Data Mining for Managing and Using Online Information on Facebook

Nidal Al Said

College of Mass Communication, Ajman University, Ajman, United Arab Emirates
Email: n.alsaid@ajman.ac.ae

***Abstract*—The problem under the study of this work is investigating data mining algorithms for intelligent analysis of data written in Arabic. The study comprised instead involves several stages, including Data Collection and Pre-Processing; Data Mining Algorithms (Multinomial Naïve Bayes Classifier, Naïve Bayes Classifier, Support Vector Machine and Modified K-Means); Study Results Processing and Software Implementation. A total of 16,732 Facebook posts written exclusively in Arabic were downloaded. Almost two-thirds of them (namely 11,155 items) were used to train algorithms, while the rest (5577 items) were subject to research. The training data were categorized into five groups based on subjects (politics, entertainment, medicine, science, and religion) with five keywords used for testing in each group. Most posts (4736 items) were related to politics. The most accurate algorithm proved to be the multinomial Naïve Bayesian classifier for the maximum number of test data, while the minimum values of this feature were recorded for the Support vector machine. The effectiveness of the multinomial Naïve Bayesian classifier algorithm was most remarkable for the maximum amount of data, while the Support Vector Machine was most effective for the minimum amount. The argument's fit score is maximum at 5577 data points for the multinomial Naïve Bayesian classifier and 1394 data points for K-means. To improve and refine the results of data mining, the sample must be expanded, adding more data classes and keywords. Other machine learning models, such as deep learning algorithms, could also be used. The significance of investigation is very important because it expands our knowledge about the use of Machine Learning Algorithms to mine Arabic texts on social media platforms.**

***Keywords*—social networks, data mining, classification, accuracy, multinomial naive Bayes classifier**

I. Introduction

In the 21st century, the trend towards digitization and the transition to information technology is especially popular and widespread. Authors of work [1] confirm that the "Internet and computer connect all over the world with web communication, information technology trends to digitalization". It has profoundly modified human reality, the way of communicating and transferring information [2].

The Internet is a communication and data-sharing medium. Social networks were created for personal communication between people depicted as charts, where nodes are users and connections are their relationships and interactions [1, 3]. Social networks (Twitter, Facebook, Instagram, and others) have become a new way to share information online: news, stories, blogs, podcasts, tips, etc. Originally, social networks were intended only for personal use expressed in the exchange of messages between users, that is to say, communication between them or between a group of people [1–3]. These days, they are used by organizations, companies, politicians, or other influencers to promote services or goods, identify potential clients, quickly post information in the right place at the right time, or promote social/election campaigns [4, 5]. In addition, networks provide an opportunity to receive feedback and communicate public attitudes to these messages [3].

In its time, Facebook became a conceptually new social network, which began to grow and change rapidly, adapting to users' needs and desires. In particular, "Users can create a customized profile to share information about themselves with others that have agreed to be their "friend" [6]. Over 18 years of existence, active participants exceeded 2.8 billion people a month [7]. With Facebook, boundaries between users are removed, making it easier to communicate. It provides an opportunity to share events, life stories, or activities with some degree of reliability. With so many accounts, the platform can be a target for scammers. Mainly when the platform's popularity has led people to share personal life events on different websites. In creating alleged "fake" accounts, they use social media to steal personal information, finances, or other illegal activities [1, 6, 8]. Cybercrime is a reasonably common term nowadays, and it's growing steadily. Besides direct attacks against online information, there are indirect actions involving web pages and blogs to psychologically influence people by publishing statements or images to discredit an individual or incite hatred among a group of people [9].

Furthermore, the activity of online media platforms in crises such as natural or man-made catastrophes (floods, earthquakes, and so on) is increasing significantly, which enables news to be learned more quickly than from Television (TV) [10, 11]. They also make it possible to find relatives and friends sooner. In some cases, social networks even play a crucial role in searching for missing

Manuscript received December 29, 2022; revised February 13, 2023; accepted March 27, 2023; published August 3, 2023.

persons [9]. Based on the foregoing, social networks contain a significant amount of information, the quantity of which continues to grow. Therefore, a software code—Intelligent Data Analysis (IDA) [5, 12] was written to facilitate data structuring, understanding, and retrieval, that is for data mining. This algorithm has been subject to numerous changes such as MapReduce, Graph, MongoDB, and others [4, 13].

Large businesses that can maintain innovation acquire a huge competitive edge in the business sector. Businesses can use the information provided by these enterprises' keen attention to following and monitoring news sources in forums, social media, and e-commerce when making decisions. Due to the abundance of data given in these resources, sentiment analysis from reviews of goods and services, user emotions, and significant feedback may all be gleaned. Proper sentiment analysis is crucial in making all of this possible. In the study [14], fresh data sets for Turkish, English, and Arabic were produced, and for the first time, comparative sentiment analysis was carried out on texts in the three languages. Also, a very thorough analysis of the results of the deep learning and machine learning models, as well as the pre-trained language models for Turkish, Arabic, and English, was provided to the researchers.

