Arabic Stock News Sentiments Using the Bidirectional Encoder Representations from Transformers Model

Eman Alasmari[1], Mohamed Hamdy[1,2], Khaled H. Alyoubi[1], and Fahd Saleh Alotaibi[1],

[1] The Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia.
[2] The Faculty of Computer and Information Sciences, Ain Shams University,11566 Abbassia, Cairo, Egypt.

Summary

Stock market news sentiment analysis (SA) aims to identify the attitudes of the news of the stock on the official platforms toward companies' stocks. It supports making the right decision in investing or analysts' evaluation. However, the research on Arabic SA is limited compared to that on English SA due to the complexity and limited corpora of the Arabic language. This paper develops a model of sentiment classification to predict the polarity of Arabic stock news in microblogs. Also, it aims to extract the reasons which lead to polarity categorization as the main economic causes or aspects based on semantic unity. Therefore, this paper presents an Arabic SA approach based on the logistic regression model and the Bidirectional Encoder Representations from Transformers (BERT) model. The proposed model is used to classify articles as positive, negative, or neutral. It was trained on the basis of data collected from an official Saudi stock market article platform that was later preprocessed and labeled. Moreover, the economic reasons for the articles based on semantic unit, divided into seven economic aspects to highlight the polarity of the articles, were investigated. The supervised BERT model obtained 88% article classification accuracy based on SA, and the unsupervised mean Word2Vec encoder obtained 80% economic-aspect clustering accuracy. Predicting polarity classification on the Arabic stock market news and their economic reasons would provide valuable benefits to the stock SA field.

Keywords:
Machine Learning, Deep Learning, Classification, Prediction, Stock Market.

1. Introduction

Stock price decline is a concern for stock market investors and an exciting issue for stock analysts. Stock market news sentiment analysis (SA) allows investors to make the correct investment decision through thorough and thoughtful deliberation. Whereas knowing the best time to buy or sell stocks is the main aim of stock market prediction as there are various factors that may affect stock prices, these factors depend on stock market news indicators, which economists take advantage of to analyze the factors that may affect the market, such as public news sentiment [1]. Undoubtedly, everyone is familiar with sharing news with others in social communities either individually or as a company [2]. Currently, online microblogs' news makes vast amounts of news moods available. At the same time, microblogs may contain large amounts of data about companies that share their information, such as that about their stocks and the stock market, through their microblogs. The articles on these microblogs are used to classify their stocks' orientations and the economic causes of these.

Analysts and investors still have difficulty predicting the future price of a company's stocks. The sentiment of news regarding a company's stocks is a vital interest of financial analysts, investors, and other competitors [3, 4]. SA aims at sentiment recognition and public-opinion checking, which are considered text-mining research fields [5]. Although there are many types of research for analyzing individuals' behaviors in social microblogs, few studies have analyzed the classification of stock news in financial-market microblogs [6]. These microblogs allow investors and analysts to share financial news and opinions with other investors. The SA of microblog posts helps investors perceive the risks related to companies' stocks, which assists them in their investment decision making. Hence, the investor sentiment shown in microblogs has a significant impact on stock prices relative to growth stocks [7]. Therefore, the prediction of stock market behavior depends on the polarity prediction classification of microblogs' news sentiments. Such news can be categorized as positive, negative, or neutral for the polarity prediction of SA, which affects the prediction of stock news classification [3, 4, 8, 9]. Using companies' information from microblogs can therefore improve the accuracy of polarity prediction. Thus, there is a need for SA of companies' stock market articles for stock price prediction.

The previous studies on SA in deep learning (DL) and machine learning (ML) heavily focused on analyzing several features and ignored the sentences' structures and the relations between words in any targeted language, such as the Arabic language [10, 11]. Also, the recent studies have focused on applying DL in SA for multiple tasks based on the Arabic language [11]. However, to the best of the authors' knowledge, few of the previous DL studies were on the Arabic SA in the stock market fields. Moreover, there has been no study that bridges the gaps in Arabic SA by

Manuscript received February 5, 2024
Manuscript revised February 20, 2024
https://doi.org/10.22937/IJCSNS.2024.24.2.13

using DL with MSA features in the Arabic sentiments mirrored in stock news even though DL SA approaches such as the Bidirectional Encoder Representations from Transformers (BERT) model currently offer the best performance for Arabic stock news [12, 13]. Besides, no study on the extraction of the main economic aspects of the reasons for the sentiment polarity depending on Arabic news features has yet been conducted. Thus, there is a need to develop ML and DL models that can classify stock news sentiments and determine their polarity on the basis of labeled data, and that can extract the economic causes of such sentiments. The current study focused on the prediction and classification of Arabic stock market news sentiments and their main economic causes. It sought to develop a supervised model based on logistic regression and the BERT classifier for labeled Arabic data, and an unsupervised model based on k-means clustering.

2. Related Work

2.1 Arabic Sentiment Analysis Techniques for Microblog News

The Arabic SA models and techniques consist of steps such as preprocessing and analysis of linguistic, part-of-speech (POS), semantic, and lexicon-derived features [14]. Preprocessing is often the first step in a text-processing system. It consists of steps for facilitating classification. For instance, in Arabic text preprocessing, the text is tokenized; the letters at the beginning of names (e.g., "الـ") are removed; the letters are normalized, such as by converting "أ" ("Hamza") to "ا" ("Alef"); and then the stop words, such as "كان, إن, في" ("was, that, in"), are removed [15]. Thus, preprocessing cleans the textual data by removing the undesirable elements therefrom to increase the accuracy of the future SA results. Ignoring preprocessing such as spelling corrections will make systems disregard the main words [14, 16, 17], but overdoing preprocessing will make systems lose important data.

Preprocessing of microblogs increases the sentiment prediction accuracy in any field. It considers the variations of letters, such as the basic Arabic letter "Alef" or "أ," which includes the derivative letter "Alef Maksoura" or "ى" often confused with the other basic Arabic letter "Ya" or "ي." Also, there is usually confusion in writing the basic Arabic letters "Ta marbota" or "ة" and "Ha" or "ه." Moreover, the letter "Hamza" is an interchangeable letter based on its word and position, which include "أ", "ؤ", and "ئ." However, additional preprocessing approaches need to be used for some microblog data. These additional techniques involve several tasks, such as the elimination of hashtags, mentions, URLs, and special characters and reposting after the elimination of these elements. Thus, it is essential to handle the Arabic dialects or to provide an MSA alternative [14].

