BORSAH: An Arabic Sentiment Financial Tweets Corpus

Mohammed Alshahrani[1,2], Fuxi Zhu[1*], Mohammed Alghaili[3], Eshrag Refaee[4], Mervat Bamiah[5]
Computer School, Wuhan University[1], College of Computer Science and IT, Albaha University [2], Computer School, Hunan University[3], Computer Faculty, Jazan University[4], Faculty of computer science, Prince Sultan University[5], Wuhan, China[1], Albaha, Saudi Arabia[2], Changsha, China [3], Jazan, Saudi Arabia[4], Riyadh, Saudi Arabia[5]
*Corresponding author: fxzhu@whu.edu.cn

Abstract
Impact of social media networks such as Twitter, Facebook, Instagram, etc. on business is vital since people opinions and attitudes may affect the success or failure of a product or a service. This study is a part of continues research project entitled "Evaluating the influence of Twitter on the Saudi Arabian Stock market (TADAWUL)[1] to investigate the impact of Twitter financial tweets on the Saudi Arabia stock market. This paper presented BORSAH an Arabic financial sentiment analysis dataset (corpus) crawled from Twitter. The collected dataset consists of (41,455) Arabic gold-standard annotated Twitter feeds gathered from (118,283) tweets tagged manually from total crawled dataset that consists of (277,453) tweets. The experiment went through three steps, Firstly, we labeled the corpus for Subjectivity and Sentiment Analysis (SSA). Secondly, we applied three machine learning algorithms on part of the corpus. Thirdly, we calculated the accuracy rate of each algorithm. A first sub-corpus will be released via the European Language Resources Association (ELRA) repository with this submission. As far to our knowledge, this is the largest manual annotated Arabic tweets corpus for SSA and the first Arabic financial corpus that will be available for the research community.

Keywords: Arabic Corpus, Sentiment Analysis, Stock Market, Tweets Dataset Analysis, Twitter

1. Introduction

Researchers have been investigating the impact of Twitter on diverse fields such as politics, healthcare, public opinion and stock markets among others. The main challenge was understanding the behavior of users and trends accurately. Several research works were performed to identify users' preferences to predict stock markets prices trends (Fama et al., 1969; Qiu and Song, 2016). This paper presented BORSAH an Arabic financial sentiment analysis dataset crawled from Twitter. It is the Arabic synonym of "Souq Alashom" which means stock exchange market. The word BORSAH has inherited its name from the word "Bourse" that refers to the stock exchange. BORSAH is part of a previous research project for investigating the influence of Twitter on TADAWUL All Shares Index (TASI)[1].

In this study, we used Twitter Application Programming Interface (API) for crawling (277,453) Arabic tweets related to the stock market. We implemented three types of correlations, Pearson correlation coefficient, Kendall rank correlation, and Spearman rank correlation to prove the correlation between Twitter and the Saudi stock market. Furthermore, we considered also the variable mention role for identifying twitter accounts, whose tweets contributed towards market trends. The experiment deployed three machine learning algorithms on a training (test) dataset for crawling (14,000) tweets from the gold-standard annotated feeds. However, during the annotation process, we observed a phenomenon of Twitter selling the Twitter[2] followers tweets to increase the count of followers and illustrate popularity for a specific Twitter account.

This paper is organized as follows: section 2 discusses the related works in Arabic sentiment analysis for the Saudi stock market, and Twitter. Section 3 presents the corpus collection steps including data collection, pre-processing, annotation, and analysis. Section 4 discusses the results. Section 5 views the correlation between BORSAH and TASI, whereby Section 6 describes the release format. Finally, Section 7 concludes the study and the findings.

2. Related Works

Researchers have applied machine learning techniques in Natural Language Processing (NLP) for Subjectivity and Sentiment Analysis (SSA). They annotated their corpora by disregarding the grammar or the lexicon-based methods. Moreover, based on their observations they stated that there is an urgent need to build an Arabic corpus by harvesting different types of Arabic texts on the web to fulfill the demands of scientists in the Arabic NLP research field (Al-Sabbagh and Girju, 2012; Abdul-Mageed and Diab, 2012). Refaee et al. (2014) have crawled and annotated a general Twitter Arabic corpus for SSA. Duwairi et al. (2014) stated that sentiment analysis was implemented in financial industry to illustrates the sentiment data of targeted companies which enables better decision in real time.

