Performance Evaluation in Arabic Sentiment Analysis during the Covid-19 Pandemic

Ahmed Saber Sakr[*1], Tamer S. EL Grwany[**2], M Amin [**3]

[*] *Department of information systems, Faculty of computers and information, Menoufia University, Egypt*

[1] a.ssakr@yahoo.com

[**] *Department of Mathematics and Computer Science, Faculty of Science, Menoufia University, Egypt*

[2] vipcompgroup55@gmail.com

[3] mohamed_amin110@yahoo.com

Abstract: ***This paper classifies sentiment analysis in Arabic language and mining sentiment in relation to the COVID-19 pandemic in the period (2019 - 2021). Three large data sets are collected from tweets, hotel and restaurant reviews for building the proposed sentiment analysis model. We compared eight machine learning algorithms, Multinomial Naïve Bayes (MNB), Bernoulli Naïve Bayes (BNB), Decision Tree (DT), K-nearest neighbour classifier (KNN), Support Vector Machines (SVM), Linear Support Vector Classifier (LSVC), Random Forest Classifier (RFC) and Stochastic Gradient Descent Classifier (SGD) on three cases: n-gram unigram, bigram, and trigram for each algorithm. The performance evaluations are compared according to precision, recall, and F-measure. The polarity prediction results in sentiment analysis models were achieved by linear SVC using hotel dataset with bigram case, with the accuracy of 0.966, precision of 0.967, recall of 0.966 and F-measure of 0.966 . The rest algorithms give normal execution on all datasets. It may very well be reasoned that the AI calculations need the right morphological components to upgrade the classification exactness when managing various words that assume various parts in the sentence with a similar letter.***

Key words: ***Sentiment analysis, Arabic language, Tweets, COVID-19 Pandemic.***

1 INTRODUCTION

With the onset of the Covid-19 pandemic in late 2019, World Health Organization (WHO) has suggested that isolation and self-quarantine are ones of the fundamental methods to stop this pandemic from spreading at an alarming rate. In the meantime, isolation and self-quarantine make Internet and social media communication one of the most important ways to unleash and share opinions and ideas. Accordingly, transforming these conclusions and posts into resources is profoundly important, which prompted the extraction of human feelings and diversion from online media networks for utilizing in global public powers, business choices and strategy improvement using sentiment analysis. Sentiment analysis (SA) or "opinion mining" is a natural language processing (NLP) technique to identify the users' feelings towards a certain vision through various electronic sources based on their opinions. With the increment of Arabic substance via web-based media, and the development of Arab politics significantly in the Middle East and North Africa, prompting researchers to direct attention towards techniques for processing natural languages in the Arabic language, including the analysis of feelings in Arabic. However, there are very limited studies studying emotions in the Arabic language, therefore the current study aims to classify Sentiment analysis in Arabic language using different methods of Machine Learning (ML) algorithms. Eight methods were used, namely, Bernoulli Naïve Bayes (BNB), Multinomial Naïve Bayes (MNB), Decision Tree (DT), K-nearest neighbor classifier (KNN), Support Vector Machines (SVM), Linear Support Vector Classifier (LSVC), Random Forest Classifier (RFC) and Stochastic Gradient Descent Classifier) SGD). To compare these methods, four indicators were used namely accuracy, recall, precision, and F-measure. All methods were performed on three large data sets, up to 105k, collected in the Covid-19 pandemic period from 2019 till 2021.

2 RELATED WORK

In [1], the occurrence of different types of infectious diseases during the past 10 years such as epidemics, pandemics, and viruses like COVID-19 or disease outbreaks are reviewed and analyzed. The point was understanding the utilization of notion investigation and acquire the main discoveries of the writing. Articles on related points were deliberately looked in five significant data sets, specifically, ScienceDirect, PubMed, Web of Science, IEEE Xplore and Scopus, from 1 January 2010 to 30 June 2020.

