DOI: 10.15849/IJASCA.230720.10

Int. J. Advance Soft Compu. Appl, Vol. 15, No. 2, July 2023
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Harnessing Machine Learning for Arabic COVID-19 Omicron News Classification: A Comparative Study

Abdulkareem Alzahrani

Computer Engineering and Science Department,
Faculty of Computer Science and Information Technology,
Al-Baha University, Al-Baha, 65779, Saudi Arabia
e-mail: ao.alzahrani@bu.edu.sa

Abstract

The exponential growth of web-based content has prompted governments and businesses to seek new strategies and tools for effective governance, navigation, and management of complex data. Since the start of SARS-CoV-2 (COVID-19), there have been numerous news related to coronavirus and its variants published on various platforms. Thus, it is important to introduce an intelligent classification technique for classifying the news related to COVID-19. Automated news classification refers to the process of categorising news articles into predetermined categories based on their content, using the confidence gained from the training news dataset. Therefore, this paper intends to develop a model capable of classifying Arabic news related to the COVID-19 Omicron variant. The model can classify related news into five classes: statistics, vaccines, restrictions, economy and health. The news corpus is comprised of 220,000 instances gathered from various Arabic newspapers. The dataset is pre-processed, and then the n-gram features are identified to train and test the models. This study experimented with various machine learning (ML) classifiers, i.e., Multinomial Naïve Bayes (MNB), Support Vector Machine (SVM), and Random Forest (RF) with various n-grams when creating the model. The results show that MNB with character n-gram outperformed baseline methods in terms of recall, precision and F1-score with average values of 86%, 87% and 86%, respectively.

Keywords: *Natural Language Processing, Machine Learning, Arabic Dataset, Arabic Text Classification.*

1 Introduction

Coronavirus SARS-CoV-2 (COVID-19) is an infectious disease that has spread around the world, affecting medical systems and human health [1]. A number of COVID-19 variants have been discovered, including Alpha [2], Beta [3], Gamma [4], Delta [5] and Omicron [6, 7]. The last was first discovered in South Africa in November 2021 [6]. According to the World Health Organization (WHO), there have been 762,201,169 confirmed cumulative cases of COVID-19 globally, including 6,893,190 cumulative deaths ‘up to the 12th of April 2023’ [8].

Received 29 January 2023; Accepted 22 May 2023

Recent statistics (from March 6 to April 2, 2023) show that there were ≈ 3.3 million new cases and >23,000 deaths reported globally. This represents a reduction of 28% in new cases and 30% in deaths compared to the preceding 28-day period (from February 6 to March 5, 2023). Although there has been a general decrease in new cases and deaths globally, it's worth noting that 31% of countries (74 in total) have reported a rise in new cases of 20% or more during the last 28 days compared to the preceding 28-day period. As of April 2, 2023, there have been more than 762 million confirmed cases, and over 6.8 million deaths reported worldwide [9].

This rampant inflation of the globally infectious disease has led to an unprecedented surge in the number of published newspaper articles and other news items on the subject. These articles and news have reported statistics, announced the restrictions and vaccines, and studied the impact of the disease in different domains, such as education, health, politics, and economy. Hence, such a dramatic increase in publications has made the individuals facing issues in following the publications they are looking for and suit their individual interests. This is where an intelligent model capable of classifying this flood of publications can prove to be immensely valuable and crucial. This can be accomplished by identifying patterns in the coverage. For this, it is valuable to construct a model that can contribute to categorising these articles and news items through the adoption of natural language processing (NLP).

As the name suggested, NLP is a sub-field of artificial intelligence (AI) that focuses on enabling computers to process and analyse natural languages in written, spoken and sign forms. Thus, computers can understand, interpret and generate human languages. In addition, various tasks can be fulfilled through the use of NLP, such as sentiment analysis, name entity recognition, information extraction, conversational agents and chatbots, text summarisation, machine translation, question-answering and text classification [10].

The field of NLP holds great importance in today’s world, owing to its practical applications across various industries such as healthcare [11, 12], education field [13, 14], marketing [15, 16], finance [17], and customer service [18]. Yet, NLP is faced with a significant challenge due to the innate ambiguity of natural language. The meaning of words can vary considerably based on the context in which they are used. Thus, the interpretation of sentences differs significantly depending on the syntax and semantics of the language being employed [10]. Accurately analysing and comprehending natural language data can, therefore, pose a challenge and it remains an area of ongoing research for practitioners in the field. This includes exploring the utilisation of various techniques such as statistical models [19], rule-based systems [20], ML [21, 22], deep learning [23], and hybrid techniques [24]. These techniques allow computers to process and examine extensive quantities of data, enabling them to identify connections and patterns present within natural language usage. This in turn, has led to significant breakthroughs in NLP and made the models more accurate.

Building on the previous points, it is important to consider exploring NLP techniques for classifying COVID-19 news and articles, particularly in languages that face issues, such as Arabic. These issues include the ambiguities and the dialects variation [25], the richness of the morphology [26] and the lack of resources [27], as discussed by Manal Mostafa Ali [28].

To the best of the author’s knowledge, most studies use social media platforms to classify or investigate the public’s sentiments [29–31]. However, there seems to be a lack of use for the news and articles published in formal newspapers.

In [32], the paper conducted an analysis of the dissemination of both factual and false information on Twitter from the time period spanning 2006 to 2017. The findings revealed that false information exhibited a considerably greater reach, speed, depth, and overall impact across all categories when compared with factual information. Furthermore, the study demonstrated that robots played an equal role in disseminating true and false news, indicating that human beings are the primary drivers of the spread of false information.

Hence, it can be contended that false news and rumours disseminate more effortlessly on social media, in contrast to news disseminated through traditional newspapers. Professional newspapers maintain high standards to ensure the accuracy and quality of their published articles. Furthermore, several countries have put in place stringent regulations to prevent the dissemination of rumours through newspapers.

The aim of this paper is to create a corpus of Arabic news and articles related to Omicron, which was scraped from opted Arabic Gulf newspapers. Another objective is to develop a model capable of classifying the news into one of five categories: statistics, vaccines, restrictions, economy, and health. To achieve this, a comparative analysis is carried out utilising recall, precision, F1 score, and accuracy between three well-known ML classifiers (i.e., MNB, SVM, and RF).