AL-Rubaiee et al. (2015) have classified Arabic text in stock trading using Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Naive Bayes algorithms. They presented the relations between Arabic tweets and stock market movements based on observation of both stock market closing price and daily sentiment tweets graph. However, their dataset was small, and they did not prove the correlation.

In their extended research AL-Rubaiee et al. (2016) studied MUBASHER the leading stock analysis software provider in Gulf Cooperation Council (GCC) region using Twitter for opinions mining purposes. They extracted feedbacks from MUBASHER by designing a model for the Saudi Arabic tweets sentiment analysis. Their model

[1]https://www.tadawul.com.sa/wps/portal/tadawul/markets/equities/indices/today

[2]https://www.theatlantic.com/technology/archive/2018/01/all-followers-are-fake-followers/551789/

combined machine learning and NLP approaches for classifying the Arabic tweets into sentiment polarity classes: Positive, Neutral and Negative.
This research aims to mitigate the gaps in related works at the Arabic NLP community regarding finance by proving the correlation between Twitter and TASI as continuous work of a previous project. Moreover, this paper provides the largest gold annotated Arabic manual available that contains (41,455) Arabic tweets corpus for SSA, also it provides the first Arabic tweets dataset for SSA experts in the stock market for mining other objects of tweets such as Mention, Retweets, Following Count, Followers Count and its impact on stock market trends.

3. Twitter Financial Corpus

Twitter API[3] is used by developers and researchers for retrieving or modifying data. The Twitter Search API provides relevant results to ad-hoc user queries from a limited corpus or recent tweets. The Representational State Transfer (REST) API allows access to Twitter texts for reading the timeline, tweeting and following. A study conducted by Alsing and Bahceci (2015) stated that there are several issues when using Twitter Search API including the complexity of restricted tweets quires and the availability of data that cannot be older than seven days. Since the Search API only uses indices which contains most recent or popular tweets. Another issue is that Twitter Search API is used for relevance and not for completeness which results in some missing tweets in the query results.

The Twitter Search API Developers Page[4] states that Streaming API is more suitable for completeness-oriented queries. However, this is not the case of gathering data for sentiment analysis where high completeness is required to analyze the whole data. Twitter provides various types of streaming endpoints, each type is customized to a specific use such as a) Public streaming, b) User streaming and c) Site streaming. Public streaming refers to long-lived Hypertext Transfer Protocol (HTTP) and parsing requests incrementally. The streaming API allows requests for various parameters including language, location, follows, track, count, and delimited to define what data is supposed to be returned. However, the request is different for each language or framework based on HTTP library.

The used Twitter API supports UTF-8 for the Arabic language that may cause characters counting problems for the Arabic dialectal tweets. The main drawback of Twitter API is that a user may crawl tweets based on a country code or enters a location longitude and latitude using API search. Unfortunately, API streaming does not provide this option. Moreover, Twitter does not provide country code location of the user unless he activates the location tracking.

This study aims to identify the minority of users who have activated their locations, also to mitigate the geographical attribute challenge for API streaming, since Saudi Arabia geographical shape is semi-rectangular from north to south. The tweets were crawled within a radius of Saudi Arabia from central point covering the whole country, also the radius distance include some countries such as Jordan, Iraq, Egypt, Sudan, GCC, and Iran.

3.1 Data Collection

In this study several Twitter API's were used in parallel to crawl Arabic tweets regarding TADAWUL. However, only (10) keywords were accepted for each query, due to the limitation of crawlers which may lead to missing tweets about TADAWUL that do not contain the designed keywords. A set of search queries were generated to enhance the possibility of acquiring tweets which convey emotions, attitudes, and opinions towards a specified entity. Table 1 shows the keywords used to retrieve the Arabic financial tweets. Crawling live stream tweets was conducted between (27th August to 23rd December 2015), with total of (277,453) collected tweets.