Linguistic-feature analysis, the second step in the Arabic SA models and techniques, is divided into two techniques: analysis of n-grams and analysis of the POS features. N-gram is a chain of n-elements from some textual data, such as words, letters, and syllables. The frequently used n-gram elements are words, which are categorized into three types: unigrams (one word), bigrams (two sequential words), and trigrams (three sequential words) [14, 16, 17]. In Arabic SA, unigrams lead to a high prediction performance, along with the syntactic features. Syntactic features such as word roots, word n-grams, and punctuation impact SA. The variations in the Arabic syntactic features lead to the use of the word roots because although Arabic words have many forms, they all originate from the same root. For instance, all the Arabic words "سلمت", "يسلم", and "سلام" come from the root word "سلم." Thus, there are two configuration settings for extracting the roots of Arabic words: lexeme (LEX) and lemma (LEM). LEX is a configuration setting of all word forms containing the same meaning through tokenization and morphotactics. LEM, on the other hand, is the exact form selected to represent LEX; thus, the Arabic noun is the masculine singular default form but the verb is the third-person masculine singular perfective [14].

POS tagging consists of dividing a word into grammatical categories: nouns, pronouns, verbs, adverbs, adjectives, conjunctions, interjections, and prepositions. The POS tags in Arabic-text analysis include information about the morphology of the word. As for the semantic features, they contain contextual features that point to the semantic orientation of the surrounding textual data. Grammatical features have annotation approaches that add a polarity score to phrases or words by measuring the overall correlation of a multilabel of elements. This semantic orientation is based on various concepts of elements with a given sentiment polarity. Hence, if this element did not appear in the prior dataset with the association of the larger group of elements, polarity can be detected. Some researchers have found that semantic features outperform the unigram and POS-tagging features. Finally, the lexicon-derived features of the Arabic language are lexicons automatically created from some microblogs. However, Arabic lexicons are different from English lexicons as the Arabic language has various dialects and various word forms originating from a single root word [14].

2.2 Targeted News Data

2.2.1. Importance of the news data in microblogs

Microblogs have become essential channels for news dissemination. An increasing number of users are expressing their feelings and sharing other information about the social news in microblogs [18]. Therefore,

microblogs allow investors to share financial news and opinions with other investors. SA of microblog posts helps investors perceive the risks involved in buying a company's stocks and make a decision regarding whether to invest in a company or not to. Hence, stock market microblog news SA has a significant impact on the stock price relative to growth stocks and is associated with stock price movement [7, 10, 16]. The decision making regarding buying/selling orders is based on the market mood determined by technical indicators, which are measurements based on stock prices' time series [19]. Also, unexpected events related to companies affect their performance in a positive or negative direction, such as their stock prices [10]. Consequently, for investment purposes, stock price forecasting has attracted increasing attention in recent years [20], and microblog news data may be among the essential inputs for such forecasting. Thus, stock market prediction values have a significant impact on the financial sector [21].

Stock price prediction is important due to the high volatility of stock market prices [22]. Therefore, to minimize the risk of stock market investment, an accurate stock market price prediction model is needed. The investor sentiment index has been used for stock market prediction in recent years. It is based on stock market news, which some vendors provide as a service. To avoid buying overrated or high-risk stocks, investors decide whether to buy or sell stocks depending on the main stock news. Stock prices and their movements are thus forecasted by analyzing the related news, which outperforms an investor's forecasting of the upcoming short-term trends. Thus, automated decision making supports the prediction of the upcoming stock price trends [10, 23, 24].

2.1.2. Challenge of obtaining news data

The primary challenge faced by researchers with regard to data collection from microblogs is the recent changing of some microblogs' terms of service, disallowing public hosting of old textual data or blogs and public extraction of these [16]. Besides, microblogs have increased the number of orders on API to retrieve specific data, such as on Twitter, as they can no longer directly obtain many data resources. Thus, microblog crawling tools are used to obtain many data resources. There are specific crawling processes: (1) selecting the number of microblogs with particular features, such as language; (2) crawling for a specific type of microblog data and filtering the textual microblog after obtaining each piece of data; (3) archiving the data obtained from crawling on the basis of the ID of each microblog in a specific period; and (4) processing the microblog texts by removing the miscellaneous elements, extracting the themes and feelings, segmenting the words, analyzing the syntax, and processing the text [18]. Accordingly, the Arabic microblogs of some news websites that use a crawler to extract data contain many labels within a specific period [15].

The user-written textual dataset is usually a massive volume of noisy and unstructured data. These data make SA a difficult and challenging task; thus, there is a need to process the unstructured data computationally to extract and determine the sentiments embedded in them [24]. Each language used to write microblog contents has a unique form or word structure [18]. The sentiment of the microblog is affected by the word sequence features, textual-language features, and grammatical-relation features [18]. These challenges can be dealt with through feature filtering (by analyzing some feature sets, such as n-grams) and by replacing certain words and processing the text, which become more significant as the data increase [16].

The processing of textual data involves removing any word that is irrelevant to the text's sentiment. Therefore, the preprocessing of such textual data involves scoring the text's sentiment after excluding the noise therein caused by the aforementioned words. Text processing is done using Python's Natural Language Toolkit on the basis of two methods: tokenization and removing stop words and symbols. Tokenization involves dividing texts by spaces to make a list of individual words per word package. Each word is then used as a feature to train or learn the classifier. Stop words are removed from the list of words because they have neutral meanings and are inappropriate for SA in the targeted language, according to its dictionary. Stop words may include prepositions, symbols such as "@" and URLs, and other words that have no sentiment value [16].

Unigrams are one-word n-grams for each unique tokenized word made for the classifier. For instance, a microblog can be classified according to whether it contains the word "bad" or does not. As this unigram is associated with a negative microblog, the classifier will classify microblogs containing the word "bad" as negative. Bigrams (two-word n-grams) and trigrams (three-word n-grams) can also be classifiers; they classify microblogs on the basis of whether they contain two or three words, respectively, or do not. For example, if a microblog contains the bigram "not bad," which is associated with a positive microblog, it will be classified as a positive microblog [16, 17].