The remaining sections of the paper are organised as follows. Section 2 surveys studies related to the paper's scope. Section 3 explains the research methodology followed. The experiment is discussed in Section 4, and the results are presented and discussed in Section 5. Finally, the conclusion is presented in Section 6.

2 Related Work

The Omicron variant of the COVID-19 infection (B.1.1.529) was first recognised in South Africa, and it was viewed as a worrying variant by the WHO on the 26th of November 2021 [33]. From that point forward, it spread around the world. By mid-January, it was the most overwhelming disturbance on earth, causing an impressive expansion in COVID-19 cases. In numerous nations, the Omicron variant represented a resurgence of the pandemic, upsetting the pattern of diminishing quantities of COVID-19 cases and passing.

Many models have been proposed to depict the elements of the scourge. Some of them can be characterised into two classes: aggregate models [34] and network-based models [35]. A few examinations have researched how inoculation and non-pharmacological methodologies can affect the course of the illness [36].

In [37], the utilisation of a straightforward numerical model was proposed in light of the traditional SIRD model to change and anticipate the COVID-19 pandemic's conduct in three countries: Brazil, Italy and Korea, which are instances of altogether different situations and stages of the COVID-19 pandemic. The model utilised in this work is additionally founded on the work-of-art, compartmental SIRD model and expands the models proposed in [38].

In [39], the authors employ NLP and a supervised k-nearest neighbours (KNN) classification algorithm to conduct sentiment analysis on public discourse regarding approved COVID-19 vaccines on Twitter. Their analysis offers insights into the proportion of (positive, negative or neutral) tweets concerning the Pfizer, Moderna, and AstraZeneca vaccines, thereby providing valuable information on public sentiments towards these vaccines. They utilise the 'Tweepy' library to retrieve data from Twitter, which is then saved in CSV format. In order to ensure optimal quality, the data is pre-processed through the elimination of special characters, hyperlinks, retweets, emojis, and stickers. This meticulous approach serves to enhance the overall accuracy and reliability of the information being analysed. After the pre-processing stage, their approach involves implementing a tokenisation technique to accurately classify the data into distinct categories (i.e., positive, negative, or neutral) based on the respective sentiment. According to the gathered data, Pfizer garnered the most favourable feedback on social media platforms, with Moderna and AstraZeneca following closely behind. It is noteworthy that Moderna and AstraZeneca received the highest percentage of negative tweets. Additionally, the study presents comprehensive word clouds and tables that demonstrate the most frequently used words and sentiment scores for each vaccine.

In [40], the present study endeavours to assess the perceptions and experiences of individuals who visited Bangkok during the COVID-19 lockdown in Thailand. Twitter was employed as the primary source for obtaining the opinions and viewpoints of tourists throughout the period of nationwide lockdown between April 3 and 30, 2020. The insights garnered from this research are significant as they offer a better understanding of the sentiments and perspectives of some travellers regarding the tourism industry. Their findings could be used to furnish recommendations to stakeholders involved in tourism. The data was first subjected to pre-processing, wherein non-tourism-related information was removed. Subsequently, the data were labelled based on their sentiment, i.e., positive, negative, or neutral. To undertake this classification, they employed the support vector machine algorithm (SVM) for sentiment analysis. Moreover, they established specific criteria for accurately distinguishing between each sentiment class. The results witness that the SVM classified tweet sentiment with 71.03% accuracy, 81.65% precision, 55.14% recall, and 0.58 F-1 score. Furthermore, 29 neutral tweets out of 61 were accurately predicted, resulting in a 47.54% accuracy rate. However, the study only employs one sentiment analysis algorithm, and the outcomes may differ with the use of other algorithms.

In [41], a method is proposed to identify the major concerns and trending topics related to the COVID-19 pandemic, according to Twitter users. The authors suggested a framework that utilises a combination of clustering algorithms and transfer learning to categorise similar tweets based on their semantic information and meaning. They employed several methods to analyse Twitter data. Initially, the Twitter API was utilised to gather relevant tweets, and the data was subsequently cleaned to eliminate irrelevant information. The Universal Sentence Encoder was then utilised to extract the feature representation of the sentences within the tweets. Then, the K-means clustering algorithm was utilised for grouping similar tweets based on their semantic information and meaning. Finally, a text summarisation algorithm based on deep learning was employed to generate a concise summary of the cluster and to determine the primary topics within each cluster. The proposed framework showed high effectiveness in identifying trending topics related to COVID-19 on Twitter. In addition, the authors analysed the coherence and informativeness of their work and compared it with other methods (i.e. Inverse Document Frequency (TF-IDF) and Latent Dirichlet Allocation (LDA)). The results witnessed that their approach outperformed the other two methods. Furthermore, the framework provides a valuable tool for gaining a deeper understanding of people's concerns and opinions about COVID-19, which can be extended to other social media platforms and contexts to detect trending topics and opinions of interest. One aspect that needs to be considered is the potential impact of irrelevant and noisy information found in the data collected from Twitter on the accuracy of the proposed approach.

The work in [42] presents a novel approach to identifying anomalous events related to COVID-19 by analysing a vast amount of real-time tweets. The proposed methodology employs a distributed Directed Acyclic Graph topology and a lightweight algorithm to identify such events automatically. Furthermore, the system is capable of grouping and presenting significant keywords identified in the tweets. The related experiment was run to compare single and dual machines during the period of 1st to 30th August 2020. It involved 15 tests, with each machine being tested 15 times for a total of 30 days. The results of the study showed that the dual machine outperformed the single machine, with the distributed machine reaching its peak on the 10th test by generating 3372 tweets, compared to the standalone machine, which generated 1654 tweets. As a result, the study concludes that the proposed framework is effective in detecting anomalous events associated with COVID-19 from large-scale real-time tweets related to COVID-19. The study is subject to two limitations. Firstly, it solely relies on Twitter data, which may include extraneous or erroneous information. Secondly, a comprehensive evaluation of the proposed algorithm's effectiveness in identifying anomalous events needs to be included in the paper.