The work aims to explore the use of text extraction technology in Facebook for data mining of information in Arabic. To achieve this goal, the following tasks need to be completed the following tasks to be completed:

- Selecting algorithms to be tested based on a literature review;
- Creating a data set based on Facebook publications for training and testing,
- Identifying the best option to work with the Arabic text.

In the following section, a literature review of existing data mining algorithms is provided, and the most appropriate one is selected. Materials and methods for this study are described in Section III. The results are then presented and discussed in Section IV and Section V. The last section outlines the main findings and prospects for further research based on existing uses of the data from this work.

II. Literature Review

Media platforms have become a new place to broadcast news. Thus, competition between websites and social networks has increased dramatically [2].

The mobile applications of social networks offer the possibility of connecting people who are very far apart from each other [4]. Consequently, an enormous amount of information is accumulated, particularly personal data, which is the primary target of scammers [15].

There are now numerous social media platforms, such as Instagram (for sharing images and photos with public or private audiences), YouTube (for sharing videos with the public), Twitter (for sharing short stories with the public), LinkedIn (for building professional connections), and many others [9]. Media platforms can be used in various ways for research, including collecting information, conducting surveys, linking interested individuals, etc. [16].

Gkikas and Tzafilkou *et al*. [17] looked at the relationship between brand awareness and customer engagement in sponsored Facebook picture postings. The following text attributes were looked at: (i) readability scores, (ii) text length, and (iii) the number of hashtags. An export of 135 picture posts with accompanying text descriptions from a Facebook business page for fashion retailers included post-performance indicators for engagement (measured in likes), and awareness (expressed in reaches and impressions). Positive correlations between all performance indicators and the length of the content and the number of hashtags were found. Whereas the Flesch Kincaid reading easiness score was solely connected with awareness measures of reaches and impressions, the Gunning Fog readability index showed high relationships with both engagement and awareness.

Overall, the findings showed that postings with easily readable content, lengthy (more than 31 words, or more than 321 characters), and hashtag-rich often perform better in terms of interaction and awareness. By shining light on the function of text features of branded messaging in social media and providing guidance for brand communication and social media message strategies, this research adds to the body of existing studies.

Lathan and Kwan *et al*. [18] presented a systematic review of studies using Facebook data in health research. They found variability in IRB review, consent, and use of identifying information. Authors could identify individual users in half of the studies with verbatim quotes. Greater ethical guidance is needed for research involving social media data. Considerations around sampling and analytic approaches were also discussed.

Intelligent text analysis is the process of assessing a great deal of information using fully automated or partially automated means [19]. Analytical methods deal with information in different ways: generalization, grouping, classification, description of trends, and so on [20]. For Internet pages, the IDA is commonly referred to as Internet Text Analysis (ITA) or data mining, essentially a larger version of the IDA. This algorithm makes it possible to extract text and media data (video, audio, pictures) [21, 22]. There are also reported to be three types of data mining. These include evaluating the web page structure, evaluating the information found on web pages, and comprehensive analysis of using the Internet (Fig. 1) [23, 24].

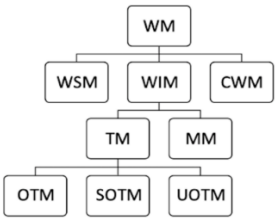

Figure 1. Structure of web mining [20, 25]. *Note:* WM—web-mining, WSM—web structure mining, WIM—web information mining, CWM—comprehensive web mining, TM—text mining, MM—multimedia mining, OTM—organized text mining, SOTM—semi-organized text mining, UOTM—unorganized text mining.

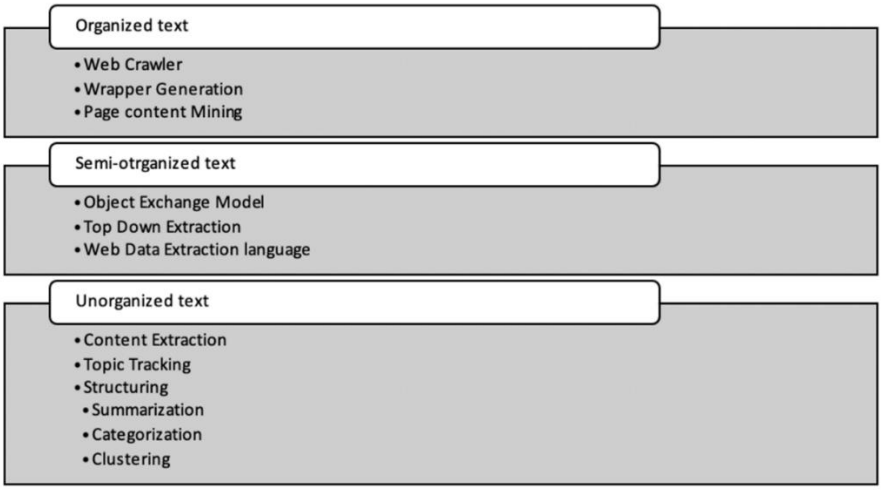

Figure 2. Techniques for working with text in ITA.

Text data is an integral part of the information in social networks, so the application of data mining is especially important for its structuring. To this end, different approaches are used (Fig. 2), most of which are designed to work with unorganized text, as its volume is four times larger than that of organized [5, 23].

