Stock News Dataset Quality Assessment by Evaluating the Data Distribution and the Sentiment Prediction

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

ealasmari0010@stu.kau.edu.sa m.hamdy@cis.asu.edu.eg kalyoubi@kau.edu.sa fsalotaibi@kau.edu.sa

[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.

Summery

This work provides a reliable and classified stocks dataset merged with Saudi stock news. This dataset allows researchers to analyze and better understand the realities, impacts, and relationships between stock news and stock fluctuations. The data were collected from the Saudi stock market via the Corporate News (CN) and Historical Data Stocks (HDS) datasets. As their names suggest, CN contains news, and HDS provides information concerning how stock values change over time. Both datasets cover the period from 2011 to 2019, have 30,098 rows, and have 16 variables—four of which they share and 12 of which differ. Therefore, the combined dataset presented here includes 30,098 published news pieces and information about stock fluctuations across nine years. Stock news polarity has been interpreted in various ways by native Arabic speakers associated with the stock domain. Therefore, this polarity was categorized manually based on Arabic semantics. As the Saudi stock market massively contributes to the international economy, this dataset is essential for stock investors and analyzers. The dataset has been prepared for educational and scientific purposes, motivated by the scarcity of data describing the impact of Saudi stock news on stock activities. It will, therefore, be useful across many sectors, including stock market analytics, data mining, statistics, machine learning, and deep learning. The data evaluation is applied by testing the data distribution of the categories and the sentiment prediction—the data distribution over classes and sentiment prediction accuracy. The results show that the data distribution of the polarity over sectors is considered a balanced distribution. The NB model is developed to evaluate the data quality based on sentiment classification, proving the data reliability by achieving 68% accuracy. So, the data evaluation results ensure dataset reliability, readiness, and high quality for any usage.

Key words:
Stock Dataset, Stock Market News, News Impact, Stock Activities, Data Quality Assessment.

1. Introduction

Financial news greatly impacts stock market activities [1]. Currently, widespread news via microblogging significantly increases stock activities and market volatility [2]. Because of its importance and dynamicity, stock market news is an exciting sector for readers and the research community [3,4]. As it is so significant, it can be used to solve challenges and enhance prediction and classification applications and approaches [5,6]. The stock market sector is connected to the financial sector because stock prices are of great interest to stakeholders. Therefore, the availability of stock price news increases the rationalization of the decision-making process for investors and analyzers [7]. Such data is vital for improving stock market analysis methodologies and their applications [3].

Saudi Arabia has the largest stock market in the Middle East [8]. It is also among the top 20 most significant countries in the global ranking of leading economies [9], and it is the largest oil producer in the world [10]. Therefore, unsurprisingly, oil production is one of the largest sectors traded on the Saudi stock market [11], and, in this way, the Saudi stock market significantly contributes to and is tightly coupled with the global stock market [10].

However, there is no dataset indicating the impact of news on the Saudi economy—especially on the stock market. This significant knowledge gap creates a great demand for collecting such data in a structured dataset. This data would enable researchers to conduct extensive studies concerning how financial news impacts Saudi stock market activities.

Data requires proper data quality metrics to assess their quality [12, 13]. There are many techniques to evaluate the dataset's quality. The diversity of these techniques leads research to focus on determining ways that help to choose and apply data quality assessment and development approaches [12]. Most data quality measures are improved based on the specific problem to be solved. However, the basic principles for developing metrics in practice are needed [12, 13, 14]. According to that, data quality assessment metrics are required to be used in this study.