An approach is developed in [43] to analyse public discussions on Twitter regarding the Centres for Disease Control and Prevention (CDC) and their response to COVID-19. For the study, data from Twitter was analysed using the latent Dirichlet allocation algorithm. The tweets were obtained from a COVID-19 Twitter chatter dataset spanning from March 11, 2020, to August 14, 2020. The tweets were cleaned using R, keeping only those containing any of five specific keywords related to the Centres for Disease Control and Prevention (CDC). Ninety-one tweets posted by the CDC itself were excluded. The final dataset analysed contained 290,764 unique tweets from 152,314 different users. Only tweets with English specified in their metadata language field were tokenised using the gsub function in R. The study identified 16 topics that the public associated with the CDC. These topics were further categorised into four main themes: government policies and actions, public perception of the CDC's credibility, understanding of the virus and the situation, and response guidelines. The paper focuses on attention and expectations from the CDC, and provides insight into public concerns and perceptions of the CDC's performance. The analysis of 290,764 tweets revealed that the COVID-19 death counts garnered the most attention, constituting 12.16% (n=35,347) of the total. Additionally, there was a significant amount of discourse relating to the credibility of the CDC and other governing bodies, as well as their respective COVID-19 guidelines, with more than 20,000 tweets each. These findings suggest that there is a high level of interest and concern regarding COVID-19, its impact, and the strategies being implemented to combat it. The paper presents two notable limitations. Firstly, there is a likelihood that private account tweets were not accounted for during the data collection process, and tweets created by bots or fake accounts may have been overlooked. Secondly, while the study successfully identified topics from the public discourse regarding the CDC, it did not delve into the changes in public attention over time or in relation to specific circumstances.

A recent research paper delves into the public conversation on social media regarding mask-wearing in the United States and its connection to the increasing number of COVID-19 cases [44]. The study collected 51,170 tweets in English, spanning from January 1, 2020, to October 27, 2020. The data was obtained by searching for hashtags that opposed mask-wearing. The tweets underwent pre-processing, which involved the removal of stop words, keywords with IDs, and hashtags. The remaining content was represented by bi- and trigrams. The latent Dirichlet allocation (LDA) algorithm was leveraged to analyse the pre-processed tweets and determine the primary topics or categories addressed in the posts. According to the findings of the paper, social media can play a crucial role in identifying significant insights regarding mask-wearing. Through topic mining, ten categories or themes of user concerns were discovered, with the top three being 1) personal freedom and constitutional rights, 2) big pharma, population control and conspiracy theories, and 3) the fakeness of pandemic, numbers and news. These three categories make up almost 65% of the tweets that oppose wearing masks. The study also revealed a strong correlation between the volume of negative tweets against mask-wearing and the number of newly reported COVID-19 cases. The increase in negative tweets preceded the rise in new cases by nine days. The study's findings should be interpreted with certain limitations in mind. While the data suggests a correlation between an increase in anti-mask sentiment on Twitter and COVID-19 cases, it cannot definitively establish causation. Furthermore, the analysis only examined English-language tweets from the United States, and additional research is needed to compare discourse across diverse platforms and countries. It should also be noted that the number of tweets collected may not be fully representative of public opinion regarding mask usage.

On the other hand, in terms of analysing articles that are published in Arabic newspapers, Qadi et al. conducted a study that aims to create a model capable of categorising news articles written in Arabic into four categories, namely: the Middle East, sports, business and technology. To reach their aim, they build a dataset containing 90,000 Arabic news articles. The articles are associated with an individual related label (category), and then the unneeded characters, such as numbers, stopping words, Latin characters and punctuation, are cleaned. They examine how effective the dataset is at using ten various well-known classifier algorithms. These algorithms include Nearest Centroid, KNN, logistic regression, SVMs, Decision Tree (DT), RF classifier, XGBoost classifier, Ada–Boost classifier, Multilayer Perceptron and multinomial classifier. In addition, they utilise the majority voting classifier, which is an ensemble learning technique to achieve the highest possible accuracy amongst the ten used classifiers. The model is tested via the confusion matrix, and the ten classifiers are compared with each other. SVM achieves the highest accuracy rate (94.4%) among the other nine classifiers. Moreover, the accuracy of the majority voting technique achieves an accuracy near that of SVM. Thus, this last may be the best choice since it outperforms the majority voting in terms of performance [45].

The surveyed literature expresses the feasibility of utilising ML to classify news articles regardless of their origin (i.e., social media or newspapers). However, there are insufficient studies on the implementation of NLP in Arabic news regarding Omicron, which have been published in formal Arabian Gulf newspapers. Hence, this study aims to bridge the highlighted research gap. This can be achieved by building an intelligent model that facilitates accurate categorisation of the enormous number of Arabic news articles on Omicron.

3 Methodology of the Proposed Work

To accomplish the goal of this research, it is crucial to pinpoint the specific newspapers being targeted, define the relevant Arabic keywords related to 'Omicron' in the articles, and establish the proper procedures for annotation. To ensure effective outreach, it is essential to carefully consider the circulation, reputation, and online archive accessibility of the newspapers targeted for publication. Furthermore, it is important to identify keywords related to the Omicron variant, such as ‘أوميكرون’ or ‘Omicron’. To ensure optimal linguistic analysis of gathered articles, it is recommended to comply with a standardised annotation framework encompassing both semantic and syntactic elements. Adherence to such a scheme can facilitate accurate and consistent analysis, promoting greater efficiency and reliability. Afterwards, the collected text should be pre-processed using standard NLP techniques like tokenisation, lemmatisation, stop words removal, etc. The n-gram features ranging from [2–4] characters to 'word' should be extracted. The dataset will be used to train ML models for text classification to analyse the public sentiment and stances regarding the Omicron variant in the opted newspapers. The performance of different models will be evaluated and discussed. Fig. 1 depicts the workflow of the proposed work. In addition, the workflow is explained in the following sections.