Data mining plays an important role in various human activities because it extracts unknown useful patterns (or knowledge). Due to its capabilities, data mining became an essential task in a large number of application domains such as banking, retail, medical, insurance, bioinformatics, etc. [26].

The telecommunication industry, being one of the major sectors in the world, has also been infiltrated by frauds. Telecommunication fraud is a combination of a variety of illegal activities like unauthorized and illegitimate access, subscription identity theft and international revenue share fraud etc. Frauds have proven to be detrimental to the prosperity of a company and impact customer relations and shareholders. This paper presents the implementation details for detecting telecommunication fraud using Data Stream Analytics and Neural Network classification-based Data Mining [27].

Different Machine Learning (ML) approaches are used to data mining, including Decision Trees (DT), Support vector Machine (SVM), K-Nearest Neighbor (kNN), Naive Bayes Classifier (NBC), and others [23].

Demand forecasting has always been a concern for business owners as one of the main activities in supply chain management. Traditional statistical methods and techniques are biased in demand prediction and are not accurate; so, machine learning algorithms as more popular techniques have been replaced in recent researches in the literature. As of the time of conducting these researches [26–28], extreme learning machine has not been used for intermittent demand prediction, so the novelty of our research is to adopt this algorithm and also other machine learning algorithms such as K-nearest neighbors, decision tree, gradient boosting, and Multilayer Perceptron to examine its accuracy and performance in comparison to other approaches.

The DT consolidates all data into groups. It is a very straightforward method by its nature, which is easy to interpret the results. DT is used for checking corner information and allows the best way to determine solution accuracy [29]. The SVM addresses the search for surface solutions to divide data points into classes using carrier vectors that are the effective components of the sample [30]. The NBC requires an enormous quantity of memory to implement the algorithm, making it difficult to use. Moreover, it is one of the best solutions for handling textual data [5]. The kNN, as an excellent example of nonparametric classification, measures the distance between training examples and the minimum value is determined [23]. For more specific studies on the frequency of individual phrases or words appearing in the text, a MNBC has been developed [31]. Where it is necessary to process a considerable quantity of data, clustering methods are used, such as X-means, K-mean, hierarchical clustering (also called hierarchical cluster analysis or HCA).

The HCA algorithm is the simplest of the proposed algorithms, dividing the data set into very detailed clusters. At the same time, it is not applied in cases where a strict hierarchy between data points does not exist [32]. X-means and K-mean are similar in their basis. K-mean splits a set of information into n-pieces of clusters, where the values of each cluster are fairly identical and differ from the others at the same time. And X-means is an improved version of the earlier method, where points are progressively distributed by the method of periodically repeated attempts, preserving, and the best solution option [33].

It appears that Machine Learning (ML) techniques can be applied to data mining on social networks (commentary under publications and publications themselves) [34]. Fatima and Srinivasu [30] applied SVM algorithms for document classification. The 4:1 ratio of data for training and analysis has been demonstrated to improve accuracy [30]. A large part of the existing work in this area deals with categorizing data in English [33–35]. Elhoseny [36] suggested the possibility of using the K-means method for working with Arabic text. Their work

underlines that working with Arabic text complicates the application of ML due to the different variations in word spelling [36]. Most publications in which they worked with Arabic employed conventional algorithms (NBC, SVM, kNN) [35]. Some studies use more advanced algorithms such as the deep neural network, which can be applied to analyze the semantic load of the text "mood" [37].

III. Materials and Methods

This study involves several stages, as shown in Fig. 3.

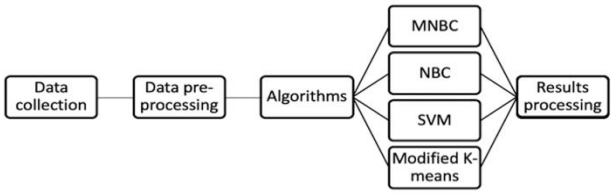

Figure 3. Steps in the research process.

A. Data Collection and Pre-processing

The data were collected using Visual Scraper software. This free program with a simple interface allows extracting information from multiple web pages using URLs. The extracted data is provided as a CSV Excel file.

Thus, 16,732 Facebook posts written exclusively in Arabic and published between April 1, 2021, to September 1, 2021, were downloaded. Of them, almost two-thirds (namely 11,155 items) were used to train algorithms, and the rest (5577 items) were subject to research. Data collected for the study is collected ethically.

Tokenization involves several stages:

- text identification (words, symbols, numbers);
- extracting stop words and non-informative symbols from the text that may affect the readability of program data;
- highlighting keywords for subsequent work.

Also, words repeated in one position were eliminated to expedite the program.

B. Data Mining Algorithms

1) Multinomial naive bayes classifier

This method is often used to determine how many times a word is repeated in a text. The likelihood of the word occurring is calculated as follows:

$$P(c|d) \propto P(c) \prod_{1 \le k \le n_d} P(t_k|c) \qquad (1)$$

where (c) is the probability of a priori hit of the data in class c; $(t_k|c)$ is the probability of conditional hit of the word t in the text d of class c.

The probability of the text d was calculated as:

$$P(c) = \frac{N_c}{N} \qquad (2)$$

where N_c is the total number of documents in class c; N is the total number of documents.