In [2], an extensive relative investigation on certain methodologies utilized for Arabic opinion examination is introduced. They re-carried out some current methodologies for Arabic SA and tried their adequacy on three datasets for Arabic SA. Their outcomes showed that the best model accomplishes the best F-score scores on ArSAS benchmark datasets. In [3] a correlation between three AI calculations is introduced. The creators looked at Support Vector Machine (SVM), Naïve Bayes (NB), and Decision Tree (DT) by means of a grouping strategy. The correlation was done on four evaluation estimations, to be specific Accuracy (ACC), Precision (PRE), Recall (REC), and F-measure (F-MES). They created Arabic feeling dataset from tweets, item audits, lodging surveys, film surveys, item attractions, and café surveys from various sites, which physically named for preparing the assessment analysis model. The outcomes showed that (SVM) and (NB)

classifiers delivered good outcomes while (DT) just accomplished normal outcomes. Polarity prediction gives the best result by (SVM) on product attraction dataset, with an ACC of 0.96, PRE of 0.99, REC of 0.99, and F-MES of 0.98. This is followed by average performance from NB and DT. In [4] the author examined a balanced and unbalanced dataset where a Large Arabic Books Review (LABR) dataset was created. This is an enormous dataset with 63,257 examples produced using surveys of 64 books in Arabic. The examination isolated the dataset into the uneven dataset and the reasonable dataset (for example the quantity of audits are equivalent in every one of the three classes), and tried different well known AI classifiers like the Bernoulli (BNB), Multinomial Naïve Bayes (MNB) and SVM with various blends of gauging models (TF-IDF, unigram, bigram, and trigram). The MNB classifier was observed to be the best utilized classifier in the exploration. In any case, this investigation has just covered the book audits space, where the surveys have an alternate in general extremity contrasted with issues like politics. In [5] the authors introduced the benefits of combining CNNs and LSTMs networks in an Arabic sentiment classification task. Also, they showed the usefulness of using multi levels of SA due to the complexities of morphology and orthography in Arabic. They increased the number of features by using character level in tweet DS. Their model, using Word-level and Ch5gram-level, showed better sentiment classification results as their approach has improved the sentiment classification accuracy for their Arabic. Health Services (AHS) dataset to reach 0.9424 for the Main-AHS dataset, and 0.9568 for the Sub-AHS dataset.

In [6] the creator proposed two enhancements for WOA calculation. They utilized four Arabic benchmark datasets to assess the estimation examination. The calculation was contrasted and six notable advancement calculations and two profound learning calculations. The thorough tests results showed that there is a calculation that beats any remaining calculations as far as feeling investigation arrangement precision through tracking down the best arrangements, while likewise limiting the quantity of chose highlights.

In [7], (ML) and (SVM) advanced computing algorithms were utilized to train the automatically collected dataset through ArabiTools and Twitter API. The contents of dataset are classified automatically and manually, in order to maintain efficient detection of CyberBullying tweets. The dataset is automatically labelled with respect to the nature of the tweet. If a tweet contains one or more CyberBullying words, it is labelled as CyberBullying, while if no word with an offensive meaning found, it is marked as NonCyberBullying.

3 PROPOSED MODEL

This investigation introduces the classification methodology for Arabic SA. The classification model comprises of four principle levels (Fig. 1) as follows. The first level is pre-processing, the second level is feature extraction, the third level is classification and the fourth level is evaluation. Python3 is utilized to fabricate the assessment investigation models since it has an ideal capacity to deal with regular dialects quite well and all the more explicit strings. The language likewise upholds Natural Language Processing (NLP) libraries as needed in this investigation. The detailed stage can be shown as follows:

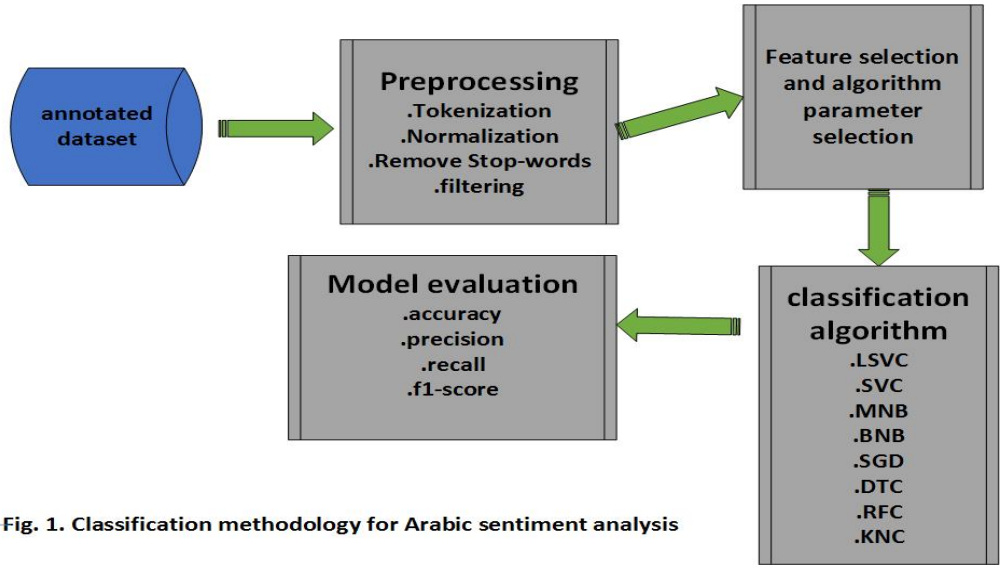


Fig. 1. Classification methodology for Arabic sentiment analysis

A. Annotation of Dataset

In this work we used three different Arabic data sets. The first is for hotel reviews and collected from several hotels during Covid-19 pandemic and it consists of 105694 balanced reviews (52848 positive, 52846 negative). and the second is for reviewing restaurants. It was collected from several reviews during the Covid-19 pandemic and consists of 5450 unbalanced reviews (3820 positive, 1630 negative) and the third Arabic data set is sourced from [8]. This dataset was collected in April 2019 during Covid-19 pandemic, and it contains 58751 balanced Arabic tweets (29849 positive, 28902 negative). The dataset was collected using positive and negative emoji lexicon. All datasets were divided into 70% for training and 30% for testing.

B. Pre-Processing

Pre-preparing is the main stage which is applied subsequent to getting information from the source to diminish mistakes, increment precision and eliminate uproarious components. The tasks that are completed incorporate tokenization, standardization, eliminating stop words, and sifting. Tokenization alludes to the way toward changing over a whole content into a progression of tokens with the goal that every token is isolated and free of the other. Standardization implies eliminating unvital images and letters from the dataset. Eliminating stop-words ordinarily allude to well known words in the language messages, for example, (من ,هذا ,عليه ,لما ,حتى ,عن ,هو), sifting (for example eliminating terms that show up in under 1% of the archives).

C. Feature Extraction

After the pre-processing, the information highlight extraction stage is applied to produce the component vectors. Three distinctive element extraction techniques were utilized in the investigation; the language models unigram, bigram and trigram. The situation of any term in the portrayal of a solitary dataset is critical since this term position recognizes and in some cases flips the expression extremity. N-gram strategy is utilized to handle the refutation issue in Arabic since nullification in Arabic is utilized to switch the extremity of a word (for example "لا", "ما", "لم") just as the area of these particles toward the start of the sentence. For instance, "هو لم يعجبه مشاهدة هذا الفيلم" signifies "He didn't care for watching this film". The action word "like" is named a good inclination however "didn't" changed extremity of the sentence from positive to negative. Then, at that point, the Term Frequency-Inverse Document Frequency (TF-IDF) is utilized to scale each element (term, unigram or bigram) in the vector. The TF-IDF gives factual data that actions the meaning of a word in a dataset or a bunch of datasets. The worth of the TF-IDF expands relatively to the occasions a given word is rehashed in the dataset. The TFIDF score for a given term is determined by Equations from 1 to 3.

$$TF(i, j) = (F(i,j))/(N(j)) \quad (1)$$

where F(i,j) represents the frequency of the term i in the dataset j and N(j) represents the total number of terms in the dataset j

$$IDF(i) = \log N/(N(i)) \quad (2)$$

where N represents the total number of datasets and N(i) represents the total number of datasets containing the term i

$$TF - IDF(i,j) = TF(i,j) * IDF(i) \quad (3)$$

D. Classification Algorithms

After the feature vectors were generated, eight classification algorithms were used in the comparison as follows:

1)K-Nearest Neighbor Classifier (KNN)

This classifier picks the K number for the closest neighbors in the preparation reports and orders a clarified archive dependent on these K neighbors. Specifically, it ascertains the comparability between the unlabeled report and the excess records in the preparation dataset. From there on, the names of the most K comparative reports are thought of. The last mark of the new not really settled utilizing a greater part vote or the weighted normal of the names of these K neighbors.