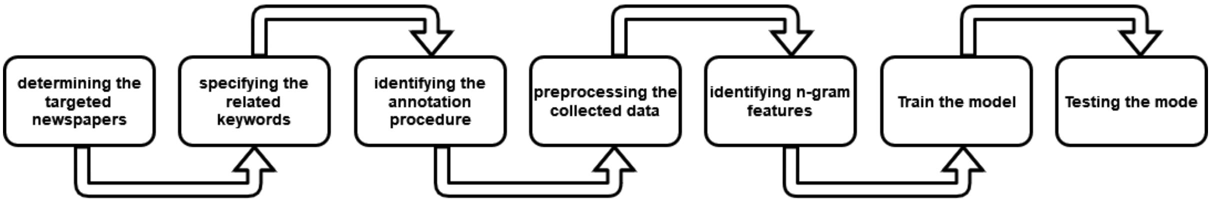

determining the targeted newspapers
specifying the related keywords
identifying the annotation procedure
preprocessing the collected data
identifying n-gram features
Train the model
Testing the mode

Figure 1: The workflow of the proposed work

3.1 Data Collection

The data for this study was collected from various Arabian Gulf newspapers: Al Riyadh (Saudi Arabia), Al Bilad (Bahrain), An-Nahar and Al-Alanba (Kuwait), Alwatan (Oman), Al Sharq (Qatar) and Al Watan (United Arab Emirates). The fastest way to find articles written about Omicron was to use a specific keyword in the search field: ‘أوميكرون’ or ‘Omicron’. After that, all related articles on each search page were crawled.

As shown in Table 1, each newspaper had different data (tags) for each item. Therefore, it is necessary to standardise the tag for each news item by choosing the most crucial tags, such as URL, title, article or date. Hence, the dataset contained more than 220,000 words from different newspapers.

Table 1: Dataset Tags

Country	Newspaper	URL	Tag
Saudi Arabia	Al Riyadh	alriyadh.com	Title, Title_URL Location, Author Article(Body), Date, Category
Bahrain	Al Bilad	albiladpress.com	Title_URL, short_description Title, Article(Body), Date
Kuwait	An-Nahar	annaharkw.com	URL, Date Title, Article(Body)
	Al-Alanba	alanba.com.kw	URL, Date, Title Article(Body), Source
OMAN	Alwatan	alwatan.com	URL, Date, Title Article(Body)
Qatar	Al Sharq	al-sharq.com	Title, Title_URL Article(Body) Location, Date
UAE	Al Watan	alwatan.ae	URL, Date, Title Category, Article(Body)

3.2 Annotation

This paper considers classifying the article based on its title. Therefore, for each newspaper, ten titles were randomly chosen, along with their related article bodies. This step was taken to determine the category that describes the articles. Hence, in scanning the collected articles, it was found that each article could be classified into one of the five categories (statistics, vaccines, restrictions, economy and health) as described in Table 2. These five categories were specifically chosen based on their significance to the variant as well as their potential impact on the wider community. The native annotators were requested to evaluate the list of categories and share their thoughts on them. They collectively concurred that the five categories were fitting and inclusive enough to encompass the fundamental elements of the published articles. By conducting a validation step, the researcher ensured that his selection of categories truly represented the perspectives of native Arabic speakers and was not influenced by his own biases. This step enhanced the credibility of the category selection and improved the overall methodology and findings of the study.

On the other hand, the annotation was completed manually by a native speaker using Table 2 and the source of information. A second annotator annotated more than 60 random articles and titles to measure inter-annotator agreement. It has been shown using Cohen's Kappa that there was a robust agreement of 80% for the tag category.

Table 2: Annotation categories and descriptions.

Category	Description
Statistics	Number of infections, recovery cases and deaths
Vaccine	Information about vaccines and PCR
Restrictions	Quarantine/open–close events, football matches and flights
Economy	News related to oil and the economy
Health	Health information about Omicron and symptoms

In addition, based on the given distribution presented in Fig. 2, the majority of the articles in the dataset are related to the health category at 35%, followed by the economy category at 22%. The statistics category is the third most common with 19%, while the vaccines and restrictions categories appear less frequently, accounting for 13% and 11% of the articles, respectively. The high percentage of articles related to health suggests that the public was highly concerned about the impact of the Omicron variant on public health. This is not surprising given that the Omicron variant is highly transmissible, and there was a lot of speculation about its potential impact on public health. The significant percentage of articles related to the economy suggests that the public was also concerned about the economic impact of the Omicron variant. This is likely because the Omicron variant's spread has resulted in many businesses shutting down or operating at reduced capacity, leading to job losses and financial insecurity. The relatively small percentage of articles related to vaccines and restrictions may suggest that the public had a relatively good understanding of vaccination strategies and government-imposed restrictions at the time the articles were published. It may indicate that these topics were not as high of a concern for the public as health and economic impacts.

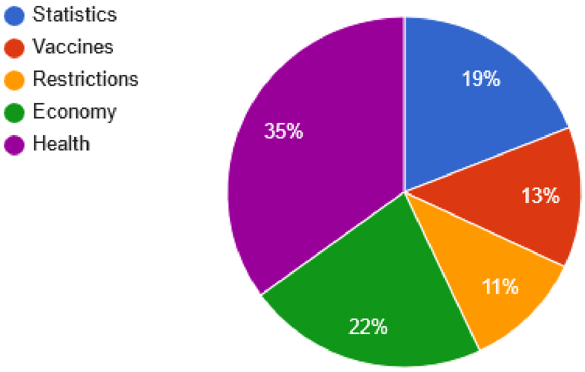


Figure 2: Category distribution

3.3 Pre-processing

Text pre-processing is a crucial step for text classification, and it is strongly recommended when dealing with text collected from the web. Hence, pre-processing can reduce errors and enhance the accuracy of the classification [46]. Text pre-processing involves tokenisation, normalisation, and diacritics removal for Arabic texts. In this study, CAMeL Tools was employed to process the text by removing diacritics (such as ً , ْ ٌّ), removing punctuation and unifying different forms of the same letter (normalisation), for example, normalising 'alef', 'alef marbuta' and 'alef maksura' [47].

4 Experiment

After gaining certainty of the sufficiency and effectiveness of the collected dataset, the experiment was run in various configurations. This included the adopted n-gram features and the chosen classifiers. Hence, for each configuration, the performance measures (i.e., recall, precision, F1 score and accuracy) were utilised for the sake of the comparison.

4.1 N-gram Features

An n-gram is a sequence of consecutive symbols extracted from a long string. A symbol can be a character or a word [48]. Unigram refers to n-grams of length one. Bigram refers to n-grams of length two. Trigram refers to n-grams of length three, etc. N-grams of texts are extensively used in text mining and NLP tasks. Therefore, the n-gram methods were applied based on words and [2–4] characters in the experiment.