Conditional probability is the relative frequency of using a particular word in a text belonging to class c is as follows:

$$P(t|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}} \qquad (3)$$

where T_{ct} is a number of words in the text of class c; $\sum_{t' \in V} T_{ct'}$ is an amount of words in the text of class c [32].

2) Naive bayes classifier

This algorithm implies creating a tuple of training data. Each of them is represented as an n-dimensional vector of elements $x = (x_1, x_2, ..., x_n)$. It is taken into account that the number of classes is m: $c_1, c_2, ..., c_m$. The classifier assigns X (unknown tuple) to class c_i only when $P(c_i|x) > P(c_j|x)$ for $1 \le j \le m$ and $i \ne j$. The posterior probability of the target class is:

$$P(c|x) = \frac{P(x|C) \cdot P(c)}{P(x)} \qquad (4)$$

where $(x|c)$ is the predictor probability of class c; (c) is a posterior probability of class c;(x) is the posterior probability of the predictor class [38].

3) Support vector machine

This algorithm implies complying with Mercer's theorem expressed by evaluating the kernel on all pairs of data points, resulting in:

$$K(u, x) = \sum_r \varphi_r(u) \varphi_r(x) \qquad (5)$$

where (u) refers to Hilbert space.

Separating function as a linear combination of kernels connected with vectors:

$$f(x) = \sum_{x_j \in S} a_j \, y_j K(u, x) + b \qquad (6)$$

where $y_j \in \{+1, -1\}$ denotes the corresponding class labels; a_j represents the corresponding coefficients; b is the displacement.

The probability of classification is correct at:

$$P_c = \sum_{i=1}^{m} P(c_i) \int_{R_i} p(x|c_i) \, dx \qquad (7)$$

The value $(x|c_i)$ must be greater than 1/m, as otherwise, the probability value will be less than 1 [39].

4) Modified k-means

The method is capable of clustering data from a large sample. The values used are as follows: D $(d_1, d_2, ..., d_n)$ is a data set; n is a number of data points K is a cluster; and X $(x_1, x_2, ..., x_n)$ are data points [40].

C. Study Results Processing

The reliability of the methods was determined using a non-conformity matrix (Fig. 4) by entering true and false cases of the estimate.

		Projected values	
		Positive	Negative
Existing values	Real	***RP***	***RN***
	Imaginary	***IP***	***IN***

Figure 4. Non-conformity matrix. Note: RP—really positive, RN—really negative, IP—imaginary positive, IN—imaginary negative.

The matrix values allowed calculating *Accuracy* (the overall algorithm correctness (the frequency of correct results), *Recall* (the algorithm effectiveness in case of detecting positive readings), *Precision* (conformity with the set of readings results), and *Specificity* (the algorithm effectiveness in case of detecting negative readings [41]:

$$Accuracy = \frac{RN+RP}{RP+RN+IP+IN} \quad (8)$$

$$Recall = \frac{RP}{RP+IN} \quad (9)$$

$$Specificity = \frac{RN}{RN+IP} \quad (10)$$

$$Precision = \frac{RP}{RP+IP} \quad (11)$$

D. Software Implementation

The algorithms were compiled in MATLAB software version 9.4. The results were collected and processed using Microsoft Office 2016.

IV. Results

The training data were divided into five groups by subjects (politics, religion, science, entertainment, and medicine). The relevant number of values is provided in Fig. 5.

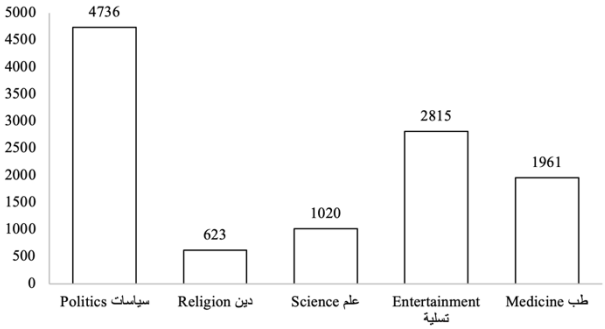

Figure 5. Distribution of posts by subjects.

As seen from the figure, most posts (4736 items) are devoted to politics, followed by entertainment (2815 items), including various stories and poems, lyrics of songs, and other posts. The subject of medicine comprises 1961 posts. The next to last place was taken by science with 1020 positions. The lowest number of publications was religious, at 623 units.