Arabic	Transliteration	Translation
السوق	Assouq	Market
تاسي	TASI	TASI
تداول	TADAWUL	TADAWUL
الأسهم	Al-as-home	Shares
السوق السعودي	Assouq Assaoodi	Saudi Market
سوق الأسهم	Sooq Al-as-home	Stock Market
سعر السوق	Se'er Al-sooq	Market Price
سعر الإغلاق	Se'er Al-eGh-laaq	Closing Price
ارتفاع	Ertifaa	Growth
هبوط	Hoboot	Fall

Table 1: Arabic financial related tweets keywords

3.2 Data Pre-processing

The tweets were stored in MongoDB[5] which contains the objects of each tweet. During the crawling process, several filters were added to a) block spam, and b) to blacklist spammers' accounts, also c) to remove duplicate tweets from the same IDs, as well as e) to remove the tweets that contain long words. These steps were conducted to ensure that the source was safe from any automatic spam tweets. We created a blacklist for spammers IDs and keywords of their spamming tweets from the initial test. Moreover, (5000) tweets were collected and manually classified their contents based on the same keywords to detect random spammers IDs and spamming keywords, to reduce spam tweets in the main intended crawling process.

Figure 1. illustrates the process of crawling, filtering, inserting, tagging and exporting Arabic tweets regarding TADAWUL. The tweets contained public objects such as User ID, Screen Name, Name, created at, Text, Mention, Retweets, Following Count, Followers Count. Moreover, Figure 1. presents the platform built for annotators to tag each tweet based on its polarity, region, spam, and non-related meaning. It illustrates the mechanism for retrieving the tweets from the database for simple statistical analysis, and easier manual annotation process, whereby the annotator clicks on one of the annotation icons colored Green to choose his/ her preference, either Green, Red, or

[3] https://dev.twitter.com/rest/public/search
[4]https://developer.twitter.com/en/docs/tweets/filterrealtime/guides/connecting.html
[5] https://www.mongodb.com/

Orange. The Green is positive, Red is negative, and Orange is neutral. Furthermore, the "No go icon" indicates it is spam. The "3-connected-dots icon" represents non-related tweets that do not refer to the stock market. The "location icon" refers to stock market tweets related to another country besides Saudi Arabia.

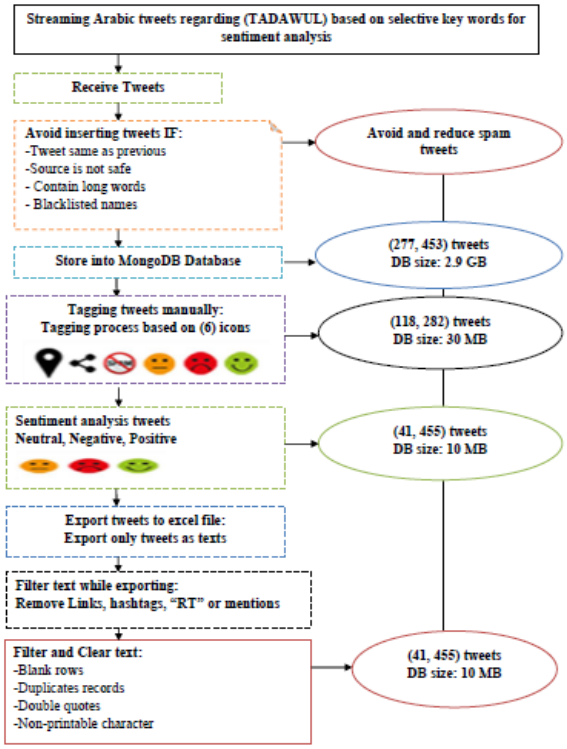

Figure 1: Arabic tweets streaming API.

3.3 Data Annotation

Table 2 illustrates the tweets statistics. Total of manual tweets tagged were (118,282), and total of gold-standard (41,455) tweets were in the dataset based on their sentiment polarity. Moreover, total of (3,756) spam tweets were tagged manually including spamming IDs and keywords lists, this process prevented thousands of spammers from storing their tweets in our database. Furthermore, (58,840) tweets did not have any related meaning even it contained the research keywords due to Arabic multiple meanings of the same word.

Type of Tweets	No. of Tweets
Total tweets crawled	277,453
Total tweets tagged manually	118,282
Positive tweets	5,449
Negative tweets	9,469
Neutral tweets	26,537
Spam	3,756
Non-related meaning tweets	58,840
Non-Saudi Stock tweets	14,231

Table 2: Tweets Statistics

However, a total of (14,231) tweets were excluded from the dataset due to their irrelevant content to stock market as they only contained some stock-market-related keywords. The annotated set of (14,231) tweets were excluded since they were discussing other major stock markets (e.g. GCC) with no mention of the Saudi stock market.