4.2 Classification Algorithms

Several ML classification algorithms have been performed to classify Arabic text. This experiment utilised three of the commonly used ML classifiers, namely the MNB classifier, SVM classifier and RF classifier.

4.2.1 MNB Classifier

MNB classifier is acceptable for discrete feature classification (e.g. word counts for text characterisation). The multinomial distribution typically requires whole-number element counts. In machine learning, the MNB algorithm is used in NLP. The algorithm depends on the Bayes theorem and identifies the tag of a text, such as a piece of newspaper article or email. The MNB algorithm works with categorical input variables, as compared to numerical variables [49]. It is efficient at providing forecasting and identification data based on historical results. Some popular utilisations of MNB are for sentimental analysis, spam filtration and article classification.

4.2.2 SVM Classifier

SVM is a special linear classifier that depends on the margin maximisation rules. It implements structural risk minimisation, which works on the intricacy of the classifier, fully intent on accomplishing great generalisation execution [50]. SVM represents a set of supervised learning methods used for detection, classification, and regression. It is efficient for high-dimensional spaces but still viable in situations where the number of aspects is more noteworthy than the number of tests. There are many utilisations of SVM, including texture classification, inverse geo-sounding problems and speech recognition.

4.2.3 RF Classifier

RF is widely recognised as one of the most efficient and reliable classification algorithms in the field. Its robustness in handling both regression and classification tasks, even in the presence of large datasets, has been shown to yield highly accurate results. It is a ML model in which the defined DTs are determined in training time and outcomes of the predicted model by the individual trees [51]. RF can limit overfitting without increasing error to bias. Thus, it is a very powerful method [52].

5 Results and Discussions

The rotation validation (k-fold cross-validation) technique was adopted and set in ten folds.

After running the experiment, the results showed the effectiveness of character n-gram [2–4] over word n-gram in terms, particularly when utilising MNB.

Table 3 and Table 4 show the precision, recall, F1 score and accuracy for each classifier using 10-fold cross-validation. Accuracy is almost the same as the F1 score.

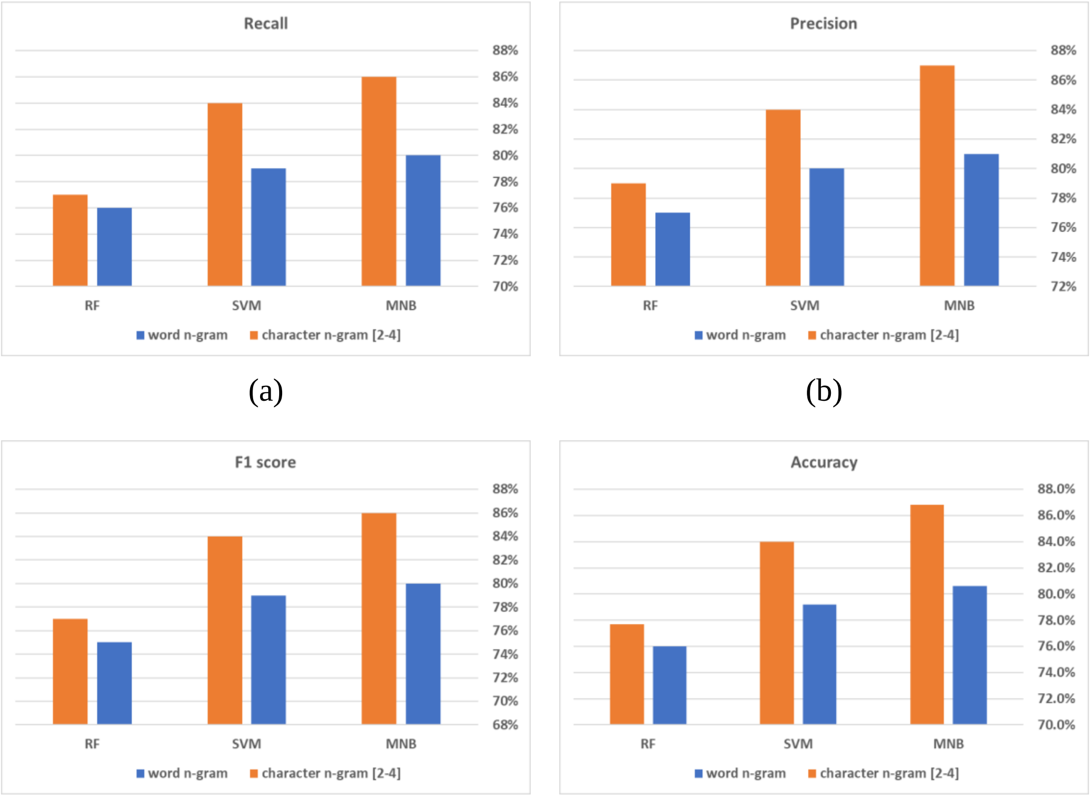
Table 3: The experiment results for Word N-Gram.

	Recall	Precision	F1 Score	Accuracy
MNB	80%	81%	**80%**	80.6%
SVM	79%	80%	79%	79.2%
RF	76%	77%	75%	76.0%

Table 4: The experiment results for Character N-Gram [2-4].

	Recall	Precision	F1 Score	Accuracy
MNB	86%	87%	**86%**	86.8%
SVM	84%	84%	84%	84.0%
RF	77%	79%	77%	77.7%

The RF produced a lower F1 score than the MNB and SVM, whereas the other two classifiers generated good results between 79% and 80% for word n-gram. In addition, using the character n-gram [2–4] improved the performance of the MNB by 6%. The MNB classifier achieved the best performance using the character n-gram, as shown in Figure 3.



(a) (b)

(c) (d)

Figure 3: The results of (a) recall, (b) precision, (c) F1 score, and (d) accuracy for MNB, SVM and RF in word n-gram and character n-gram

There are multiple factors that could have contributed to the presented results.