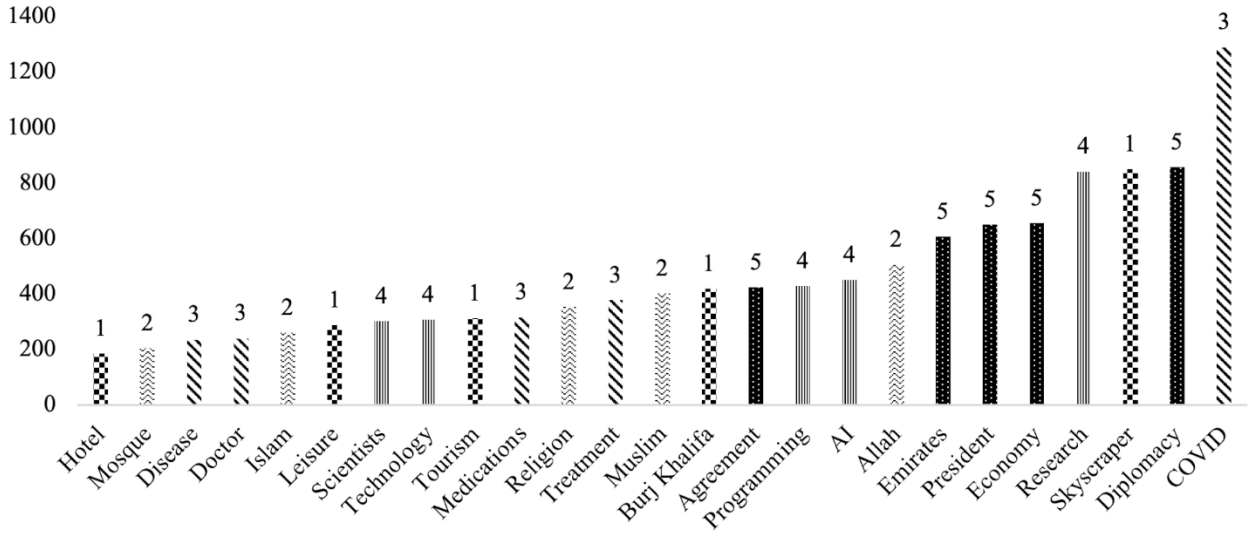

Figure 6. Number of keywords extracted from posts. Note: 1—Entertainment; 2—Religion; 3—Medicine; 4—Science; 5—Politics.

Fig. 6 shows that the number of words related to political topics is: diplomacy (26.85%), economy (20.53%), president (20.4%), migration (19%), and agreement (13.27%). The subject of medicine included such keywords as COVID is (52.42%), treatment (15.36%), medications (12.84%), doctor (9.79%), disease (9.59%). Among science keywords, research accounted for 36.11%, artificial intelligence (AI) for 19.34%, programming for 18.4%, technology for 13.21%, and s cientists for 12.95%. The group of entertainment comprised skyscrapers (41.37%), Burj Khalifa (20.42%), tourism (15.17%), leisure (14.00%), hotel (9.04%). The words related to religions included Allah (29.27%), Muslim (23.25%), religion (20.47%), Islam (15.15%), and a mosque

(11.86%). Due to the lack of conformity, these words can be used for training algorithms.

Data mining algorithms were tested using test and learning data to see exactly how the templates manage the text Table I.

TABLE I. ACCURACY (A), RECALL (R), PRECISION (P) AND SPECIFICITY (S)

Algorithm	Value	Number of units			
		5574	4183	2788	1394
MNBC	A	59.60	59.36	58.61	58.46
	R	59.50	68.75	76.75	77.07
	P	53.59	47.83	46.30	46.51
	S	59.68	53.54	48.01	47.43
NBC	A	57.74	57.92	58.03	58.25
	R	56.46	63.78	67.50	63.95
	P	48.30	51.98	51.76	49.47
	S	58.61	53.28	49.60	52.36
K-Means	A	56.75	56.20	55.87	55.87
	R	62.76	59.09	55.35	55.36
	P	52.70	55.63	55.39	48.00
	S	51.58	53.35	56.75	56.37
SVM	A	51.80	51.71	51.69	51.51
	R	48.61	55.43	48.00	48.08
	P	43.42	45.95	55.05	54.17
	S	54.11	48.78	55.84	55.27

As seen, the accuracy of the MNBC, K-means, and SVM algorithms drops with decreasing amounts of data (in %): 59.6, 59.36, 58.61, and 58.46; 56.75, 56.2, 55.87 and 55.87; 51.8, 51.71, 51.69, and 51.51, accordingly. In contrast, the accuracy increases for NBC (in %): 57.74, 57.92, 58.03, and 58.25. There are no strict correlations in the case of other indicators, and the changes are more fluctuating. Thus, in MNBC, the values for *Recall* are (in %): 59.5, 68.75, 76.75, and 77.07; for *Precision* 53.39, 47.83, 46.30, and 46.51. In NBC, the values are the following (in %): for *Recall*, 56.46, 63.78, 67.5, and 63.95; for *Specificity*, 58.61, 53.28, 49.6, and 52.36; and for *Precision*, 48.30, 51.98, 51.76, and 49.47. In the case of K-means, the following values (in %) were observed: *for Recall*, 62.76, 59.09, 55.35, and 55.36; for *Specificity*, 51.58, 53.35, 56.75, and 56.37; and for *Precision*, 52.70, 55.63, 55.39, and 55.68. When conducting SVM testing, the following values (in %) were obtained: for *Recall*, 48.61, 55.43, 48.00 and 48.08; for Specificity, 54.11, 48.78, 55.84, and 55.27; and for Precision, 43.42, 45.95, 55.05, and 54.17.