2) Multinomial Naïve Bayes (MNB)

Naïve Bayes (NB) Classification utilized in broad applications in business, ham/spam sifting [16], wellbeing, internet business, online media opinion, item assumption among clients and so forth, Bernoulli Naïve Bayes(BNB) Classification and MNB Classification is two famous methodologies of NB Text Categorization [10]. The multinomial model is intended to decide the recurrence of a term for example the occasions a term happens in an archive. Considering the way that a term might be essential in choosing the supposition of the report, the property of this model settles on it a fair decision for archive order. MNB Classifier can be detailed in Eq. 4:

A dataset 'd' with polarity 'P' is calculated as follows:

$$(P|d) \propto (P)\ \Pi P(tk1 \leqslant k \geqslant nd|P) \quad (4)$$

where P(tk|P): represents the conditional probability that whether the term tk occurs in the dataset of polarity p which is calculated according to Eq. 5:

$$(tk|p) = (count(tk|p)+1)/(count(tp)+|V|) \quad (5)$$

Here, check (tk|p) implies the occasions the term tk happens in the dataset having p extremity and tally (tp) implies the absolute number of tokens present in the dataset of extremity p. Additionally, 1 and |V| are added as smoothing constants which are added to keep away from computational setbacks when the term doesn't happen at all in the dataset or the dataset is vacant or invalid. This idea is also called Laplace Smoothing. |V| is the quantity of terms in the all out jargon of the dataset. P(p): addresses the earlier likelihood of dataset being of extremity p which is determined as Eq. 6:

$$P(p) = (\text{Number of dataset of polarity p})/(\text{Total number of dataset}) \quad (6)$$

nd: represents number of tokens in the dataset

tk: represents the kth token in the dataset

The probability P(p|n) is calculated for both i.e., the positive polarity as well as the negative polarity and the maximum is considered to be the predicted polarity for dataset.

3) Bernoulli Naive Bayes

In the Bernoulli Naïve Bayes Classifier calculation, highlights are autonomous paired factors addressing that whether a term is available in the archive viable or not. Being somewhat like the multinomial model in the characterization cycle, this calculation is additionally a well known methodology for text order undertakings however varies from the multinomial methodology in the viewpoint that multinomial methodology considers the term frequencies though Bernoulli approach is just keen on concocting that whether the term is available or missing in the archive viable. Bernoulli Naïve Bayes Classifier can be formed as displayed in Eq. 7:

A dataset 'n' being of polarity 'p' is calculated as:

$$(p|n) \propto (p)\ \Pi P(tk1 \leqslant k \geqslant nd|p)(1 - P(tk'\ |p)) \quad (7)$$

where P(tk|p): represents the conditional probability of the occurring term tk in a dataset of polarity p and P (tk'|p) represents the conditional probability of non-occurring term tk' in a dataset. Both of these contingent probabilities are given as shown in Eq 8,9:

$$(tk|p) = (count(tk|p)+1)/(count(Np)+2) \quad (8)$$

$$(tk'|p) = (count(tk'|p)+1)/(count(Np)+2) \quad (9)$$

Here, count (tk|p) indicates the count of occurrences of the term in the dataset of polarity p where the value for a given dataset can be 0 or 1 and count (Np) means the total number of dataset having polarity as p.

P(p): represents the prior probability of dataset being of polarity p which is studied as Eq10:

$$(p) = (\text{Number of dataset of polarity p})/(\text{Total number of dataset}) \quad (10)$$

nd: represents number of tokens in a dataset.

tk: represents the kth token in the dataset.