This dataset can be used to solve many problems; however, it is used for stock market news impact and sentiment classification in this study. Therefore, the proper data quality measure is evaluating its polarity distribution

Manuscript received February 5, 2022
Manuscript revised February 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.2.1

for stock market news impact. Also, a simple sentiment classification prediction is applied to assess dataset news quality and manual annotation reliability. So, the data quality must be tested to examine the categories' distribution balance and sentiment accuracy, which affect data quality in several applications. The Plotly Python packages [15] and Naive Bayes (NB) [16] are applied to predict the distribution of polarity over each class and the sentiment accuracy. The data polarity distribution clarifies the dataset's diversity and unbiasedness for a particular category. Also, the high-quality data is the strong base for a strong classifier with reasonable prediction accuracy. So, estimating the first result should be reasonable, i.e., high accuracy but not significantly above the expectations [17]. Analysis and understanding of the model task help determine the study's specific goal and discard other downstream tasks [18]. Furthermore, creating a simple model clarifies how challenging the problem is and its dimensions. Before applying the deep complex models makes things go wrong sometimes, it is necessary to realize the signals of using basic methods in the beginning [19]. Therefore, creating a simple first model needs a few steps to be followed, starting with a model that only uses the automatic text features. This step eases moving from raw data to prediction results fast.

This study provides a pre-processed, classified, and evaluated Saudi stock market dataset. Thus, the assessed dataset proves its quality and reliability level for the research community as a valuable resource. The qualified dataset encourages the researchers and other users to trust implementing their studies and applications based on it smoothly.

2. Data Description

This work presents a reliable, high-quality, classified Saudi stock market dataset, describing the effect of financial news on market behavior in a usable way for the research community. This dataset has been created via a structured methodology, considering many features, including the sensitivity and dynamicity of the data. Therefore, a process was applied through data collection, merging, processing techniques, and descriptive statistics approaches.

2.1 Data Gathering

The dataset was gathered from the Saudi stock market news platform, Tadawul, which is the only news entity authorized for the Saudi stock exchange. The Tadawul website offers many data types based on user requirements [20]. The collected data was drawn from the Corporate News (CN) and Historical Data Stocks (HDS) databases for the years 2011 to 2019. The date range of the collected news was selected based on the news' availability in Tadawul's legal subscription for the reference data. It allows the subscribers to extract only the last ten years, but the 2020 year will be discussed in a separate study. CN was chosen because it shows market-related news for each company on the Saudi stock exchange, alongside important corporate attributes and details. HDS was chosen to link daily stock fluctuations to corresponding daily stock news. Together, the databases show the relationships between, and impacts of, corporate news and Saudi stock exchange fluctuations for the selected date range. Figure 1 represents the Tadawul database contents, from which the CN and HDS datasets were extracted and subsequently merged. The dataset attributes were defined as the following news features: Sector, Company Name, Company ID, Date, Time, News Title, News Details, Opening, Highest, Lowest, Closing, Change %, Change, Quantity Handled, Total Current Value (SAR), and Number of Deals. Table 1 summarizes the dataset elements—variables, type, description, and source. It shows the extracted variable descriptions, their sources, and the main procedures employed in the dataset creation.

The dataset presented here includes 30,098 published news pieces, alongside stock fluctuation information across nine years (2011–2019).

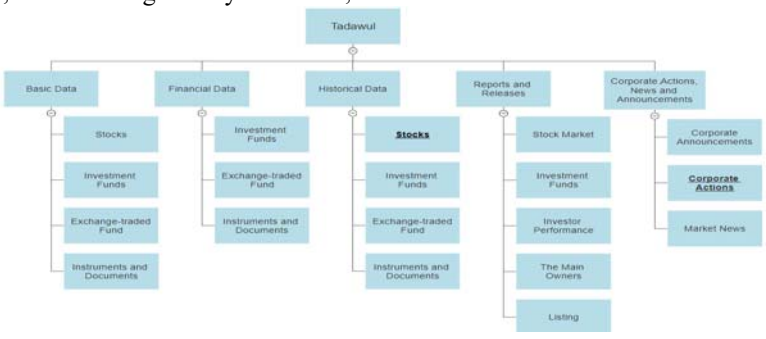

Fig. 1 shows the database contents of the Tadawul website, where CN Data and HDS have been extracted.