Finally, word replacement involves replacing a company's stock symbols and positive and negative words. N most similar words according to a cosine similarity can be used to replace negative and positive words. Stock symbols can be replaced with a common word and can be removed from the textual blog (e.g., $AAPL can be replaced with "company") [17].

2.3 Sentiment Classification Approaches

2.3.1. Neural network models

The artificial neural network (ANN) models have achieved high performance in Arabic classification and prediction on the basis of the related work chapter. The ANN has neurons in each layer. Each neuron calculates

each input and the sum of the weights, adds the bias, and performs the activation function (e.g., the sigmoid function). The ANN model develops algorithms to solve complex problems such as prediction problems. Neural network models can be generalized by learning from the inputs' relationships to predict the unseen relationships on new data. This enables the model to predict and make generalizations regarding the unseen data. Neural-network prediction approaches perform better with high-volatility data by learning the unobserved relationships in the data without setting any specific relationships in the data. Thus, neural-network models are useful in financial time series prediction, such as in the prediction of stock prices, which contain high-volatility data [25, 26].

2.3.2. Deep-learning model

The DL techniques are used in SA because of their high performance in prediction [27]. They can yield a high result in financial applications. Their impact depends on training complex nonlinear models on the basis of massive datasets [19]. There have been studies on SA approaches based on DL for accuracy improvement, but some DL techniques ignore the words' meanings and order. Therefore, recurrent neural networks (RNNs) such as long short-term memory (LSTM), gated recurrent unit, and BERT are used to train models with solid architectures. The RNNs' architectures extract features in sequential and non-sequential data [28, 29].

The DL approaches obtain the highest results in the binary-data strategies. Thus, the previous studies' continuous-data outcomes showed that RNN and LSTM have the best classification performance [30]. Also, the sentiment classification by Bidirectional Encoder Representations Transformer-Bidirectional models such as BERT and BERT-BiLSTM are superior in accuracy to other classifiers [31]. The BERT model takes advantage of bidirectionality and adds a masked language model (LM) to hide the word predicted and for next-sentence prediction (NSP) [32, 33]. Using such model addresses the limitations of the previous LMs, which work from left to right and do not capture the bidirectional contexts. Also, such model generalizes LM for easy fine tuning for any downstream task. The BERT model mainly performs two steps in its framework: pre-training and fine tuning. Pre-training is done through a couple of unsupervised tasks: masked LM and NSP. The BERT model adds two tokens: Class (CSL) and two Separation Sentences (SEPs) for the input sequences as a separate structure feature and target, to be used in one or two sequences [33].

3. Materials and Methods

The proposed model determines the classification polarity of data and the economic-reason polarity through data gathering, data preparation, and data splitting. Then the supervised and unsupervised models are applied.

3.1. Data Gathering

The data that were used in the present study were collected from the Saudi stock market platform Tadawul, containing the Corporate Articles and Historical Data Stocks datasets. The total dataset covered articles published within nine years (2011–2019). It had 34,386 rows of news and 16 columns of variables: sector, investor name, investor ID, date, time, article title, full article, opening value, highest price, lowest price, closing value, change %, change value, quantity handled value, total price, and total of deals.

3.2.Data Preparation and Annotation

In the proposed approach, there are multiple phases of data preparation: data evaluation, data cleaning, and data validation. Some issues (e.g., noisy, incorrect data and differentiated data formats) are fixed in the data evaluation stage. The data-cleaning stage handles these issues. Noisy and incorrect data are filtered and erased manually and by Pandas DataFrame. Arabic tokenization is performed to fix the differentiated data formats and to standardize all the data sections. In the end, a manual test is performed to validate the whole dataset by testing 30,089 observations.

The polarity annotation adopts the multi-class classification system, which includes the neutral class to increase the model's prediction accuracy [34, 35]. Stock news polarity is annotated manually as negative, positive, or neutral by native Arabic speakers associated with the stock market field.

3.3.Data Splitting

Some data-splitting methods, such as k-fold, grid search, and cross-validation, split the dataset into a couple of sets. These methods are correct but take a long time to train and evaluate each hyperparameter value, which is ineffective in DL models. Therefore, splitting the dataset into three subsets (train-validation-test splitting) increases the training and evaluation speed. Train-validation-test splitting chooses the best hyperparameters by evaluating the validation set outcomes after the training level. Then it re-evaluates the test set outcomes after passing the validation level. Thus, train-validation-test splitting allows selecting the best model, features, and hyperparameters on the validation and test levels. It decreases the mistake cost and performs DL fast for long-time training and evaluation [36].

3.4.The Applied Models

Two models were applied in the current study: the supervised DL model of SA classification and the unsupervised DL model of economic-aspect clustering. Stock news sentiment classification was performed using the BERT model; the articles were classified as positive, negative, or neutral. The unsupervised cluster was improved by performing semantic k-means clustering to categorize the articles into main economic aspects. Thus, these models classify the stock news polarity on the basis of labeled data and extract the economic reason for each stock news article's sentiment.

3.4.1. Supervised model

Two models were applied to the polarity classification, as shown below.

- Baseline model (BL): Logistic regression
- DL model: BERT

3.4.1.1. Logistic regression model

Starting with a simple BL such as the logistic regression model before the complex models helps test the data quality, estimate the primary result, and determine the problem dimensions. The logistic model works effectively with multi-class classification and can be generalized [37]. Also, it performs better with a massive amount of data, such as the dataset used in the present study [38]. The steps below are applied to obtain the best prediction result.

- The data are lemmatized to remove the noisy features.
- Trigrams add the important features.
- The TFIDF vectorizer is used to give the essential words more weight and vice versa.
- The maximum number of features is limited according to the best result.