Hence, the most accurate algorithm proved to be the MNBC for the maximum number of test data, while the minimum values of this feature were recorded for the SVM. If small samples are required, it is preferable to use NBC. The model's effectiveness in identifying positive values is excellent in the K-means approach, especially for large amounts of information. The NBC is suitable for smaller data amounts, and SVMs proved to be the most effective in similar features. The effectiveness of the MNBC algorithm was most remarkable for the maximum amount of data, while the SVM was most effective for the minimum amount. The argument's fit score is maximum at 5577 data points for MNBC and 1394 data points for K-means.

V. DISCUSSION

Yang *et al.* [34] studied data mining from Twitter and Weibo using the K-means model. It has been shown to help analyze Chinese content on Weibo or English content on Twitter. The algorithm, then, assigns the data to one of the four clusters. The subject of specific clusters is broader than in the case of Arabic text and includes several categories, such as the media and young people. It was used to streamline the algorithm and enhance its overall accuracy. This study showed that the K-means algorithm could be used to analyze publications in Arabic with fairly good precision [34].

The work by Bozkır *et al.* [19] aimed to ensure that data mining, particularly the SVM algorithm, can detect factors that affect the frequency and duration of Facebook usage. The amount of data used for training is four times higher than for the test. In the present work, this model was used to classify the information obtained from the network, and the number of data formed was equal to two-thirds of the total.

Elnagar *et al.* [35] applied deep neural network patterns to data mining of Arabic text from different databases. The accuracy of the algorithms used is above 85%, indicating that they can be applied to aggregate information. The models used in the present study demonstrate a 25–35% lower accuracy. It may be explained by networks' use of informal discourse, making readability difficult [35].

Seah and Shim [42] used data that has been classified into several groups using machine learning, which may be associated with suicide. At the same time, this aims to establish the feasibility of algorithms in data classification [42].

A machine learning algorithm was also applied to detect fake Facebook profiles. The K-means method has been shown to be sufficiently accurate (67.31%) to split into two groups. It has also been excellent for detecting positive values and cannot be used to identify negative values. This algorithm can also be used to classify information in Arabic [6].

MNBC can be employed to data mining from Twitter with a maximum possible accuracy of 73.15%. This method is also well-suited to processing Arabic datasets [31].

The author compares the results with the work by [43]. In this work, machine learning algorithms are used for classifying clinical reports into four different classes. After performing classification, it was revealed that logistic regression and Multinomial Naïve Bayesian Classifier (MNBC) give excellent results by having 94% precision, 96% recall, 95% f1 score and accuracy 96.2%. Various other machine learning algorithms that showed better results were random forest, stochastic gradient boosting, decision trees and boosting. The efficiency of models can be improved by increasing the amount of data. Also, the disease can be classified on gender-based such that we can get information about whether the male is affected more or females.

The author compared the data and results [43] in Table II.

TABLE II. ACCURACY (A), RECALL (R), PRECISION (P)

Algorithm	Study	Value, %		
		A	R	P
MNBC	Our	59.60	59.50	53.59
	[43]	96.20	96.00	94.00

It can be concluded that the sample used significantly affects the values of the accuracy, recall and precision.

Analyzing text has become an essential part of our lives because of the increasing number of text data which makes text classification a big data problem. Arabic text classification systems become significant to maintain vital information in many domains such as education and health sector, and public services. In the work [44], the Arabic text classification model is developed using various algorithms namely MNBC, Bernoulli Naïve Bayesian (BNB), Stochastic Gradient Descent (SGD), Logistic Regression (LR), SVC, Linear SVC, and convolutional neural networks (CNN). These algorithms have been implemented utilizing the Al-Khaleej dataset. The experiments are carried out with various representation models and it is observed that CNN with character level model outperforms others. The result of CNN exceeds the state-of-the-art machine learning method with an accuracy equal to 98%. Authors of work [44] have said that this information will be useful in different domains, particularly on social media.

VI. CONCLUSIONS

The aim of this work was to explore the use of text extraction technology in Facebook for data mining of information in Arabic. The paper analyzed the feasibility of applying machine learning techniques to data mining text obtained from the Facebook social network. The study involved several stages, including Data Collection and Pre-Processing, Data Mining Algorithms (Multinomial Naïve Bayes Classifier, Naïve Bayes Classifier, Support Vector Machine, and Modified K-Means), Study Results Processing, and Software Implementation. A total of 11,155 data units were divided into five groups by subjects (politics, religion, science, entertainment, and medicine), with five keywords used for testing in each group. It was found that the highest level of precision is characteristic of the maximum possible number of data and increases as the number of data increases. The Multinomial Naïve Bayes Classifier (MNBC) appeared to be the best option due to its highest Accuracy, Recall, and Specificity values, which amounted to 59.6%, 59.36%, 58.61%, and 58.46%, 59.5%, 68.75%, 76.75%, and 77.07%, and 59.68%, 53.54%, 48.01%, and 47.43% respectively, in descending order relative to the data set (5577 units, 4183 units, 2788 units, and 1394 units).

The study has limitations concerning social media platforms (Facebook) and language (Arabic). The future work is to improve and refine the data mining results: the sample must be expanded by adding more data classes and keywords. Other machine learning models, including deep learning algorithms, may also be used to further investigate the use of machine learning algorithms for mining Arabic texts on social media platforms.