1- When facing the challenge of analysing the Arabic language, characterised by intricate vocabularies and numerous morphological variations, the usage of character n-grams can prove to be more effective than word n-grams in capturing crucial linguistic features. This approach allows the model to better comprehend the subtle intricacies of the language and enhance its overall performance.
2- The MNB algorithm has demonstrated efficacy in handling high-dimensional text data, although it is predicated on the assumption that features are independent, which may not always hold true in real-world applications. Nevertheless, this algorithm frequently yields satisfactory results. In contrast, SVM and RF represent more intricate models that may necessitate greater amounts of data and hyperparameter tuning to achieve optimal performance.

The results showed that the aim of this study has been achieved. Hence, through the proposed model, individuals can reach their desired and required publications (i.e., news and articles). For instance, an economist may be inclined towards articles that discuss the economic consequences of the pandemic, whereas an individual working in healthcare may find articles concerning the medical facets of the pandemic more relevant. Furthermore, the model can aid policymakers and researchers in comprehending the scope and characteristics of the pandemic and its impact on various domains. For example, through the examination of news coverage and its classification, researchers and policymakers can detect areas that have not received enough attention and might require further investigation or resources.

6 Conclusion and Future Directions

This paper has contributed to creating a textual dataset for Arabic articles written in formal Arabian Gulf newspapers about the Omicron variant. Another contribution involves the development of a model capable of effectively categorising news articles into five distinct domains: statistics, vaccines, restrictions, economy, and health. The paper highlighted the methodology for creating the dataset and the opted techniques and algorithms (e.g., character n-gram, word n-gram, MNB, SVM and RF). The results showed the effectiveness of the followed methodology and how MNB outperformed the other classifiers. Moreover, the adoption of character n-gram [2–4] succeeded with a high F1 score as compared to word n-gram (>7% when using MNB, >5% when using SVM and >2% when using RF). The highest achieved accuracy in the whole experiment was 86.8% for MNB with character n-gram, and that was a good indication of the efficiency of the proposed model. The created model could prove invaluable in keeping individuals informed and up to date regarding the latest developments in these critical areas.

On the other hand, it is worth noting that the results of this study may not be generalisable to all Arabic-speaking populations. This is because the dataset used in this study was collected from Arabic newspapers in the Gulf region, where there are many non-Arabic speakers living. As a result, the news articles may not entirely reflect the perspectives and concerns of the entire population (foreigners as well as citizens). Furthermore, the study only evaluated three ML classifiers, which may not be entirely representative of all possible classifiers. Future research could explore other classifiers to improve the accuracy of classification and ensure the robustness of the findings.

In future work, it might be possible to collect non-Arabic articles on Covid-19 and analyse their most common topics, then compare the analysis with what has been achieved in this paper. Another direction is to do what has been done in this paper with other COVID-19 variants or other epidemics, and then inspect the possibility of generalising the methodology of this study. Furthermore, future research could explore other classifiers to improve the accuracy and robustness of the classification results. This could be done by comparing the performance of the classifiers used in this study with other ML algorithms or by employing deep learning models. By exploring different classifiers, it may be possible to achieve even higher accuracy in classifying Arabic news articles related to the Omicron variant.

References

[1] Tian, D., Sun, Y., Xu, H., & Ye, Q. (2022). The emergence and epidemic characteristics of the highly mutated SARS-CoV-2 Omicron variant. Journal of Medical Virology, 94(6), 2376–2383. https://doi.org/10.1002/jmv.27643

[2] Domingo, P., & Benito, N. de. (2021). Alpha variant SARS-CoV-2 infection: How it all starts. eBioMedicine, 74. https://doi.org/10.1016/j.ebiom.2021.103703

[3] Reincke, S. M., Yuan, M., Kornau, H.-C., Corman, V. M., van Hoof, S., Sánchez-Sendin, E., … Kreye, J. (2022). SARS-CoV-2 Beta variant infection elicits potent

lineage-specific and cross-reactive antibodies. Science, 375(6582), 782–787. https://doi.org/10.1126/science.abm5835

[4] da Silva, J. F., Esteves, R. J., Siza, C., Soares, E. P., Ramos, T. C., Campelo, E. C., … Naveca, F. (2022). Cluster of SARS-CoV-2 Gamma Variant Infections, Parintins, Brazil, March 2021. Emerging Infectious Diseases, 28(1), 262–264. https://doi.org/10.3201/eid2801.211817

[5] Mlcochova, P., Kemp, S. A., Dhar, M. S., Papa, G., Meng, B., Ferreira, I. A. T. M., … Gupta, R. K. (2021). SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. Nature, 599(7883), 114–119. https://doi.org/10.1038/s41586-021-03944-y

[6] He, X., Hong, W., Pan, X., Lu, G., & Wei, X. (2021). SARS-CoV-2 Omicron variant: Characteristics and prevention. MedComm, 2(4), 838–845. https://doi.org/10.1002/mco2.110

[7] Duong, D. (2021). Alpha, Beta, Delta, Gamma: What’s important to know about SARS-CoV-2 variants of concern? CMAJ, 193(27), E1059–E1060. https://doi.org/10.1503/cmaj.1095949

[8] WHO Coronavirus (COVID-19) Dashboard. (2023, December 4). World Health Organization. Retrieved April 12, 2023, from https://covid19.who.int

[9] Weekly epidemiological update on COVID-19 - 6 April 2023. (n.d.). Retrieved April 12, 2023, from https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---6-april-2023

[10] Chowdhary, K. R. (2020). Natural Language Processing. In K. R. Chowdhary (Ed.), Fundamentals of Artificial Intelligence (pp. 603–649). New Delhi: Springer India. https://doi.org/10.1007/978-81-322-3972-7_19

[11] Patra, B. G., Sharma, M. M., Vekaria, V., Adekkanattu, P., Patterson, O. V., Glicksberg, B., … Pathak, J. (2021). Extracting social determinants of health from electronic health records using natural language processing: a systematic review. Journal of the American Medical Informatics Association: JAMIA, 28(12), 2716–2727. https://doi.org/10.1093/jamia/ocab170

[12] Morin, O., Vallières, M., Braunstein, S., Ginart, J. B., Upadhaya, T., Woodruff, H. C., … Lambin, P. (2021). An artificial intelligence framework integrating longitudinal electronic health records with real-world data enables continuous pan-cancer prognostication. Nature Cancer, 2(7), 709–722. https://doi.org/10.1038/s43018-021-00236-2

[13] Alzahrani, A., Alzahrani, A., Al Arfaj, F. K., Almohammadi, K., & Alrashidi, M. (2015). AutoScor: An Automated System for Essay Questions Scoring. aut, 2(5), 182–187.