Table 1: Variable's description.

Variable	*Type*	*Description*	*Source/Procedure*
Sector	Categorical (text)	The general sector name of the published news and stock numerical information. Type of sector, assuming one of 10 categories as main sectors, contains 20 sub-sectors.	CN and HDS merged by using python libraries based on the Date, Sector, Company Name, and Company ID as the foreign keys.
Company Name	Categorical (text)	The specific company name of the published news and stock numerical information. Type of company name, assuming one of 204 categories as companies' names.	CN and HDS merged by using python libraries based on the Date, Sector, Company Name, and Company ID as the foreign keys.
Company ID	Categorical (text)	The general sector ID of the published news and stock numerical information. Type of company id, assuming one of 204 categories as companies' identities.	CN and HDS merged by using python libraries based on the Date, Sector, Company Name, and Company ID as the foreign keys.
Date	Date	Date of publishing news and stock numerical information. Type of date, assuming one day of 9 years from 2011 till 2019.	CN and HDS merged by using python libraries based on the Date, Sector, Company Name, and Company ID as the foreign keys.
Time	Time	Time of publishing news.	CN.
News Title	Text	Title of publishing news.	CN.
News Details	Text	The details of publishing news include the stock news, stock changing, Arabic semantics, … etc.	CA.
Opening	Numeric	The opening price is the daily price of the security first trades at the exchange-opening on a trading day. For the Saudi market, daily trading sessions start at 9:00 am based on South African Standard Time (SAST) Time.	HDS.
The Highest	Numeric	The highest price is the largest price stock trades reached through a daily trading session.	HDS.
The Lowest	Numeric	The lowest price is the largest price stock trades reached through a daily trading session.	HDS.
Closing	Numeric	The closing price is the smallest price at which a stock trades through a daily trading session. For the Saudi market, daily trading sessions end at 2:00 am based on South African Standard Time (SAST) Time.	HDS.
Change%	Numeric	The stock percentage change in the stock price based on a specific period or daily.	HDS.
Change	Numeric	The stock price change between the market closing today and the previous day's closing.	HDS.
The quantity Handled	Numeric	The number of current stocks on a trading day.	HDS.
Total Current Value (SAR)	Numeric	The total value of current stocks on a trading day, in Total Current Value's (Saudi Riyal).	HDS.
Number of Deals	Numeric	The number of stock market deals daily.	HDS.
Polarity	Categorical (text)	The polarity of the news Arabic semantics. Its type is assuming one of 3 categories: positive, negative, and neutral.	CN, and manually classification.
Polarity Words	Text	The polarity words of the news which have Arabic semantics.	CN, Arabic experts' analyzing, and manual classification.

2.2 Data Preprocessing

The dataset developed in this way presents not only Saudi stock news data, but also their characteristics, semantics, and related stock behaviors.

The construction of this dataset involved multiple phases: gathering the data, merging the datasets, and then contents' assessing, cleaning, validating, analyzing, and visualizing the data. The data were gathered via a paid subscription, which allowed the author to legally download them from the Tadawul website as Excel and comma-separated values (CSV) files [21]. The CN and HDS datasets were merged based on matching feature criteria, such as Sector, Company ID, and Dates by using features of Excel and Power Query Add-on. Considering the Arabic semantics of the news variables, polarities, and their corresponding polarity word variables were added to the dataset by native Arabic speakers.

Challenges were faced during the data contents' assessment stage, such as missing, corrupt, incorrect, or noisy data and differentiated data formats. These issues led the author to apply different data cleaning methods. The missing data values were estimated based on formulas or via a minor data scraping process. During the pre-processing phase, data filters were used to remove inconsistent data. The noise, including extra spaces, unusable elements, and scattered elements, was erased via the pandas DataFrame [15]. Incorrect data, such as incorrect dates and misspellings, were manually corrected, and some Arabic tokenization was used to handle the differentiated data formats. This Arabic tokenization filtered the HTML codes and other extra elements to standardize the data sections. Then, a set of manual comprehensive validation tests was applied to examine the correctness of 30,089 observations.