Weight balancing can be learned using the stochastic gradient descent (SGD) algorithm; thus, the alpha value equals 0.0001 because the TFIDF is the one that controls the importance of features to give proper weights to the important words. The logistic regression model's performance is tested on the basis of the max_features value, by changing the size of the vocabulary. Therefore, many values are tried, and the results decrease at 20,000. Thus, 25,000 is the best choice, with the best result and the fewest maximum features . The maximum number of features is limited using the argument "max_feature" by the fewer features and best result simultaneously. The multi-classes in the training stage are balanced using the argument "class_weight" to prevent the imbalanced-learning effects on the classification performance. Table 1 shows the parameters' values which are stetted to perform logistic regression on multi-class classification, by tuning the best result.

Table 1: The logistic regression model multi-class classification parameters.

Model	***Parameter***	***Value***
Logistic regression	"CountVectorizer ngram_range"	1, 2
	"CountVectorizer max_features"	250,000
	"SGDClassifier __ loss"	"log"
	"SGDClassifier __ n_jobs"	-1
	"SGDClassifier __ class_weight"	"balanced"
	"SGDClassifier __ alpha"	0.00001
	"SGDClassifier __ random_state"	1

3.4.1.2. The Bidirectional Encoder Representations from Transformers model

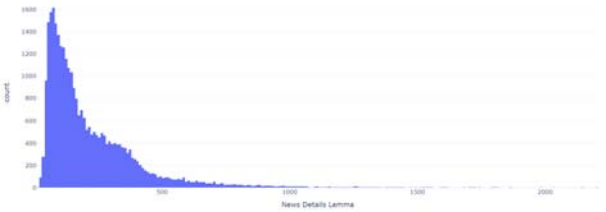
BERT comes up with a general understanding to be used in a downstream task for text classification. It trains a strong model language and then adds an output layer in each downstream task of such language. This output layer is suitable for the downstream task, allowing the fine tuning of this layer and of the other layers [33]. Using the BERT model along with WordPiece embedding will solve this problem. WordPiece takes care of the embedding for the whole token, and the embedding also divides pieces of it. Thus, the model does not discard the newly faced word in the test data but takes the embeddings of the token pieces in the training data [33, 39]. An example is if the model faced the term "going" in the training data. In the previous models, the word "going" is a full token that has only one embedding. Thus, if the previous models faced the new tokens "go" and "ing" in the test data, they would discard such tokens. However, WordPiece gives embeddings for "go" as the first piece and "ing" as the second piece. Thus, with WordPiece, when new tokens like "go" and "ing" are faced in the test data, both will be predictable on the basis of the set embeddings' values, and there is no need to discard them.

Table 2 shows that the fine tuning of the optimal hyperparameter values is task specific. Still, the ranges of possible values for the following are stetted to work well across all tasks on the basis of the BERT study [33]: batch

size, learning rate, and number of epochs, where the best epoch equals 4 in the tests and experiments. The maximum length of documents is chosen on the basis of the quantile value, which equals 0.9; that is, 90% of the documents are less than or equal to 425 words, as the articles' maximum length (see Figure 1). The wall time of training equals 1 minute and 3 seconds.

Table 2: Fine-tuning procedure of the Bidirectional Encoder Representations from Transformers model.

Model	***Parameter***	***Value***
Logistic regression	Batch size	16, 32
	Learning rate (Adam)	5e-5, 3e-5, 2e-5
	Number of epochs	2, 3, 4
	Classifier (Dense)	2,307
	Quantile	0.9
	Best_epoch	4



Fg.1 Maximum length of documents based on the quantile value.

BERT prediction of multi-class classification was applied to compare the results of the validation set with the actual target data. Table 3 shows the evaluation outcomes of the data whose classifications were predicted.

Table 3: Multi-class classification prediction results of the Bidirectional Encoder Representations from Transformers model.

	Negative	***Neutral***	***Positive***
Negative	0.8842	0.0450	0.0707
Neutral	0.0218	0.8667	0.1115
Positive	0.0277	0.0748	0.8975

3.4.2. Unsupervised model

As shown in the previous sections, the BERT model was trained to predict the correct polarity of stock market news articles. However, although this is useful by itself, it misses some parts of the needed picture. The preceding section showed how the polarity of an article is determined but does not elaborate on why the article carries such polarity. For example, the BERT classifier classifies an article as positive, but the reason for its classification of the article as positive is unknown. The reasons for polarity can include the company profits and the fact that the company is to start a new project or production line, increase its capital, or embark on a merger or acquisition. Such assessment is invaluable to the investors, the company owners, and the company's board of directors. The determination of the economic reason for an article's polarity in the literature is called aspect-based SA, which is performed through the procedure described below.

- Annotating each article according to the aspect involved and the underlying sentiment
- Training a supervised ML or DL model on the annotated data
- Using the trained model to predict the aspect and sentiment of new articles

The aforementioned procedure, however, is extremely time consuming and costly. Therefore, the approach proposed herein is a novel and universal approach that significantly reduces both time and cost by reformulating the problem of aspect extraction from a supervised text classification to an unsupervised determination of text similarity, hence eliminating the hand-annotating step from the typical approach described earlier. The proposed approach is performed through several steps. The first step is discovering the overall prevailing categories of aspects in the entire body of articles. Semantic k-means clustering is applied on the most important features per the trigram-TFIDF logistic regression model [40]. Also, instead of performing k-means clustering on the full articles, k-means clustering is limited to the reasons for the negative or positive polarity, as understood by the model. The second step is representing each category of aspects through a set of n-grams carrying the distinct semantics of the category. The third step is encoding the n-grams for each category into an embedding using a sentence-embedding procedure (transformer-based multilingual universal encoder and the mean Word2Vec encoder). Next, the articles are encoded using the same encoder chosen in the third step. Instead of encoding the whole article at once, the article is encoded after paying attention to the relationships between the words and after aspect embedding for improving the goals. Then, for each article, the arc-cosine similarity between the embedded articles is computed, and each of the category embeddings in the third step is computed. Also, a score from 0 to 1 is given for how much each category is related to the article. Finally, the two aspects most associated with the highest scores are picked. Thus, the number 2 is a hyperparameter picked on the basis of reasoning, as will be explained later.

K-means clustering uses the mean Word2Vec vectorizer and needs to choose the number of clusters required by the algorithm for separating the features into. In the present study, after many possibilities were experimented with, ten groups were found to provide good semantic separation between the aspect categories. Hence,

some overall categories of reasons (or aspects) were detected for positive and negative sentiments on the basis of k-means clustering. The following are the ten clusters that resulted from k-means clustering using the mean Word2Vec vectorizer based on the semantic unit.