CONFLICT OF INTEREST

The author declares no conflict of interest.

REFERENCES

[1] M. Ganesan and P. Mayilvahanan, "Cyber crime analysis in social media using data mining technique," *Int. J. Pure Appl. Math.*, vol. 116, no. 22, pp. 413–424, 2017.

[2] S. A. Salloum, C. Mhamdi, M. Al-Emran, *et al.*, "Analysis and classification of Arabic newspapers' Facebook pages using text mining techniques," *Int. J. Inf. Technol. Lang. Stud.*, vol. 1, no. 2, pp. 8–17, 2017.

[3] F. Stahl, M. M. Gaber, and M. Adedoyin-Olowe, "A survey of data mining techniques for social media analysis," *J. Data Min. Digit. Humanit.*, 2014.

[4] J. Oliverio, "A survey of social media, big data, data mining, and analytics," *J. Ind. Integr. Manag.*, vol. 3, no. 3, 1850003, 2018.

[5] M. Panda, "Developing an efficient text pre-processing method with sparse generative naive bayes for text mining," *Int. J. Modern Educ. Comput. Sci.*, vol. 11, no. 9, pp. 11–19, 2018.

[6] M. B. Albayati and A. M. Altamimi, "Identifying fake Facebook profiles using data mining techniques," *J. ICT Res. Appl.*, vol. 13, no. 2, pp. 107–117, 2019.

[7] Facebook Investor Relations. (2021). Invesotor.fb. [Online]. Available: https://investor.fb.com/investor-news/press-release-details/2021/Facebook-Reports-First-Quarter-2021-Results/default.aspx

[8] J. A. Obar and S. S. Wildman, "Social media definition and the governance challenge: An introduction to the special issue," *Telecomm. Policy*, vol. 39, no. 9, pp. 745–750, 2015.

[9] K. Domdouzis, B. Akhgar, S. Andrews, *et al.*, "A social media and crowdsourcing data mining system for crime prevention during and post-crisis situations," *J. Syst. Inf. Technol.*, vol. 18, no. 4, pp. 364–382, 2016.

[10] Y. Shibuya, "Mining social media for disaster management: Leveraging social media data for community recovery," in *Proc. 2017 IEEE International Conference on Big Data (Big Data)*, IEEE, 2017, pp. 3111–3118.

[11] T. Spielhofer, R. Greenlaw, D. Markham, *et al.*, "Data mining Twitter during the UK floods: Investigating the potential use of social media in emergency management," in *Proc. 2016 3rd International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, IEEE, 2016, pp. 1–6.

[12] V. Medvedev, O. Kurasova, J. Bernatavičienė, *et al.*, "A new web-based solution for modelling data mining processes," *Simul. Model Pract. Theory*, vol. 76, pp. 34–46, 2017.

[13] G. Bello-Orgaz, J. J. Jung, and D. Camacho, "Social big data: Recent achievements and new challenges," *Inf. Fusion.*, vol. 28, pp. 45–59, 2016.

[14] P. Cavci and B. Das, "Prediction of the customer's interests using sentiment analysis in e-commerce data for comparison of Arabic, English, and Turkish languages," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 35, no. 3, pp. 227–237, 2023.

[15] P. Kaur, A. Goyal, K. Sharma, *et al.*, "Social media in data mining review paper," *Int. J. Adv. Stud. Sci. Res.*, vol. 3, no. 11, pp. 286–289, 2018.

[16] J. Taylor and C. Pagliari, "Mining social media data: How are research sponsors and researchers addressing the ethical challenges?" *Res. Ethics*, vol. 14, no. 2, pp. 1–39, 2018.

[17] D. C. Gkikas, K. Tzafilkou, P. K. Theodoridis, *et al.*, "How do text characteristics impact user engagement in social media posts: Modeling content readability, length, and hashtags number in Facebook," *Int. J. Inf. Manage. Data Insights*, vol. 2, no. 1, 100067, 2022.

[18] H. S. Lathan, A. Kwan, C. Takats, *et al.*, "Ethical considerations and methodological uses of Facebook data in public health research: A systematic review," *Soc. Sci. Med.*, vol. 322, 115807, 2023.

[19] A. S. Bozkır, S. G. Mazman, and E. A. Sezer, "Identification of user patterns in social networks by data mining techniques: Facebook case," in *Proc. International Symposium on Information Management in a Changing World*, Springer, 2010, pp. 145–153.