The first round of tagging process was carried out by finance graduate researcher as the tweets contained finance related vocabulary and daily stock market information. Afterward, a Saudi linguistic expert double checked the tagged tweets. When there was a conflict between the two annotators about such tweet, a third assessor had evaluated the tweet to which class should be assigned to. Moreover, since the tagging process is time-consuming we divided the dataset into two parts. The first dataset consists of (118,282) tweets which were tagged and analyzed in this paper. The rest will be released in the future after completing manual tagging and analysis.

Identifying the related tweets was costly in terms of time and efforts as it needed several months to tag the collected tweets manually since the required data has specific nature regarding the stock market in Saudi Arabia. Twitter has inserted enormous tweets from GCC countries, it was extremely tough in some cases to identify the tweet that refers to which stock market in the region Saudi or GCC. Additionally, the same keywords of the stock market can be used in diverse fields not only stock market.

We observed a phenomenon that anonymous users are selling the Twitter[6] followers to increase the number of followers and illustrate popularity for a Twitter account. This observed phenomenon prevented us to from storing such tweets by false popular accounts that may lead to a false analysis of this dataset. For example, a Twitter account following 180k accounts and 190k followers most probably this Twitter account user does not read the tweets of those accounts he follows due to the huge number of tweets that are generated hourly by the users. Additionally, to reduce the impact of such cases, we set a threshold of 5k accounts that each user can follow. A total of (38,432) tweets or IDs were excluded into a separate spam dataset. The Main attributes of the dataset are defined in Table 3.

Attribute Name	Attribute Description
User_ID	Unique ID assigned by Twitter to each user
Created_at	Date and time on which tweet as posted.
Screen_name	Account name displayed on twitter for each user
Followers_count	Number of other users following this account
Retweets_count	Total No of retweets posted for this tweet
Mentions	Total No of mentions of this tweet
Following_count	No of twitter accounts a user is following
Status	The polarity of the tweet text. It mainly consists of three emotions, negative, positive and neutral
Text	Arabic text 140 letters

Table 3: Main Attributes of Data Set

[6] https://www.theatlantic.com/technology/archive/2018/01/all-followers-are-fake-followers/551789/

Figure 3. illustrates the Arabic tweets daily volume and amount of positive, neutral and negative tweets. These tweets were gathered from (27th August to 20th October 2015). Negative tweets are presented in red, Positive tweets in green and daily volume in blue color.

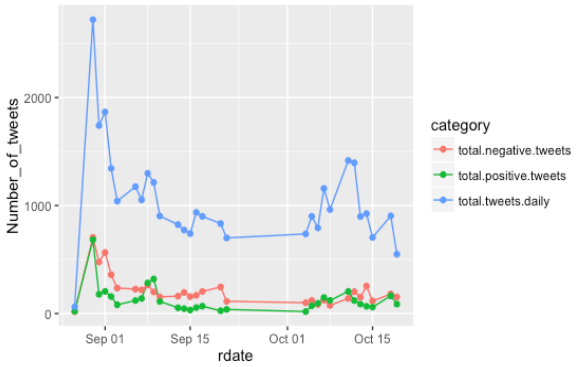

Figure 3: Volume of Daily Tweets

3.4 Data Analysais

A machine learning system was built for Arabic sentiment analysis. The used algorithms to explore the polarity of a given data were Naive-Bayes, SVM, and KNN, which provide the best accuracy in sentiment analysis for Arabic text in relation to the Saudi stock market. First, part of the gold-standard corpus of (14,000) tweets were used as a training dataset. After that, each one of the three machine learning algorithms was applied to the training dataset. Finally, the experiment was processed on (625) tweets test dataset to evaluate the accuracy rate of each one of the algorithms. The implementation of Naïve Bayes was divided into three stages.

1- *First stage:* representing sentences by a vector of words called "victory" for each feature in the dataset. Victory refers to a large array that contains each word with its frequency in the dataset denoted by $|VOC|$.

2- *Second stage:* learning the dataset by training the text test to estimate the probability of each word (w_k) in the text test with each feature in the dataset. To perform this a calculation process must be performed as follows:

- Calculating the average of each feature (F_j) in the dataset by dividing the number of each feature by the whole number all features in the dataset.
- Calculating the probability of each word (w_k) within each feature (F_j) in the dataset separately. By the following equation:

$$P(w_k|F_j) = \frac{n_k+1}{(n+|VOC|)} \quad (1)$$

Whereby n is the number of words in the a given feature F_j in the dataset. n_k is the number of times the word k occurs in each featureF_j in the dataset.