4) Support Vector Machines (SVM)

It is a sort of directed learning calculation; it is a compelling conventional book classification structure. The principle thought of SVM is to track down the hyper-plan, which is addressed as a vector that isolates record vectors in a single class from the report vectors in different classes. The SVM model attempts to grow the distance between the two classes by making a distinct choice limit. This works by characterizing an isolating hyperplane or set of hyperplanes. The yield of this calculation is an ideal hyperplane that amplifies the detachment distance between the two positive and negative hyperplanes used to sort new models when utilized with marked preparing information. An ideal detachment is the point at which the hyperplane contains the biggest distance to the closest preparing information points of any class, with the base extent of the vector | w |. The equation is displayed in Eq.11 where w is a weight vector, x is input vector, and b is the predisposition.

$$\min|w| = yi\ (w \cdot xi + b) \geqslant 1;\ i = 1; \ldots, N \quad (11)$$

5) Linear Support Vector Classifier (LSVC)

LSVC method applies a linear kernel function to execute the classification. The linear SVC has extra parameters such as penalty normalization which applies 'L1' or 'L2' and a loss function. The kernel method cannot be changed in the linear SVC, because it is based on the kernel linear method [17].

6) Stochastic Gradient Descent (SGD)

SGD [13] is an iterative strategy for enhancing a target function with reasonable perfection properties (for example differentiable or sub-differentiable). It tends to be viewed as a stochastic guess of inclination drop advancement, the factual assessment considers the issue of limiting a target work having the type of an aggregate as shown in Eq.12., where the boundary (W) that limits Q(w) is to be assessed. Each summand work (Qi) is commonly connected with the (I-th) perception.

$$Q(w)=\frac{1}{n}\sum_{i=1}^{n} Qi(w) \quad (12)$$

7) Random Forest Classifier (RFC)

RFC are a troupe learning strategy for grouping, relapse and different errands that work by developing a large number of choice trees at preparing time and yielding the class that is the method of the classes (characterization) or mean/normal forecast (relapse) of the individual trees Random choice woodlands right for choice trees' propensity for over-accommodating their own preparation set.

8) Decision Tree (DT)

(DT) used to take care of numerous arrangements and relapse issues. The premise of its work is to build a tree outline that contains arrangement models and afterward partition the dataset into a more modest incomplete dataset and afterward foster the tree of choices in continuous stages. The yield of this calculation is a tree, leaf hubs and the choice hubs. From that point, utilizing a given size of data the tree is created, here this requirement is to utilize the greatest degree of data when the two primary leaves are equivalent in number. In choosing which element to part at each progression in building the tree, data acquire is determined as displayed, whereby (J) addresses the classes and (p) are the things as show in Eq .13.

$$IG(P)=1-\sum_{i=1}^{J} P_i^2 \quad (13)$$

E. Evaluation

Performance Metrics in most SC problems, three measures of classification effectiveness are most used: accuracy, precision, and recall. We use them and the F-Measure to measure the accuracy of the test data as it considers both the precision and the recall of the test in computing the score.

i) **Accuracy** **Accuracy** $(ACC) = \frac{(TP+TN)}{(TP+TN+FP+FN)}$ (14)

where:
TP (True Positives): the number of positively-labelled test sentences that are correctly classified as positive.
TN (True Negatives): the number of negatively-labelled test sentences that are correctly classified as negative.
FP (False Positives): the number of negatively-labelled test sentences that are incorrectly classified as positive.
FN (False Negatives): the number of positively-labelled test sentences that are incorrectly classified as negative.

ii) **Precision:** defines the probability that if a random sentence should be classified as positive, then this is the correct decision as shown in Eq. 15.

$$\textbf{Precision } PRE = \frac{TP}{(TP+FP)} \quad (15)$$

iii) **Recall:** is the probability that if a random sentence should be classified as positive, then this is the taken decision as shown in Eq. 16.

$$\textbf{Recall } REC = \frac{TP}{(TP+FN)} \quad (16)$$

iv) **F-Measure:** determines the weighted average for both the precision and recall obtained. We use the F1 measure, so that both the recall and the precision are evenly weighted, as shown in Eq. 17.