- Clusters 0, 1, and 3 contain articles about profit/revenue increase/decrease, including stock appreciation/depreciation.
- Cluster 4 contains articles about starting/stopping a project or a production line, or anything related to production in general.
- Cluster 6 contains articles about the increase/decrease of the company's capital.
- Cluster 7 contains articles on dividend distribution or non-distribution.
- Cluster 8 contains articles on insurance operation surplus/shortage.
- Cluster 2 contains articles on the approval of various insurance policies.
- Cluster 9 contains articles on the board of directors' signing of agreements/memoranda of understanding, approval, or disapproval on critical issues or on anything related to managerial matters.
- Cluster 5 has miscellaneous articles that have been not considered.

Most of the clusters exhibited sentimental unity rather than semantic unity, which was the present study's aim for the economic aspects. The Universal Sentence Encoder was also applied to cluster the articles into ten groups. However, unlike the mean Word2Vec encoder, it cannot separate the semantics from the sentiments. The mean Word2Vec encoder is thus the best choice for exhibiting only semantic unity, as required. The central economic aspects manually extracted on the basis of the previous results were profit/revenue, projects/production lines, capital, dividends, insurance operations, insurance approval, and managerial/contracts.

Testing the proposed approach using individual articles showed a large percentage for each aspect based on the article's relevance to each aspect. There is testing for the applied algorithms: k-means clustering using the mean Word2Vec vectorizer and attention algorithm. The attention mechanism enhances the encoding of the article instead of treating the constituent words in the article equally. Attention is a method of selectively encoding an article by assigning higher weights to the words that are semantically similar to each corresponding aspect. However, although the two algorithms' results are similar, the normal algorithm has more sensible results in many examples. Table 4 contains two tested examples for both algorithms. The first article is about insurance and production lines; the normal algorithm accurately predicts its aspects. Also, the attention algorithm is close to the result, but not to the best one. The second article is about profit, insurance approval, and production. Both algorithms accurately predict the aspects, but with some differences in the arrangement. Thus, it is essential to detect the highest aspect or two aspects to limit the probabilities and to validate the method performance. Thus, the proposed approach predicts the highest uni-aspect and bi-aspects besides the seven aspects' percentages of similarity to the article.

The results look promising but not perfect. The algorithm tends to favor the dividends aspect more than the other aspects, although it does this only when the article talks about profit or loss or financial statements as dividend distribution or non-distribution is closely tied to how a company is faring in the financial aspect. Even when the algorithm deviates from the correct aspect, it favors the closely correlated aspects.

Table 4: Word2Vec vectorizer and attention algorithm examples of testing.

Article 1 (Arabic)	***Article 1 (English)***
اشارة للاعلان السابق للشركة المتقدمة للبتروكيماويات المتقدمة المنشور علي موقع تداول بتاريخ 25 اكتوبر 2015م بخصوص اكتمال الاعمال الميكانيكية لمشروعها المشترك الخاص بمصنع انتاج البروبلين بجمهورية كوريا الجنوبية بطاقة تصميمية تبلغ 600 000 طن متري في السنة تود المتقدمة ان تعلن انه بتاريخ اليوم الثلاثاء 15 مارس 2016م بدا مصنع انتاج البروبلين المملوك لشركة اس كي ادفانسد المحدودة اس كي ادفانسد عملية التشغيل التجريبي للانتاج الذي سيخضع لاختبار الاداء طبقا لعقدي ترخيص التقنية مقاول اعمال الهندسة التوريد التشييد الذي قد يستغرق من شهر الي ثلاثة اشهر…	This is a reference to the previous announcement of Advanced Petrochemical Company published on the Tadawul website on October 25, 2015 regarding the completion of the mechanical works for the company's joint project for the propylene production plant in the Republic of South Korea, with a design capacity of 600,000 tons per year. On Tuesday, March 15, 2016, the applicant announced that the propylene production plant is owned by SK Advanced Ltd.; the process of trial operation of the production, which will undergo a performance test according to the technology licensing contracts; the engineering works contractor; and that the construction may take from 1 to 3 months…
Normal Algorithm Result	***Attention Algorithm Result***
1. Insurance approval: 0.8260993022742971 2. Profit: 0.79493785 3. Production: 0.73466593 4. Dividends: 0.70566356 5. Managerial/contracts: 0.6876711 6. Insurance ops: 0.6737051 Capital: 0.6472212	1. Insurance approval: 0.8508865377472663 2. Profit: 0.8629385 3. Dividends: 0.77983284 4. Production: 0.7792212 5. Managerial/contracts: 0.723952 6. Insurance ops: 0.7149271 Capital: 0.67267907
Article 2 (Arabic)	***Article 2 (English)***

تعلن الشركة السعودية لخدمات السيارات المعدات ساسكو انها بتاريخ 13 ديسمبر 2015م قد اكملت توقيع اتفاقية تسهيلات مرابحة متوافقة مع الشريعة الاسلامية مع بنك الخليج الدولي شركة مساهمة بحرينية بقيمة 150 مليون ريال ذلك بضمان سند لامر تتضمن هذه الاتفاقية قرض متوسط الاجل بقيمة 50 مليون ريال بمدة تمويل خمس سنوات منها سنتين فترة سماح علي ان يتم سداده من خلال اقساط ربع سنوية متساوية القيمة بالاضافة الي اصدار خطابات ضمان بقيمة 100 مليون ريال يكمن الهدف من راء هذه الاتفاقية التوسع في مشاريع الشركة دعم انشطتها الرئيسية شراء مواقع جديدة لبناء محطات قود فضلا عن تمويل راس المال العامل…

The Saudi Automotive Equipment Services Company (SASCO) announces that on December 13, 2015, it completed the signing of a Murabaha facility, which is compatible with Islamic Sharia with Gulf International Bank, a Bahraini joint-stock company worth 150 million riyals, by guaranteeing a bond. This agreement includes a medium-term loan of 50 million riyals with a financing period of 5 years, including a 2-year grace period, provided that it is repaid through quarterly installments of equal value. In addition to issuing letters of guarantee amounting to 100 million riyals. The objective behind this agreement is to expand the company's projects, support its main activities, purchase new sites on which to build fuel stations, and finance working capital, ….