2.3 Data Annotation

Based on the news detail attributes, polarities (and polarity words) were added to the dataset. Polarity was defined as emotions being expressed in the companies' textual news, published in the Arabic language. Such news was divided, based on Arabic semantics, into three categories (polarities): Positive (containing positive Arabic semantics), Negative (containing negative Arabic semantics), and Neutral (containing neither positive nor negative Arabic semantics). For instance, consider the following article:

- The article in Arabic: "يعلن بنك الجزيرة عن توصية مجلس إدارته توزيع أرباح سنوية للمساهمين"
- The article in English: "Bank Aljazira announces the recommendation of its board of directors to distribute annual dividends to shareholders."

Here, the word indicating polarity is "أرباح," which means "dividends," so the polarity for this statement is Positive.

3. Exploratory Data Analysis (EDA)

The readymade pandas and plotly Python packages [15] were applied to obtain descriptive statistics and visualize the data.

Tables 2–4 summarize the dataset's descriptive facts: date variables, categorical variables, and numerical variables.

Table 4 shows the mean values and the Standard Deviation (SD) values of the integer and numeric variables. Opening, Highest, Lowest, and Closing have similar values in calculating the mean value and the SD value. The mean values change does not exceed 1.21, and the SD values change is not exceeding 0.68. Therefore, these values lead to a slight change in the change values, which do not exceed 0.04 in mean values and 4.34 in SD values, which means that stock market activity is unpredictable but logical. The Quantity Handled, Total Current Value (SAR), and Number of Deals values show the high level of activity of the Saudi stock market.

Table 2: Dataset summary statistics – Date variable.

Variable	*Min*	*Max*	*Unique (days)*
Date	01\01\2011	31\01\2019	3240

Table 3: Dataset summary statistics – Categorical variables.

Variable	*Unique*	*Top Count*
Sector (In Arabic)	22	التامين: 6753، المواد الأساسية: 6216، السلع الرأسمالية: 2595، إنتاج الأغذية: 2126، إدارة وتطوير العقارات: 1728
Sector (In English)	22	Insurance: 6753, Basic Materials: 6216, Capital goods: 2595, Food Production: 2126, Real State M & D: 1728
Company ID	206	1330: 450, 7040: 331, 4130: 316, 8110: 300, 4090: 283
Time observation (Hours per day)	24	08:00:18: 31, 08:00:27: 31, 08:00:26: 30, 08:00:37: 28, 08:00:14: 28
Polarity	3	Positive: 15509, Neutral: 11483, Negative: 3106

Table 4: Dataset summary statistics – Integer and numeric variables.

Variable	*Mean*	*SD*
Opening	35.32	39.36
The Highest	35.89	40.04
The Lowest	34.81	38.70
Closing	36.02	38.91
Change	-0.01	1.90
Change%	0.04	4.34
The Quantity Handled	1.513561e+06	6.332587e+06
Total Current Value (SAR)	3.842513e+07	1.705804e+08
Number of Deals	915.77	5485.46

The published news distribution and the polarity distribution of published news are represented in Figure 2-7. The distribution of news published over the selected years showed periodic monthly changes, as evidenced in Figure 2. The top ten sectors and companies that published news most frequently are shown in Figures 3 and 4.

From the published news distribution perspective, the highest amount of publishing news is in 2013, and the lowest year in publishing news is in 2018. Insurance and Basic Materials have the highest amount of publishing news as sectors, and company number 1330 as a company belongs to the Capital Goods sector.

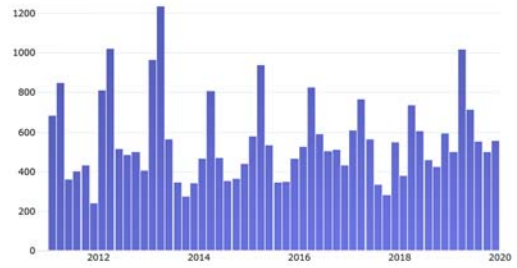

Fig.2 shows the distribution of news publishing over the years.