$$\textbf{F-Measure } F\text{-}MES = \frac{2(P * R)}{(P+R)} \quad (17)$$

4 SIMULATION RESULTS AND ANALYSIS

In this Section, the results of K-nearest neighbor classifier (KNN), Bernoulli Naive Bayes, Multinomial Naïve Bayes, Support Vector Machines (SVM), The Linear Support Vector Classifier (LSVC), Stochastic Gradient Descent (SGD), Random Forest Classifier and Decision Tree (DT) are analyzed and discussed. The exhibition is estimated utilizing precision , accuracy, recall, and F-measure.

A. K-Nearest Neighbor Classifier (KNN)

Figure 2 shows the results of KNN algorithm on the three datasets. Execution is estimated utilizing ACC, PRE, REC, and F-MES. The best outcomes accomplished by KNN were for the restaurant reviews data set when n-gram 3 with ACC of 0.882, PRE of 0.882, F-MES of 0.882 and REC of 0.882. The lowest ACC was 0.548 on hotel datasets, the minimum PRE was the tweets dataset with 0.73, the minimum F-MES and REC were the tweets with 0.629 and 0.548, Straight. Compared to [9], KNN gives the highest REC and is equal to 69.04 when (K=10).

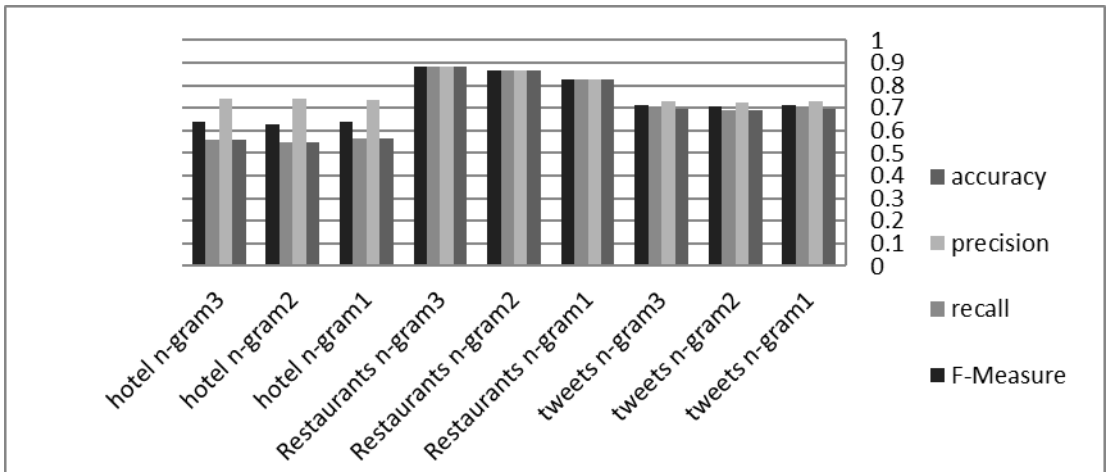


Figure 2: Results for K Neighbours Classifier

B. Multinomial Naïve Bayes

Figure 3 shows the results of the Multinomial Naïve Bayes algorithm to the three datasets. Execution is estimated utilizing ACC, PRE, REC, and F-MES. The best outcomes accomplished by Multinomial Naïve Bayes were for the hotel reviews data set when n-gram 2 with an ACC of 0.948, PRE of 0.959, an F-MES of 0.959 and REC of 0.959. The lowest ACC was 0.761 on tweets datasets, the lowest PRE was the tweets dataset with 0.762, the lowest F-MES and REC was the tweets with 0.761.