3- Third stage: Training the classifier by calculating the value of Naïve Bayes (V_{NB}) using the following equation:

$$V_{NB} = argmaxP(F_j) \prod_{w_k \in W} \quad P(w_k | F_j) \quad (2)$$

KNN is the second machine learning algorithm that has been used for:

- Calculating the Euclidean distance between vectors and the dataset sentence.
- Sorting the distances in decreasing order.
- Taking the k nearest distances to the sentence that needs to be classified.
- Finding the majority class among the selected k distances. the majority class will be the chosen prediction for the given sentences.

Assuming k = 3 (where k is the closest distance to the given sentence).

The last machine learning algorithm used in the proposed system is SVM has the common library libsvm as most researchers are using it.

4. Results

We have examined the three algorithms with text test to find out the accuracy rate. The performance metrics that were widely used to evaluate the classification results were precision and recall (Khan et al., 2014). The results were summarized in Table 4. highlights the number of tweets annotated automatically.

Algorithm	No. of Test Sentences	No. of Positive Sentences	No. of Negative Sentences	No. of Neutral Sentences	No. of Errors
Naive Bayes	625	35	42	548	155
KNN	625	67	165	393	97
SVM	625	90	142	393	58

Table 4: Numbers of tweets annotated automatically

Algorithm	TP	FP	TN	FN	TP rate (Recall)	FP rate	Kappa	Accuracy rate	Precision
Naive Bayes	35	40	590	16	0.686275	0.011091	78.3077%	75.2%	0.466667
KNN	67	31	558	34	0.663366	0.052632	87.074%	84.48	0.683673
SVM	90	6	535	37	0.708661	0.063492	91.1878%	90.72	0.9375

Table 5: Statistical significance test for each algorithm

Moreover, statistical significance test was conducted to validate comparison of results. From the analysis, SVM has the highest accuracy rate with (90.72%) while Naive Bayes is the lowest with (75.2%).

This study is based on Cohen's methods (Cohen, 1960) which measure the degree of agreements among the assigned labels correcting for agreement by chance. We have tested (625) tweets that are not included in the training dataset to calculate the performance of each algorithm. All these test tweets annotated manually before testing them. We found out that the number of errors using Naive Bayes is larger than the number of errors in KNN and SVM. Table 5 shows the statistical significance test for each algorithm.

5. BORSAH and TASI Correlation

This study uses TADAWUL for finding the Saudi stock market trends. We collected the closing prices data for TASI through TADAWUL from (27th August to 20th October 2015). This data proved the correlations between the influential users, tweets and the stock market prices. Figure 4 shows the distribution of TASI performance, positive and negative tweets at the same time.

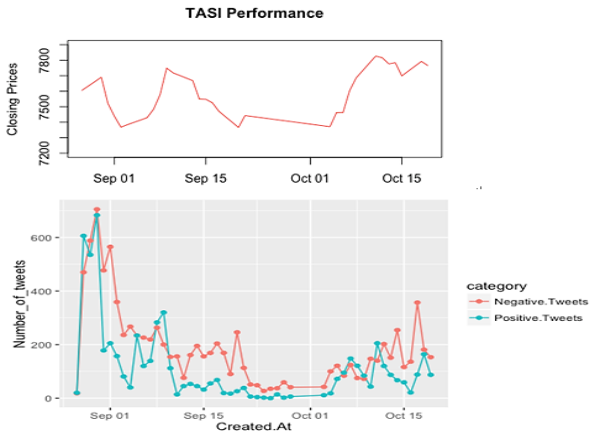

Figure 4 : Distribution of TASI performance

We illustrated in our previous work that three types of correlations were implemented, Pearson's correlation coefficient, Kendall rank correlation, and Spearman's rank correlation. Furthermore, we presented how the variable mention plays an important role in identifying twitter accounts, whose tweets contributed towards market trends. We emphasized that most influential users can be predicted in the future, who may have a significant impact on the stock market trends based on studying the followers count variable. In this paper, we observed a relationship between the daily volume of tweets and stock market index. The daily volume of tweets had increased and decreased in real time parallel with the stock market indicator during rising and falling phases. Figure 5. shows the distribution of TASI performance and daily volume at the same time.

