vol. 85, (2019), 569-578. doi: 10.1016/j.engappai.2019.07 .002
26. Li, X., Wu, P. and Wang, W., "Incorporating stock prices and news sentiments for stock market prediction: A case of hong kong", ***Information Processing & Management***, Vol. 57, No. 5, (2020), 102212. doi: 10.1016/j.ipm.2020.102212
27. Cambria, E., Olsher, D. and Rajagopal, D., "Senticnet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis", in Twenty-eighth AAAI conference on artificial intelligence. (2014). doi: 10.1609/aaai.v28i1.8928
28. Hu, M. and Liu, B., "Mining and summarizing customer reviews", in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. (2004), 168-177. doi: 10.1145/1014052.1014073
29. Deng, L. and Wiebe, J., "Mpqa 3.0: An entity/event-level sentiment corpus", in Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: ***Human Language Technologies.*** (2015), 1323-1328. doi: 10.3115/v1/N15-1146
30. Neviarouskaya, A., Prendinger, H. and Ishizuka, M., "Sentiful: A lexicon for sentiment analysis", ***IEEE Transactions on Affective Computing***, Vol. 2, No. 1, (2011), 22-36. doi: 10.1109/T-AFFC.2011.1
31. Esuli, A. and Sebastiani, F., "Sentiwordnet: A publicly available lexical resource for opinion mining", in Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), (2006).
32. Mohammad, S.M. and Turney, P.D., "Crowdsourcing a word–emotion association lexicon", ***Computational Intelligence***, Vol. 29, No. 3, (2013), 436-465. doi: 10.1111/j.1467-8640.2012.00460.x
33. Basiri, M.E., Naghsh-Nilchi, A.R. and Ghassem-Aghaee, N., "A framework for sentiment analysis in persian", ***Open Transactions on Information Processing***, Vol. 1, No. 3, (2014), 1-14. doi: 10.15764/OTIP.2014.03001
34. Basiri, M.E. and Kabiri, A., "Sentence-level sentiment analysis in persian", in 2017 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA), IEEE. (2017), 84-89. doi: 10.1109/PRIA.2017.7983023
35. Mohammad, S.M., Kiritchenko, S. and Zhu, X., "Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets", arXiv preprint arXiv:1308.6242, (2013).

36. Thelwall, M., Buckley, K. and Paltoglou, G., "Sentiment strength detection for the social web", ***Journal of the American Society for Information Science and Technology***, Vol. 63, No. 1, (2012), 163-173. doi: 10.1002/asi.21662

37. Basiri, M.E. and Kabiri, A., "Translation is not enough: Comparing lexicon-based methods for sentiment analysis in persian", in 2017 International Symposium on Computer Science and Software Engineering Conference (CSSE), IEEE. (2017), 36-41. doi: 10.1109/CSICSSE .2017.8320114

38. Sabeti, B., Hosseini, P., Ghassem-Sani, G. and Mirroshandel, S.A., "Lexipers: An ontology based sentiment lexicon for persian", arXiv preprint arXiv:1911.05263, (2019).

39. Amiri, F., Scerri, S. and Khodashahi, M., "Lexicon-based sentiment analysis for persian text, In Proceedings of the International Conference Recent Advances in Natural Language Processing, (2015), 9-16.

40. Dehkharghani, R., "Sentifars: A persian polarity lexicon for sentiment analysis", ***ACM Transactions on Asian and Low-Resource Language Information Processing***, Vol. 19, No. 2, (2019), 1-12. doi: 10.1145/3345627

41. Haddi, E., Liu, X. and Shi, Y., "The role of text pre-processing in sentiment analysis", ***Procedia Computer Science***, Vol. 17, (2013), 26-32. doi: 10.1016/j.procs.2013 .05.005

42. AleAhmad, A., Amiri, H., Darrudi, E., Rahgozar, M. and Oroumchian, F., "Hamshahri: A standard persian text collection", ***Knowledge-Based Systems***, Vol. 22, No. 5, (2009), 382-387. doi: 10.1016/j.knosys.2009.05.002

43. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D. and Kappas, A., "Sentiment strength detection in short informal text", ***Journal of the American Society for Information Science and Technology***, Vol. 61, No. 12, (2010), 2544-2558. doi: 10.1002/asi.21416

Persian Abstract

چکیده

با رشد چشمگیر رسانه‌های اجتماعی، افراد و سازمان‌ها به طور فزاینده ای از افکار عمومی در این رسانه‌ها برای تصمیم گیری خود استفاده می کنند. هدف از تحلیل احساسات استخراج خودکار احساسات افراد از آن شبکه‌های اجتماعی است. شبکه‌های اجتماعی مرتبط با بازارهای مالی، از جمله بازار سهام، اخیرا مورد توجه افراد و سازمان‌های بسیاری قرار گرفته است. افراد در این شبکه‌ها نظرات و ایده‌های خود را در مورد هر سهم در قالب یک پست یا توییت به اشتراک می گذارند. در واقع تحلیل احساسات در این زمینه، ارزیابی نگرش افراد نسبت به هر سهم است. رویکردهای مختلفی در تحلیل احساسات وجود دارد، در این مقاله یک رویکرد ترکیبی برای تجزیه و تحلیل احساسات پیشنهاد شده است. در رویکرد پیشنهادی بردار ویژگی مورد استفاده در یادگیری ماشین از واژگانی که به طور خودکار از توییت‌های کاربر استخراج می شود، به دست می آید. این واژگان با استفاده از اطلاعات قیمت سهام مربوط به نظر کاربر ساخته شده است. همچنین با استفاده از اطلاعات قیمت روز بعد هر سهم، اصلاحاتی در این واژگان پیشنهاد شده است. این واژگان برای بردار ویژگی به سه روش ایجاد شده که در هر سه روش حدود ۸ درصد بهبود نسبت به روش پایه از نظر امتیاز F گزارش شده است. روش پایه ای که برای مقایسات در نظر گرفته شده است، نسخه فارسی واژه نامه SentiStrength است که برای اهداف عمومی طراحی شده است.