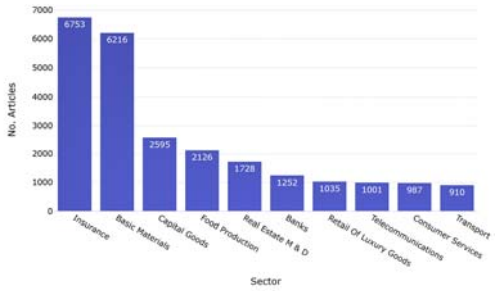

Fig.3 shows the top ten frequent sectors in publishing news.

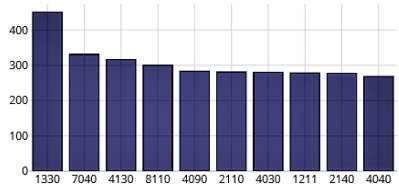

Fig.4 shows the top ten frequent companies in publishing news.

Figures 5–7 present the annual polarity distribution of published news per sector, company, and year within the market index.

Across sectors, the Basic Material sector published the greatest amount of Positive news, while the Insurance sector published the greatest amount of negative news. Also, company number 1330 in the Capital Goods sector published the greatest amount of Positive news. However, company number 4130 in the Investment sector published the greatest amount of negative news. The positive news was published most often from 2012 to 2016, but, after 2017, Negative news was published more often than Positive news.

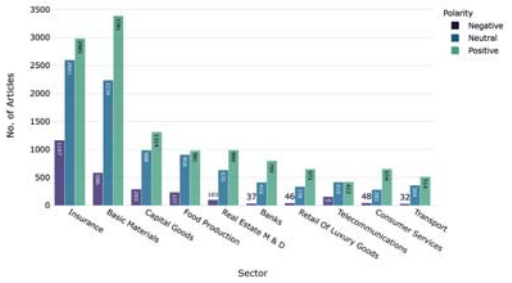

Fig.5 presents the annual polarity distribution with published news over each sector.

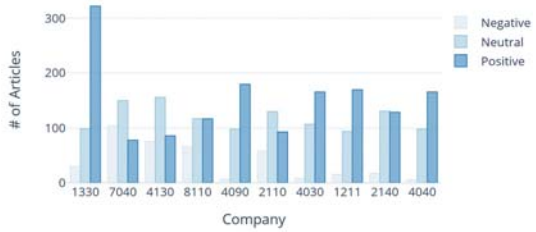

Fig.6 presents the annual polarity distribution with published news over each company.

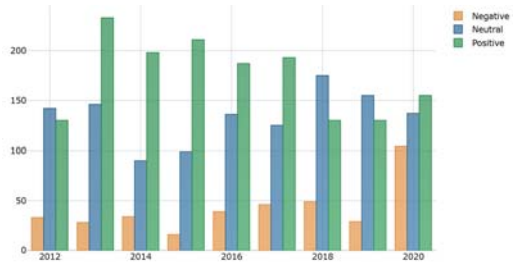

Fig.7 presents the annual polarity distribution with published news over the years for the market index.

Furthermore, Figure 8 shows the different monotonic correlation rankings of all numerical variables. For every two numerical variables, these correlations were calculated based on standard deviations. Figure 9 clearly demonstrates the correlation between the Number of Deals and Change based on the Total Current Value's size.

The rank correlations method was used to measure how strong the linear relationships were between two or more

numerical variables. Based on standard deviations, the correlations between two numerical variables were calculated. The highest correlations were found between the Opening, Highest, and Lowest dataset attribute values. There was also a high correlation between those three variables and the Closing value. A clear correlation was found between the Change and Change % values, as well as between the Quantity Handled and the Number of Deals and Total Current Value's values. Moreover, the correlation between the Number of Deals and Change based on the Total Current Value's size is calculated. If the Number of Deals increased, the Total Current Value's also increased, while the Change value decreased; and vice versa.