Normal Algorithm Result	***Attention Algorithm Result***
7. Profit: 0.7810267	1. Profit: 0.82416815
8. Insurance approval: 0.7601846408453664	2. Insurance ops: 0.85664684
9. Production: 0.74592507	3. Insurance approval: 0.7774521435437018
10. Dividends: 0.70682096	4. Production: 0.7726567
11. Insurance ops: 0.702791	5. Dividends: 0.75157106
12. Managerial/contracts: 0.68651867	6. Managerial/contracts: 0.73063576
13. Capital: 0.6522217	Capital: 0.71698725

4. Results and Discussion

4.1.Polarity Prediction

After the performance of many experiments, the trigram algorithm generated high results but a massive number of features. Therefore, a trigram is used with TFIDF to maintain the performance and limit the number of elements simultaneously. Also, the highest number of features is specified using the argument "max_feature=250000," resulting in fewer features. It chooses features on the basis of their importance, by limiting the features with the best results. Table 5 includes the logistic regression and BERT model performances which are assessed through model testing and evaluation of the classification by the best model. Model testing requires feature and target data. The classification evaluation for article prediction generates multi-class classification assessment for the best model which is shown in Table 6.

Table 5: Logistic regression and Bidirectional Encoder Representations from Transformers model testing results.

Model Name	***Dataset***	***AUC***	***Accuracy***	***Precision***	***Recall***	***F1 Score***
Logistic Regression	Test	0.837542	0.839635	0.837542	0.401299	0.8381
BERT	Test	0.958385	0.877076	0.877822	0.877076	0.877262

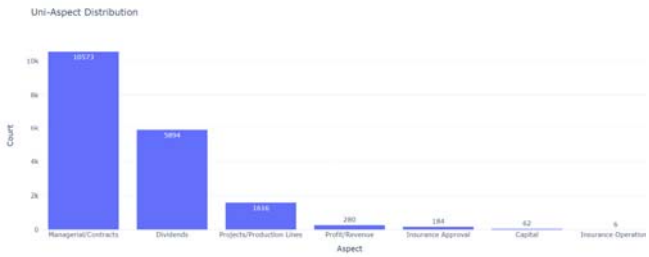
Note: AUC = area under the receiver operating characteristic curve.

Table 6: Model evaluation of multi-class classification.

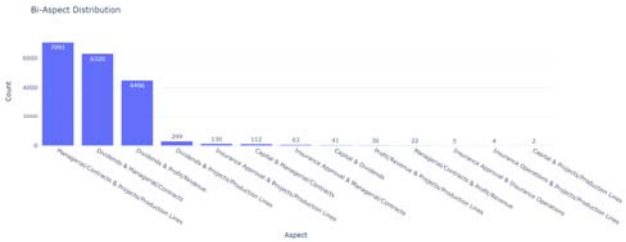
Model Name	***Class***	***Precision***	***Recall***	***F1 Score***	***Support***
BERT	Negative	0.801749	0.884244	0.840979	311
	Neutral	0.884444	0.866725	0.875495	1148
	Positive	0.902724	0.897485	0.900097	1551
	Accuracy	-	-	0.884385	3010
	Macro avg	0.862972	0.882818	0.872190	3010
	Weighted avg	0.885319	0.884385	0.884606	3010

4.2.Scoring of the Economic Aspects

The results show the percentages of similarity and the highest uni-aspect and bi-aspects for each article's seven aspects. The highest uni-aspect is the aspect that has the highest percentage of similarity. Also, the bi-aspects have the two aspects with the highest percentages of similarity to each article. After the model was applied to the data, the following were the highest uni-aspects among all the aspects: profit (16,799 articles), insurance approval (1,570 articles), capital (195 articles), and insurance operations (51 articles) (see Figure 2). Moreover, the highest bi-aspects among all the aspects were insurance approval and profit (8,641 articles), and the lowest was insurance approval and operations (1 article) (see Figure 3). Hence, in each of 2,000 articles, the two highest economic aspects were manually labeled to validate the model's performance. The model was tested on the basis of the labeled data to examine its prediction performance, and its prediction achieved 80% accuracy. The accuracy degree is calculated loosely, such that if the actual class is one of the two predicted classes (bi-aspects), the prediction is accurate. The calculation of accuracy is plausible because the aspects are not mutually exclusive and there is more than one reason in most cases. Therefore, the two highest scores are extracted according to the model, and if the actual aspect is one of the two predicted aspects, the prediction is correct.



Fg.2 Highest uni-aspect scoring.



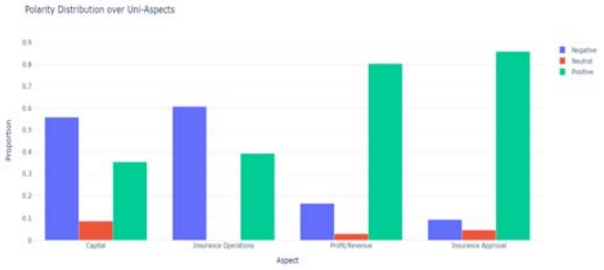
Fg.3 Highest bi-aspect scoring.

4.3.Polarity Distribution over the Economic Aspect

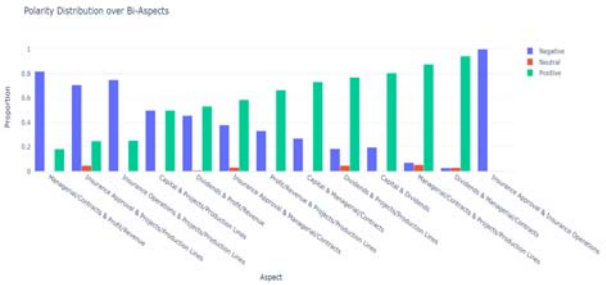
The final results of both models show the articles' polarity prediction and highest economic aspect as the reason for its negativity or positivity. for instance, the first article in the data is manually labeled positive, predicted by BERT as positive. Its predicted percentages of similarity are highest (79%) for the profit aspect and lowest (63%) for the capital aspect. Also, the highest uni-aspect is the profit aspect, and the highest bi-aspect is profit and production lines.