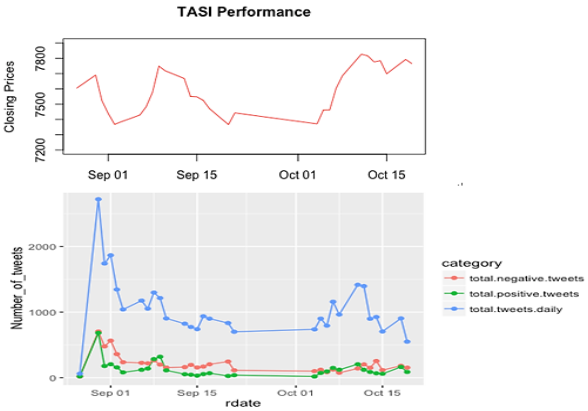

Figure 5: The daily distribution of TASI performance.

6. Release Format

The experiment consists of two datasets, the gold-standard annotated Arabic twitter feeds that consist of (41,455) tweets, and (118,283) manually tagged tweets dataset. Those two datasets will be released via the ELRA data repository which is saved in Comma Separated Values (CSV) file format. However, we removed Tweets text from the dataset due to Twitter privacy restrictions, our future submission will contain the text as encrypted. This is the first subset release among several planned releases. The aim of this study is to provide an annotated Financial Arabic Twitter dataset for the research community to investigate the influence of Twitter on the stock market.

7. Conclusion and Future Works

This paper presented BORSAH as an Arabic financial sentiment analysis dataset crawled from Twitter. We illustrated a method for harvesting Twitter Arabic finance tweets and presented the annotation process, we studied the properties and the statistics of the corpus. Moreover, we applied an Arabic text classification in finance through different algorithms such as SVM, KNN, and Naive Bayes. The SVM algorithm showed the best result.

This corpus is part of a project evaluating the influence of Twitter on the Saudi Arabian Stock market indicators TASI, whereby the correlation between TASI and Twitter was previously proved using three types of coefficients correlations. Moreover, we extended current corpus by annotating manually and automatically extra (159,171) tweets. A first sub-corpus will be released via the ELRA repository that focused only on TADAWUL Arabic tweets.

In future, we aim to investigate an observation we obtained about specific group users who tend to post many tweets targeting or mentioning specific companies. Thus, more investigation is needed to assess the impact of varying the threshold of number of accounts followed by a specific user, i.e. rather than the 5k threshold used in this study.

8. Acknowledgment

This research is supported by The National Natural Science Foundation of China with Grant No: 61272277.

9. Bibliographical References

Abdul-Mageed, M. and Diab, M. (2012). Awatif: A Multi-Genre Corpus of Modern Standard Arabic Subjectivity and Sentiment Analysis. *In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

AL-Rubaiee, H., Renxi, Q. and Dayou L. (2016). Identifying Mubasher Software Products through Sentiment Analysis of Arabic Tweets. *Industrial Informatics and Computer Systems (CIICS), 2016 International Conference on* IEEE. DOI: 10.1109/ICCSII.2016.7462396.

AL-Rubaiee, H., Renxi, Q. and Dayou L. (2015). Analysis of the Relationship Between Saudi Twitter Posts and the Saudi Stock Market. *Intelligent Computing and Information Systems (ICICIS), 2015 IEEE Seventh International Conference on.* IEEE. DOI: 10.1109/IntelCIS.2015.7397193

Al-Sabbagh, R. and Girju, R. (2012). Yadac: Yet another dialectal Arabic corpus. *In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Alsing, O. and Bahceci, O. (2015). Stock Market Prediction using Social Media Analysis. *Degree Project, In Computer Science*, First Level Stockholm, Sweden.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Duwairi, RM. (2015). Sentiment Analysis for Dialectical Arabic. *6th international conference on Information and communication systems(ICICS)*. IEEE, pp. 166–170. DOI: 10.1109/IACS.2015.7103221.

Fama, E., Fisher, F., Jensen, M., and Roll, R. (1969). The Adjustment of Stock Prices to New Information. *International Economic Review*. 10(1):1–21. DOI: 10.2307/2525569.

Qiu, M. and Song, Y. (2016). Predicting the Direction of Stock Market Index Movement Using an Optimized Artificial Neural Network Model. *PLoS ONE*, 11(5): 0155133. https://doi.org/10.1371/journal.pone.0155133.

Refaee, E. and Rieser, V. (2014). An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis. *In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).