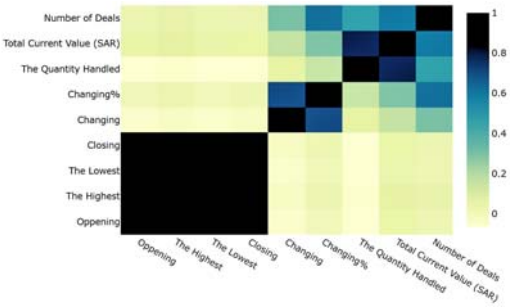

Fig.8 shows the difference of monotonic correlation ranking of all the numerical variables.

Fig.9 shows the correlation Number of Deals and Changing in price based on Total Current Value.

Figure 10 shows an excellent example of the correlation between Closing and Change across polarity, based on the Total Current Value's size.

Finally, it was essential that this dataset be examined for relationships between categorical and numerical variables. Numerical variable correlations were statistically analyzed across the polarity values to determine the degree to which these stock variables moved in relation to each other. As an example of the correlation between Closing and Change across polarity, based on the Total Current Value's size. If the Closing price increased, the Total Current Value decreased, while the Change value increased; and vice versa.

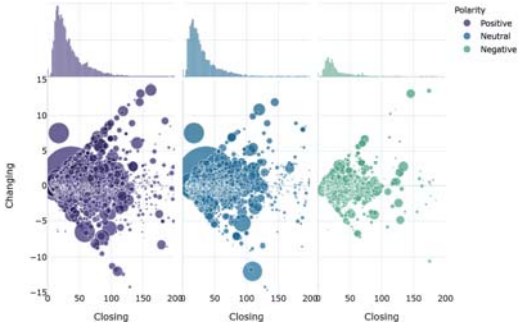

Fig.10 shows the correlation Closing price and Changing in price across polarity based on Total Current Value.

4. Data Quality Assessment

Data evaluation depends on two ways: the distribution balancing of polarity classes and the sentiment classification accuracy of the dataset. Exploratory data analysis was conducted for both categorical and numerical variables. Concerning the single categorical variables, the polarity distribution showed how the data were balanced and their polarity tendencies. The data distribution balancing is proved data unbiasedness. Most news belonged to the Positive category, and the Negative category contained the fewest publications. The published news distribution and the polarity distribution of published news are represented in Figure 11, which shows the news label values according to polarity distribution.

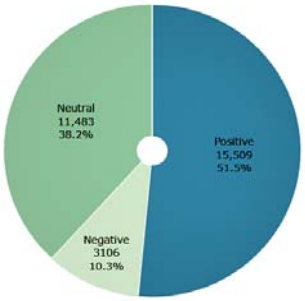

Fig. 11 shows the distribution of polarity over sectors.

Assessing data quality based on a simple model is better than complex models because of its fast and easy application. Applying a simple model as NB is used to test the dataset quality by estimating the first result and knowing the issue dimensions [19]. Therefore, evaluating data through the NB model with split data requires only a few steps to train the model, depending on the auto parameters. Dataset evaluation requires a feature dataset for learning purposes and a target dataset for validation

and testing steps based on the NB classifier. Depending on the classifier, the data evaluation generates an evaluation report of multiple sentiment categories (See Table 5).

Table 5: The Model Evaluation Metrics Report of Multi-class Classification.