The insurance approval aspect contains the highest positive news because most of its articles are about the approval of companies' insurance applications, which is good news. Also, the profit aspect has a high percentage of positive articles because most of its articles are about companies' profit and revenue news, which are mostly positive. The insurance operations aspect has the highest percentage of negative news because most of its articles are about the occurrence of problems and the need for insurance to limit issues, which is bad news. Also, the capital aspect has a high percentage of negative articles because most of its articles are about capital loss news, which are mostly negative (see Figure 4).

Furthermore, the profit and production lines bi-aspect has the highest percentage of positive news because most of its articles are about companies' profits and revenue from their new projects or production lines, which are good news. The insurance approval and profit bi-aspect also has a high percentage of positive articles because most of its articles are about companies' revenue from the approval of their insurance applications, which are mostly positive. The capital and insurance operations bi-aspect has the highest percentage of negative news because most of its articles are about the occurrence of problems, insurance need, and capital loss, which are mostly negative (see Figure 5).



Fg.4 Uni-aspect polarity distribution.



Fg.5 Bi-aspect polarity distribution.

5. Conclusions

The present study aimed to predict the polarity classification of stock market news and its economic aspect, which affects the polarity. A couple of models were executed to achieve the study aim: the supervised ML model and the unsupervised mean Word2Vec encoder. These models were applied on a prepared and annotated dataset extracted from the Saudi stock market news platform. The supervised ML model includes the logistic regression and BERT models for classifying the sentiments of stock news articles. The unsupervised model, on the other hand, conducts k-means clustering using the universal ML sentence encoder and Word2Vec. The models were applied to textual Arabic stock news by tuning the hyperparameters and features. The logistic regression model was used with lemmatization, trigrams, and the TFIDF vectorizer for the dataset's features to decrease the token numbers. Such model achieved 84% prediction accuracy by setting the maximum features' value (250,000 tokens). The BERT model, on the other hand, achieved 88% prediction accuracy, the highest outcome in a short time for the classification task. The unsupervised model prefers k-means clustering based on Word2Vec to the universal ML sentence encoder to exhibit semantic unity. Unlike the mean Word2Vec encoder, however, k-means clustering using the universal ML sentence encoder cannot separate the semantics from the sentiments. K-means clustering uses the mean Word2Vec vectorizer to choose ten groups that provide the best semantic separation between aspect categories. The manually extracted economic aspects are profit/revenue, projects/production lines, capital, dividends, insurance operations, insurance approval, and managerial/contracts. Moreover, k-means clustering was tested to predict the aspects of the article on the basis of Word2Vec and the attention algorithm performance. Some examples of the results show that the algorithms tend to prefer the dividends aspect to the other aspects. However, when the algorithm deviates from the correct aspect, it predicts correlated aspects. Finally, the mean Word2Vec encoder achieved 80% economic-aspect prediction accuracy. The developed models are valuable resources that classify the Arabic Stock Market news based on their polarity and main economic aspects. The SA and economic aspects extraction models help to perceive the risks based on the news of the stock. They support making the right decision based on the stocks articles sentiments and their main economic reasons of the polarity. Thus, automated decision-making supports predicting the upcoming stock price trends in investing or analysts' evaluation.

Data Availability

Data is available in this link based on the method of data available on the request: www.kaggle.com/dataset/bf24521a3898714597a13efa27a85fa208c96ad49620c787c52720861ddd1e6c.

References

[1] F. Jin, W. Wang, P. Chakraborty et al., "Tracking Multiple Social Media for Stock Market Event Prediction," in *Advances in Data Mining. Applications and Theoretical Aspects*, Cham, pp. 16–30, 2017.

[2] Y. W. Wanjari, V. D. Mohod, D. B. Gaikwad et al., "Automatic news extraction system for Indian online news papers," in *Proc. - 2014 3rd Int. Conf. Reliab. Infocom Technol. Optim. Trends Futur. Dir. ICRITO 2014*, pp. 1–6, 2015.

[3] A. Mukwazvure and K. P. Supreethi, "A hybrid approach to sentiment analysis of news comments," in *2015 4th Int. Conf. Reliab. Infocom Technol. Optim. Trends Futur. Dir. ICRITO 2015*, pp. 1–6, 2015.

[4] V. S. Pagolu, K. N. Reddy, G. Panda et al., "Sentiment analysis of Twitter data for predicting stock market movements," in *Int. Conf. Signal Process. Commun. Power Embed. Syst. SCOPES 2016 - Proc.*, pp. 1345–1350, 2017.

[5] F. Hemmatian and M. K. Sohrabi, "A survey on classification techniques for opinion mining and sentiment analysis," *Artif. Intell. Rev.*, vol. 52, no. 3, pp. 1495–1545, 2019.

[6] E. W. Zhang, W., Li et al., "Dynamic Business Network Analysis for Correlated Stock Price Movement Prediction," *IEEE Intelligent Systems*, vol. 30, no. 2, pp. 26-33, 2015.

[7] D. D. Wu, L. Zheng and D. L. Olson, "A Decision Support Approach for Online Stock Forum Sentiment Analysis," *IEEE transactions on systems, man, and cybernetics: systems*, vol. 44, no. 8, pp. 1077–1087, 2014.

[8] S. Krishnamoorthy, "Sentiment analysis of financial news articles using performance indicators," *Knowl. Inf. Syst.*, vol. 56, no. 2, pp. 373–394, 2018.

[9] P. Choudhari, "Sentiment Analysis and Machine Learning Based Sentiment Classification : A Review," *International Journal of Advanced Research in Computer Science*, vol. 8, no. 3, 2017.

[10] K. Min and H. Moon, "Deep Learning Approach for Short-Term Stock Trends Prediction Based on Two-Stream Gated Recurrent Unit Network," *IEEE Access*, vol. 6, pp. 55392–55404, 2018.

[11] M. Al-Ayyoub, A. Nuseir, K. Alsmearat et al., "Deep

learning for Arabic NLP: A survey," *J. Comput. Sci.*, vol. 26, pp. 522–531, 2018.