[14] Alzahrani, A., Alzahrani, A., Alarfaj, F., Almohammadi, K., & Alrashidi, M. (2014). An Automated Scoring Approach For Essay Questions. In International Conference on Education in Mathematics, Science and Technology (pp. 488–492). Konya, Turkey.

[15] Rizwan, M., Rehman, S., Nawaz, A., Ali, T., Imran, A., Alzahrani, A., & Almuhaimeed, A. (2023). Aspect-Based Sentiment Analysis for Social Multimedia:

A Hybrid Computational Framework. Computer Systems Science and Engineering, 46(2), 2415–2428. https://doi.org/10.32604/csse.2023.035149

[16] Kanwal, B., Rehman, S. U., Imran, A., Shaukat, R. S., Li, J., Alzahrani, A., … Alarfaj, F. K. (2023). Opinion Mining from Online Travel Reviews: An Exploratory Investigation on Pakistan Major Online Travel Services Using Natural Language Processing. IEEE Access, 1–1. Presented at the IEEE Access. https://doi.org/10.1109/ACCESS.2023.3260114

[17] Kwak, W., Shi, Y., & Lee, C. F. (2019). Data Mining Applications in Accounting and Finance Context. In Handbook of Financial Econometrics, Mathematics, Statistics, and Machine Learning (pp. 823–857). WORLD SCIENTIFIC. https://doi.org/10.1142/9789811202391_0021

[18] Tian, X., Vertommen, I., Tsiami, L., van Thienen, P., & Paraskevopoulos, S. (2022). Automated Customer Complaint Processing for Water Utilities Based on Natural Language Processing—Case Study of a Dutch Water Utility. Water, 14(4), 674. https://doi.org/10.3390/w14040674

[19] Silverman, G. M., Sahoo, H. S., Ingraham, N. E., Lupei, M., Puskarich, M. A., Usher, M., … Pakhomov, S. V. (2021). NLP Methods for Extraction of Symptoms from Unstructured Data for Use in Prognostic COVID-19 Analytic Models. Journal of Artificial Intelligence Research, 72, 429–474. https://doi.org/10.1613/jair.1.12631

[20] Mumtaz, R., & Qadir, M. A. (2022). CustRE: a rule based system for family relations extraction from english text. Knowledge and Information Systems, 64(7), 1817–1844. https://doi.org/10.1007/s10115-022-01687-4

[21] Imran, A., Fahim, M., Alzahrani, A., Fahim, S., Alheeti, K. M. A., & Rehman, S. U. (2023). Twitter Sentimental Analysis using Machine Learning Approaches for SemeVal Dataset. In 2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC) (pp. 1–6). Presented at the 2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC). https://doi.org/10.1109/ICAISC56366.2023.10085107

[22] Karam, M. W., Imran, A., Alzahrani, A., Ali Alheeti, K. M., Abbas, A., & Najem Al-Aloosy, A. A. (2022). Dramatic Increase in Fear-Related Discussion on Twitter during COVID-19: Analysis, Topic Modeling and Tweets Classification. In 2022 International Conference on Electrical Engineering and Sustainable Technologies (ICEEST) (pp. 1–8). Presented at the 2022 International Conference on Electrical Engineering and Sustainable Technologies (ICEEST). https://doi.org/10.1109/ICEEST56292.2022.10077859

[23] Kamath, U., Liu, J., & Whitaker, J. (2019). Deep Learning for NLP and Speech Recognition. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-14596-5

[24] Umair, M., Alam, I., Khan, A., Khan, I., Ullah, N., & Momand, M. Y. (2022). N-GPETS: Neural Attention Graph-Based Pretrained Statistical Model for Extractive Text Summarization. Computational Intelligence and Neuroscience, 2022, 6241373. https://doi.org/10.1155/2022/6241373

[25] Al-Azani, S., & El-Alfy, E.-S. M. (2020). Enhanced Video Analytics for Sentiment Analysis Based on Fusing Textual, Auditory and Visual Information. IEEE Access, 8, 136843–136857. https://doi.org/10.1109/ACCESS.2020.3011977

[26] Al-Twairesh, N., & Al-Negheimish, H. (2019). Surface and Deep Features Ensemble for Sentiment Analysis of Arabic Tweets. IEEE Access, 7, 84122–84131. https://doi.org/10.1109/ACCESS.2019.2924314

[27] Oueslati, O., Cambria, E., HajHmida, M. B., & Ounelli, H. (2020). A review of sentiment analysis research in Arabic language. Future Generation Computer Systems, 112, 408–430. https://doi.org/10.1016/j.future.2020.05.034

[28] Ali, M. M. (2021). Arabic sentiment analysis about online learning to mitigate covid-19. Journal of Intelligent Systems, 30(1), 524–540. https://doi.org/10.1515/jisys-2020-0115

[29] Mahyoob, M., Algaraady, J., Alrahiali, M., & Alblwi, A. (2022). Sentiment Analysis of Public Tweets Towards the Emergence of SARS-CoV-2 Omicron Variant: A Social Media Analytics Framework. Engineering, Technology & Applied Science Research, 12(3), 8525–8531. https://doi.org/10.48084/etasr.4865

[30] Vatsa, D. V. D., & Yadav, A. Y. A. (2022). An analytical insight of discussions and sentiments of Indians on Omicron-driven third wave of COVID-19 using twitter dat. https://doi.org/10.21203/rs.3.rs-1508291/v2

[31] Mubarak, H., & Hassan, S. (2021, March 1). ArCorona: Analyzing Arabic Tweets in the Early Days of Coronavirus (COVID-19) Pandemic. arXiv. https://doi.org/10.48550/arXiv.2012.01462

[32] Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

[33] Update on Omicron. (n.d.). World health organization. Retrieved September 9, 2022, from https://www.who.int/news/item/28-11-2021-update-on-omicron

[34] Makhoul, M., Ayoub, H. H., Chemaitelly, H., Seedat, S., Mumtaz, G. R., Al-Omari, S., & Abu-Raddad, L. J. (2020). Epidemiological Impact of SARS-CoV-2 Vaccination: Mathematical Modeling Analyses. Vaccines, 8(4), E668. https://doi.org/10.3390/vaccines8040668