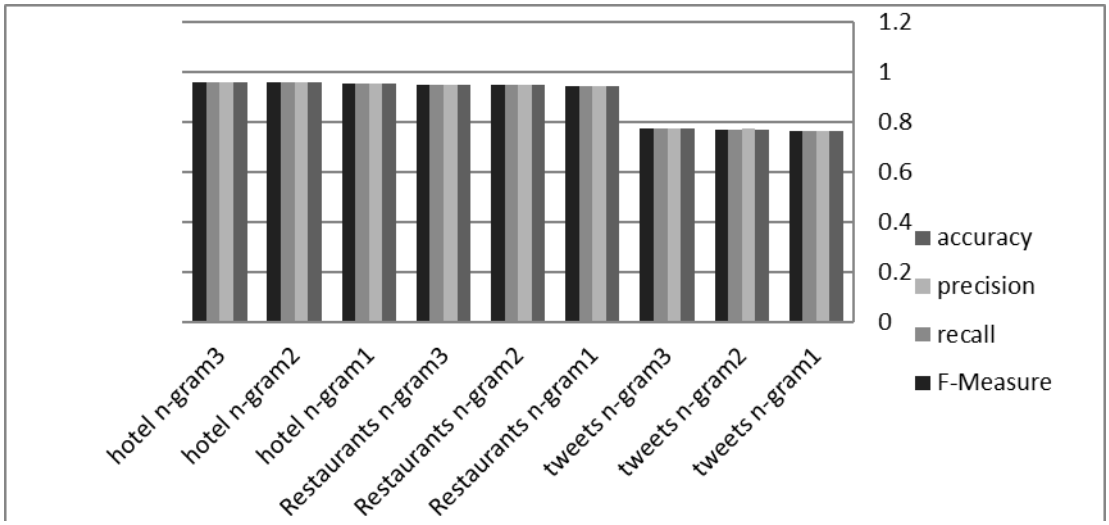


Figure 3: Results for Multinominal

C. Bernoulli Naive Bayes

Figure 4 presents the results of Bernoulli Naive Bayes algorithm on the three datasets. Execution is estimated utilizing ACC, PRE, REC, and F-MES. The best outcomes accomplished by Bernoulli Naive Bayes was for Restaurants reviews data set when n-gram 1 with an ACC of 0.948, a PRE of 0.951, an F-MES of 0.949 and a REC of 0.948. The lowest ACC was standardized with 0.758 on tweets datasets; the lowest PRE was the tweets dataset with 0.763, the minimum F-MES and REC were the tweets with 0.76 and 0.758, Straight. Compared with [10], MNB perform somewhat better compared to BNB on dataset with a smaller number of records (312 records for this situation); be that as it may, MNB arrives at an exactness of around 73% which isn't extremely proficient. This follows the way that it is truly challenging to accomplish high precision with less measure of information and more information will prompt more prominent exactness with every one of the examined calculations. In the current examination, the creators additionally reasoned that albeit MNB gives more noteworthy exactness, however the distinction in precision isn't exceptionally huge as BNB likewise gives a precision of right around 69% which suggests that the presentation of these calculations doesn't contrast much from the given dataset.