[12] Z. Obied, A. Solyman, A. Ullah et al., "BERT Multilingual and Capsule Network for Arabic Sentiment Analysis," in *Proc. 2020 Int. Conf. Comput. Control. Electr. Electron. Eng. ICCCEEE 2020*, pp. 1-6, 2021.

[13] A. Abuzayed and H. Al-Khalifa, "Sarcasm and Sentiment Detection In Arabic Tweets Using BERT-based Models and Data Augmentation," in *Proc. Sixth Arab. Nat. Lang. Process. Work.*, pp. 312–317, 2021.

[14] M. El-Masri, N. Altrabsheh and H. Mansour, "Successes and challenges of Arabic sentiment analysis research: a literature review," *Soc. Netw. Anal. Min.*, vol. 7, no. 1, pp. 1–22, 2017.

[15] M. A. Han, Hao• Hmeidi, I. et al., "A lexicon based approach for classifying Arabic multi-labeled text," *Int. J. Web Inf. Syst.*, vol. 1011, no. 17, pp. 324–342, 2016.

[16] J. Kordonis, S. Symeonidis and A. Arampatzis, "Stock Price Forecasting via Sentiment Analysis on Twitter," in *Proc. 20th Pan-Hellenic Conf. Informatics - PCI '16*, pp. 1–6, 2016.

[17] D. de França Costa and N. F. F. da Silva, "INF-UFG at FiQA 2018 Task 1: predicting sentiments and aspects on financial tweets and news headlines," In *Companion Proceedings of the The Web Conference 2018*, pp. 1967–1971, 2018.

[18] L. Qiu, Q. Lei and Z. Zhang, "Advanced Sentiment Classification of Tibetan Microblogs on Smart Campuses Based on Multi-Feature Fusion," *IEEE Access*, vol. 6, pp. 17896-17904, 2018.

[19] L. Troiano, S. Member, E. M. Villa et al., "Replicating a Trading Strategy by Means of LSTM for Financial Industry Applications," *IEEE Trans. Ind. Informatics*, vol. 14, no. 7, pp. 3226–3234, 2018.

[20] Y. Guo, S. Han, C. Shen et al., "An Adaptive SVR for High-Frequency Stock Price Forecasting," *IEEE Access*, vol. 6, pp. 11397–11404, 2018.

[21] P. Pai, S. Member and C. Liu, "Predicting Vehicle Sales by Sentiment Analysis of Twitter Data and Stock Market Values," *IEEE Access*, vol. 6, pp. 57655–57662, 2018.

[22] F. Z. Xing, E. Cambria and R. E. Welsch, "Intelligent asset allocation via market sentiment views," *IEEE Comput. Intell. Mag.*, vol. 13, no. 4, pp. 25–34, 2018.

[23] Y. Touzani, K. Douzi and F. Khoukhi, "Stock Price Forecasting: New Model for Uptrend Detecting and Downtrend Anticipating Based on Long Short-Term Memory," In *Proceedings of the 2018 2nd International Conference on Cloud and Big Data Computing*, pp. 61–65, 2018.

[24] V. K. Piryani, R., Madhavi, D. et al., "Analytical mapping of opinion mining and sentiment analysis research during 2000 – 2015," *Information Processing & Management*, vol. 53, no. 1. pp. 122-150, 2017.

[25] F. A. Y. Q. Ni, M.ASCE, H. F. Zhou et al., "Generalization Capability of Neural Network Models for Temperature-Frequency Correlation Using Monitoring Data," *J. Struct. Eng.*, vol. 135, no. 10, pp. 1290-1300, 2009.

[26] M. Anthony, and P. L. Bartlett, "Neural network l earning: Theoretical foundations," in *Cambridge: cambridge university press*, vol. 9. 1999.

[27] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: a review," *Artif. Intell. Rev.*, vol. 53, no. 6, pp. 4335–4385, 2020.

[28] S. Sachin, A. Tripathi, N. Mahajan et al., "Sentiment Analysis Using Gated Recurrent Neural Networks," *SN Comput. Sci.*, vol. 1, no. 2, pp. 1–13, 2020.

[29] J. V. Tembhurne and T. Diwan, "Sentiment analysis in textual, visual and multimodal inputs using recurrent neural networks," *Multimed. Tools Appl.*, vol. 80, no. 5, pp. 6871–6910, 2021.

[30] M. Nabipour, P. Nayyeri, H. Jabani et al., "Predicting Stock Market Trends Using Machine Learning and Deep Learning Algorithms Via Continuous and Binary Data; A Comparative Analysis," *IEEE Access*, vol. 8, pp. 150199–150212, 2020.

[31] R. Cai, B. Qin, Y. Chen et al., "Sentiment analysis about investors and consumers in energy market based on BERT-BILSTM," *IEEE Access*, vol. 8, pp. 171408–171415, 2020.

[32] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, pp. 5999–6009, 2017.

[33] J. Devlin, M. W. Chang, K. Lee et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, pp. 4171–4186, 2019.

[34] W. M. Szu, Y. C. Wang and W. R. Yang, "How does investor sentiment affect implied risk-neutral distributions of call and put options?," In *HANDBOOK OF FINANCIAL ECONOMETRICS, MATHEMATICS, STATISTICS, AND MACHINE LEARNING*, vol. 18, no. 2, pp. 1599-1636, 2015.

[35] M. Koppel and J. Schler, "The importance of neutral examples for learning sentiment," *Comput. Intell.*, vol. 22, no. 2, pp. 100–109, 2006.

[36] T. S. Ng, "Machine learning," *Stud. Syst. Decis. Control*, vol. 65, pp. 121–151, 2016.

[37] G. Hackeling, "Mastering Machine Learning with scikit-learn," in *Packt Publishing Ltd*, 2017.

[38] P. Liang and M. I. Jordan, "An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators," in *Proc. 25th Int. Conf. Mach. Learn.*, pp. 584–591, 2008.

[39] F. A. Gers, J. Schmidhuber and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Comput.*, vol. 12, no. 10, pp. 2451–2471, 2000.

[40] I. Goodfellow, Y. Bengio and A. Courville, "deep learning English version," MIT press, p. 800, 2017.