[35] Rodrigues, R. F., da Silva, A. R., da Fonseca Vieira, V., & Xavier, C. R. (2018). Optimization of the Choice of Individuals to Be Immunized Through the Genetic Algorithm in the SIR Model. In O. Gervasi, B. Murgante, S. Misra, E. Stankova, C. M. Torre, A. M. A. C. Rocha, … Y. Ryu (Eds.), Computational Science and Its Applications – ICCSA 2018 (pp. 62–75). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-95165-2_5

[36] Levine-Tiefenbrun, M., Yelin, I., Katz, R., Herzel, E., Golan, Z., Schreiber, L., … Kishony, R. (2021, February 8). Decreased SARS-CoV-2 viral load following vaccination. medRxiv. https://doi.org/10.1101/2021.02.06.21251283

[37] Reis, R. F., de Melo Quintela, B., de Oliveira Campos, J., Gomes, J. M., Rocha, B. M., Lobosco, M., & Weber Dos Santos, R. (2020). Characterization of the COVID-19 pandemic and the impact of uncertainties, mitigation strategies, and underreporting of cases in South Korea, Italy, and Brazil. Chaos, Solitons, and Fractals, 136, 109888. https://doi.org/10.1016/j.chaos.2020.109888

[38] Xavier, C. R., Oliveira, R. S., Vieira, V. da F., Rocha, B. M., Reis, R. F., Quintela, B. de M., … Santos, R. W. dos. (2022). Timing the race of vaccination, new variants, and relaxing restrictions during COVID-19 pandemic. Journal of Computational Science, 61, 101660. https://doi.org/10.1016/j.jocs.2022.101660

[39] Shamrat, F. M. J. M., Chakraborty, S., Imran, M. M., Muna, J. N., Billah, M. M., Das, P., & Rahman, M. O. (2021). Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm. Indonesian Journal of Electrical Engineering and Computer Science, 463–470.

[40] Sontayasara, T., Jariyapongpaiboon, S., Promjun, A., Seelpipat, N., Saengtabtim, K., Tang, J., & Leelawat, N. (2021). Twitter sentiment analysis of Bangkok tourism during COVID-19 pandemic using support vector machine algorithm. (Special Issue: COVID-19 and historical pandemics.). Journal of Disaster Research, 24–30.

[41] Asgari-Chenaghlu, M., Nikzad-Khasmakhi, N., & Minaee, S. (2020, September 19). Covid-Transformer: Detecting COVID-19 Trending Topics on Twitter Using Universal Sentence Encoder. arXiv. https://doi.org/10.48550/arXiv.2009.03947

[42] Amen, B., Faiz, S., & Do, T.-T. (2022). Big data directed acyclic graph model for real-time COVID-19 twitter stream detection. Pattern Recognition, 123, 108404. https://doi.org/10.1016/j.patcog.2021.108404

[43] Lyu, J. C., & Luli, G. K. (2021). Understanding the Public Discussion About the Centers for Disease Control and Prevention During the COVID-19 Pandemic Using Twitter Data: Text Mining Analysis Study. Journal of Medical Internet Research, 23(2), e25108. https://doi.org/10.2196/25108

[44] Al-Ramahi, M., Elnoshokaty, A., El-Gayar, O., Nasralah, T., & Wahbeh, A. (2021). Public Discourse Against Masks in the COVID-19 Era: Infodemiology Study of Twitter Data. JMIR Public Health and Surveillance, 7(4), e26780. https://doi.org/10.2196/26780

[45] Qadi, L. A., Rifai, H. E., Obaid, S., & Elnagar, A. (2019). Arabic Text Classification of News Articles Using Classical Supervised Classifiers. In 2019 2nd International Conference on new Trends in Computing Sciences (ICTCS) (pp. 1–6). Presented at the 2019 2nd International Conference on new Trends in Computing Sciences (ICTCS). https://doi.org/10.1109/ICTCS.2019.8923073

[46] Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. Information Processing & Management, 50(1), 104–112. https://doi.org/10.1016/j.ipm.2013.08.006

[47] Obeid, O., Zalmout, N., Khalifa, S., Taji, D., Oudah, M., Alhafni, B., … Habash, N. (2020). CAMeL Tools: An Open Source Python Toolkit for Arabic Natural Language Processing. In Proceedings of the 12th Language Resources and Evaluation Conference (pp. 7022–7032). Presented at the LREC 2020, Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2020.lrec-1.868

[48] Constantin, C., du Mouza, C., Litwin, W., Rigaux, P., & Schwarz, T. (2016). AS-Index: A Structure for String Search Using n-Grams and Algebraic Signatures. Journal of Computer Science and Technology, 31(1), 147–166. https://doi.org/10.1007/s11390-016-1618-6

[49] Abbas, M., Memon, K. A., Jamali, A. A., Memon, S., & Ahmed, A. (2019). Multinomial Naive Bayes classification model for sentiment analysis. IJCSNS Int. J. Comput. Sci. Netw. Secur, 19(3), 62.

[50] Awad, M., & Khanna, R. (2015). Support Vector Machines for Classification. In M. Awad & R. Khanna (Eds.), Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers (pp. 39–66). Berkeley, CA: Apress. https://doi.org/10.1007/978-1-4302-5990-9_3

[51] Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random Forests. In C. Zhang & Y. Ma (Eds.), Ensemble Machine Learning: Methods and Applications (pp. 157–175). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4419-9326-7_5

[52] Alzahrani, A. (2023). A Safeguard Agent for Intelligent Health-care Environments. In 2023 International Conference on Smart Computing and Application (ICSCA) (pp. 1–6). Presented at the 2023 International Conference on Smart Computing and Application (ICSCA). https://doi.org/10.1109/ICSCA57840.2023.10087746

Abdulkareem Alzahrani is an assistant professor of Computer Science (AI) at Al-Baha University. He holds MSc in Advanced Web Engineering (2011), and PhD in computer science (2017) from the University of Essex, UK. His research interests include Artificial Intelligence, Computational Intelligence, Ambient Intelligence, Autonomous agent, and multi-agent systems.