Model Name	*Class*	*Pre-cision*	*Re-call*	*F1-score*	*Support*
Naïve Bayes	Negative	0.46	0.78	0.58	311
	Neutral	0.71	0.78	0.74	1148
	Positive	0.82	0.64	0.72	1551
	Accuracy	-	-	0.71	3010
	Macro avg	0.66	0.74	0.68	3010
	Weighted avg	0.74	0.71	0.72	3010

The confusion matrix of NB is applied to compare the prediction of the feature dataset with the actual target dataset. It presents the evaluation metrics results of the classification report in the validation-set predicted data. The following figures show the prediction results for multiclass classification of NB (See Figure 12), and the normalized values for each outcome (See Figure 13). As shown in the confusion matrix, the neutral class's prediction achieved a high score, which is hardly reached at an easy level of ML such as BL. These results support the achievement of the quality level that the qualified dataset required.

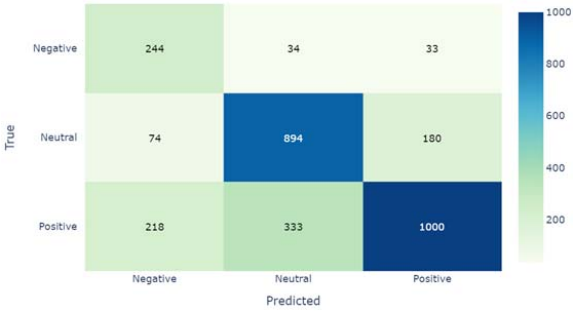

Fig. 12 Shows the Confusion Matrix of NB Multiclass Classification.

Fig. 13 Shows the Normalized Confusion Matrix of NB Multiclass Classification.

The experimental result of the NB model in predicting the sentiment classification is 68% with auto features and with no hyperparameters tuning (See Table 6). The classification sentiment accuracy is a high and sensible result which tells that the dataset is qualified and reliable. However, the dataset has a massive number of articles which is more than the ability of NB, which works better with small data [22]. So, other advanced ML and DL models may get better results by seeking a higher accuracy.

Table 6: The Evaluation Metrics Report of The Datasets.

Model Name	*Dataset*	*Accuracy*	*Precision*	*Recall*	*F1-score*
Naïve Bayes	Train	0.726929	0.761358	0.726929	0.730638
	Valid-ation	0.710299	0.742229	0.710299	0.715460
	Test	0.683389	0.714801	0.683389	0.686385

5. Conclusion

Data description determines the data import and the procedures of methodology stages. Data collection is the initial stage that is applied through Tadawul references data. Also, describing the dataset's contents eases its preparation, annotation, and performing upcoming tasks. Data preparation procedures should be performed to deal with missing data, corrupt data, incorrect data, noise, etc. Data preparation includes merging datasets, assessing data contents, cleaning, and validation via preprocessing, Excel, Python, and Arabic tokenization. Then, the data annotation process is performed manually by Arabic native speakers associated with the domain. For more understanding of the data, it explored based on EDA procedures into descriptive statistics of the published news distribution and descriptive analytics of the sentiment label. Then, the data quality is evaluated based on the distribution balancing of the different classes and sentiment classification. The data distribution balancing

examination proved data quality and its unbiasedness. At the same time, the sentiment accuracy by the NB model based on the data achieves 68%, which is a sensible result that shows the data quality and reliability. Thus, performing these sequential stages is essential to ensure data readiness for the upcoming approaches and tasks based on the provided data. The delivered dataset is preprocessed, labeled, and assessed to be a valuable resource for the research community in several usages.

References

[1] X. Li, H. Xie, L. Chen, J. Wang, and X. Deng, "News impact on stock price return via sentiment analysis," Knowledge-Based Syst., 2014.

[2] J. R. Piñeiro-Chousa, M. Á. López-Cabarcos, and A. M. Pérez-Pico, "Examining the influence of stock market variables on microblogging sentiment," J. Bus. Res., vol. 69, no. 6, pp. 2087–2092, 2016.

[3] M. C. Mariani, M. A. M. Bhuiyan, O. K. Tweneboah, M. P. Beccar-Varela, and I. Florescu, "Analysis of stock market data by using Dynamic Fourier and Wavelets techniques," Phys. A Stat. Mech. its Appl., vol. 537, p. 122785, 2020.