[20] Y. Al-Saggaf and M. Z. Islam, "Data mining and privacy of social network sites' users: Implications of the data mining problem," *Sci. Eng. Ethics*, vol. 21, no. 4, pp. 941–966, 2015.
[21] V. V. Agarkar, P. E. Ajmire, and P. S. Bodkhe. (2020). Web mining: An application of data mining. [Online]. Available: https://www.academia.edu/45351751/Web_Mining_An_Application_of_Data_Mining?auto=citations&from=cover_page
[22] V. Ermakov, V. Safonov, and D. Dogadkin, "Characteristic features of molybdenum, copper, tungsten and rhenium accumulation in the environment," *Innov. Infrastruct. Solut.*, vol. 6, p. 104, 2021.
[23] V. Safonov, V. Ermakov, V. Danilova, *et al.*, "Relationship between blood superoxide dismutase activity and zinc, copper, glutathione and metallothioneines concentrations in calves," *Biomath.*, vol. 10, no. 2, 2111247, 2021.
[24] A. Kumar and R. K. Singh, "Web mining overview, techniques, tools and applications: A survey," *Int. Res. J. Eng. Technol.*, vol. 3, no. 12, pp. 1543–1547, 2016.
[25] M. N. Doja, "Web usage mining techniques to improve the capabilities of e-learning websites and blogs," *Int. J. Adv. Res. Comput. Sci.*, vol. 8, no. 5, pp. 1847–1851, 2017.
[26] M. K. Gupta and P. A. Chandra, "A comprehensive survey of data mining," *Int. J. Inform. Techn.*, vol. 12, pp. 1243–1257, 2020.
[27] V. Jain, "Perspective analysis of telecommunication fraud using data stream analytics and neural network classification-based data mining," *Int. J. Inform. Techn.*, vol. 9, pp. 303–310, 2017.
[28] A. M. U. D. Khanday, S. T. Rabani, Q. R. Khan, *et al.*, "Machine learning based approaches for detecting COVID-19 using clinical text data," *Int. J. Inform. Techn.*, vol. 12, pp. 731–739, 2020.
[29] M. Zohdi, M. Rafiee, V. Kayvanfar, *et al.*, "Demand forecasting based machine learning algorithms on customer information: An applied approach," *Int. J. Inform. Techn.*, vol. 14, pp. 1937–1947, 2022.
[30] H. H. Patel and P. Prajapati, "Study and analysis of decision tree based classification algorithms," *Int. J. Comput. Sci. Eng.*, vol. 6, no. 10, pp. 4–78, 2018.
[31] S. Fatima and B. Srinivasu, "Text document categorization using support vector machine," *Int. Res. J. Eng. Technol.*, vol. 4, no. 2, pp. 141-147, 2017.
[32] A. R. Susanti, T. Djatna, and W. A. Kusuma, "Twitter's sentiment analysis on GSM services using Multinomial Naïve Bayes," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 15, no. 3, pp. 1354–1361, 2017.
[33] V. Cohen-Addad, V. Kanade, F. Mallmann-Trenn, *et al.*, "Hierarchical clustering: Objective functions and algorithms," *Journal of the ACM (JACM)*, vol. 66, no. 4, pp. 1–42, 2019.
[34] M. A. Al-Hagery, M. A. Alzaid, T. S. Alharbi, *et al.*, "Data mining methods for detecting the most significant factors affecting students' performance," *Int. J. Inf. Technol. Comput. Sci.*, vol. 5, pp. 1–13, 2020.
[35] Y. Yang, J. H. Hsu, K. Löfgren, *et al.*, "Cross-platform comparison of framed topics in Twitter and Weibo: Machine learning approaches to social media text mining," *Soc. Netw. Anal. Min.*, vol. 11, no. 1, pp. 1–18, 2021.
[36] A. Elnagar, R. Al-Debsi, and O. Einea, "Arabic text classification using deep learning models," *Inf. Proces. Manag.*, vol. 57, no. 1, 102121, 2020.
[37] M. Elhoseny, A. Abdelaziz, A. S. Salama, *et al.*, "A hybrid model of Internet of things and cloud computing to manage big data in health services applications," *Future Gener. Comput. Syst.*, vol. 86, pp. 1383–1394, 2018.
[38] A. M. Almeida, R. Cerri, E. C. Paraiso, *et al.*, "Applying multi-label techniques in emotion identification of short texts," *Neurocomputing*, vol. 320, pp. 35–46, 2018.
[39] A. Bayhaqy, S. Sfenrianto, K. Nainggolan, *et al.*, "Sentiment analysis about e-commerce from tweets using decision trees, K-nearest neighbor, and Naïve Bayes," in *Proc. 2018 International Conference on Orange Technologies (ICOT)*, IEEE, 2018, pp. 1–6.
[40] M. Awad and R. Khanna, "Support vector machines for classification," *Efficient Learning Machines*, 2015, pp. 39–66.
[41] M. Z. Hossain, M. N. Akhtar, R. B. Ahmad, *et al.*, "A dynamic K-means clustering for data mining," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 13, no. 2, pp. 521–526, 2019.
[42] M. D. Vicario, W. Quattrociocchi, A. Scala, *et al.*, "Polarization and fake news: Early warning of potential misinformation targets," *ACM Trans. Web*, vol. 13, no. 2, pp. 1–22, 2019.
[43] J. H. Seah and K. J. Shim, "Data mining approach to the detection of suicide in social media: A case study of Singapore," in *Proc. 2018 IEEE International Conference on Big Data (Big Data)*, IEEE, 2018, pp. 5442–5444.
[44] A. Y. Muaad, G. H. Kumar, J. Hanumanthppa, *et al.*, "An effective approach for Arabic document classification using machine learning," *Global Transit. Proceed.*, vol. 3, no. 1, pp. 267–271, 2022.