[4] F. Jareno Cebrian, "The sensitivity of sectoral returns to real interest rates and inflation," Investig. Econ., 2006.

[5] X. Ding, Y. Zhang, T. Liu, and J. Duan, "Deep learning for event-driven stock prediction," in IJCAI International Joint Conference on Artificial Intelligence, 2015.

[6] Q. Al-Radaideh, A. Assaf, and E. Alnagi, "Predicting stock prices using data mining techniques," Int. Arab Conf. Inf. Technol., 2013.

[7] A. Mittal and A. Goel, "Stock Prediction Using Twitter Sentiment Analysis," http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf, 2012.

[8] A. Badawi, A. AlQudah, and W. Rashideh, "Determinants of Foreign Portfolio Investment in Emerging Markets: Evidence from Saudi Stock Market," SSRN Electron. J., 2017.

[9] I. A. Gelil, N. Howarth, and A. Lanza, "Growth, Investment and the Low-Carbon Transition: A View from Saudi Arabia," Kapsarc, no. August, pp. 1–20, 2017.

[10] M. Alharbi, "The Reliance of the Saudi Economy and Adequacy of its Foreign Reserves with Reference to Oil Price Volatility: An Overview," Int. J. Bus. Adm. Stud., vol. 5, no. 6, pp. 329–339, 2019.

[11] I. Henriques and P. Sadorsky, "Oil prices and the stock prices of alternative energy companies," Energy Econ., vol. 30, no. 3, pp. 998–1010, 2008.

[12] Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. ACM computing surveys (CSUR), 41(3), 1-52.

[13] Huang, K., Lee, Y., and Wang, R. Quality Information and Knowledge. Prentice Hall, Upper Saddle River: N.J. 1999.

[14] Kahn, B. K., Strong, D. M., and Wang, R. Y. Information Quality Benchmarks: Product and Service Performance. Commun. ACM, (2002).

[15] "The official home of the Python Programming Language." [Online].

[16] Murphy, K. P, "Naive bayes classifiers," *University of British Columbia*, *no. 18*, vol. 60, pp. 1-8, 2006.

[17] Stehman, S. V., & Foody, G. M, "Accuracy assessment," *In the SAGE handbook of remote sensing*, London: Sage, pp. 297-309, 2009.

[18] Kieras, D. E., & Butler, K. A, "Task Analysis and the Design of Functionality," *The computer science and engineering handbook*, vol. 23, 1401-1423, 1997.

[19] Brownlee, J, "Machine learning algorithms from scratch with Python," *Machine Learning Mastery*, 2016.

[20] "The Saudi Stock Market Tadawul." [Online]. Available: https://www.tadawul.com.sa/wps/portal/tadawul/home/.

[21] Available: https://www.python.org/.

[22] A. Gaydhani, V. Doma, S. Kendre, and L. Bhagwat, "Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach," *arXiv preprint arXiv:1809.08651*, 2018.

Eman Alasmri is a master's student at King Abdulaziz University in the Department of Information System. Her research interest includes Machine Learning, Deep Learning, Data Analysis, Web Design and Development.

Dr. Mohamed ElEliemy is a professor who is affiliated to both King Abdulaziz University in Jeddah, KSA and Ain Shams university in Cairo, Egypt. His research interest is in Data Analysis and Mobile Computing. He has an extensive publication record in both fields.

Fahd Alotaibi earned his PhD in Computer Science from the University of Birmingham, United Kingdom in 2015. He is currently working as an Associate Professor in the Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia. His research interests include Natural Language Processing, Data Science, and Data Mining.

Khaled H. Alyoubi is an Associate Professor of Computer Science at Faculty of Computing and Information Technology in King Abdulaziz University. His research interests include Data sciences, Data management, IR, Data Analytics and Data mining. He has a PhD in Computer Science from Birkbeck University of London, UK.