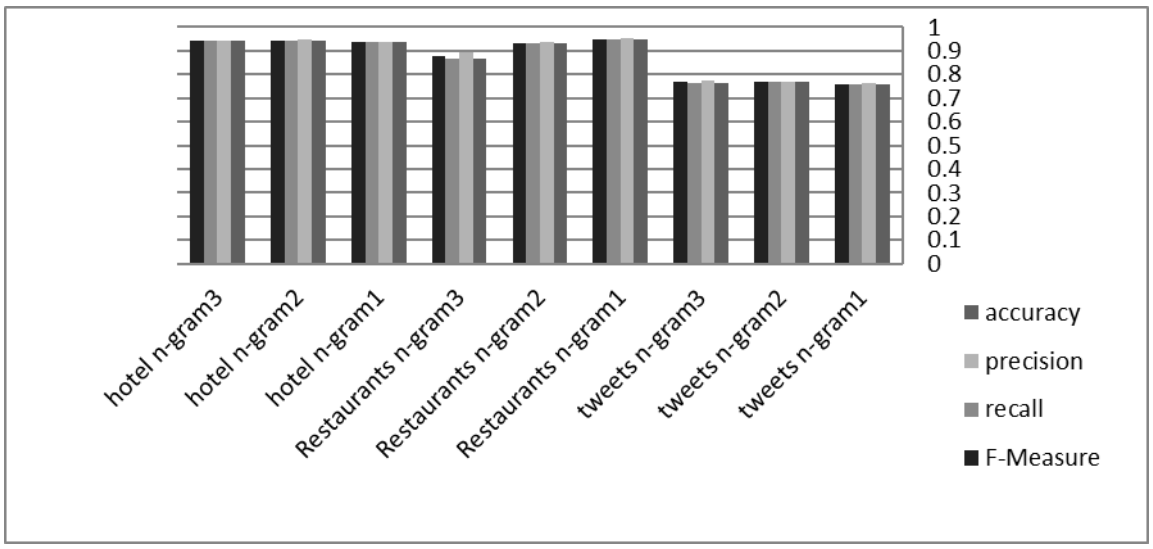


Figure 4: Results for Bernoulli NB

D. Support Vector Machines (SVM)

Figure 5 shows the results of SVM algorithm to the three datasets. Execution is estimated ACC, PRE, REC, and F-MES. The best outcomes accomplished by Support Vector Machines were for hotel reviews data set at an n-gram of 2.3, an ACC of 0.964, a PRE of 0.966, an F-MES of 0.965 and a REC of 0.966. The lowest ACC was 0.8 on tweets datasets; the minimum PRE was for the tweets dataset with 0.803, the minimum F-MES and REC were the tweets with 0.801 and 0.8, straight. Compared with the findings of [3],[9],[12],[13], SVM also gives high result.

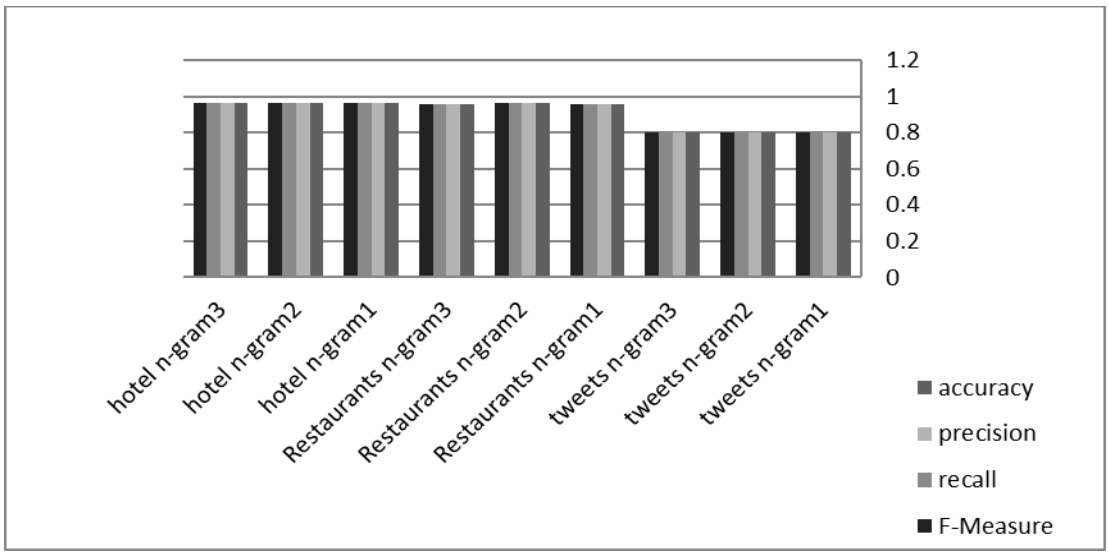


Figure 5: Results for SVM

E. The Linear Support Vector Classifier (SVC)

Figure 6 shows the results of SVC algorithm to the three datasets. Execution is estimated utilizing ACC, PRE, REC, and F-MES. The best outcomes accomplished by LSVC were for the hotel reviews data set when n-gram 2.3 with an ACC of 0.966, a PRE of 0.967, an F-MES of 0.966 and a REC of 0.966. The lowest ACC was with 0.774 on tweets datasets; the minimum PRE was the tweets dataset with 0.773, the minimum F-MES and REC were the tweets with 0.772 and 0.771, straight.

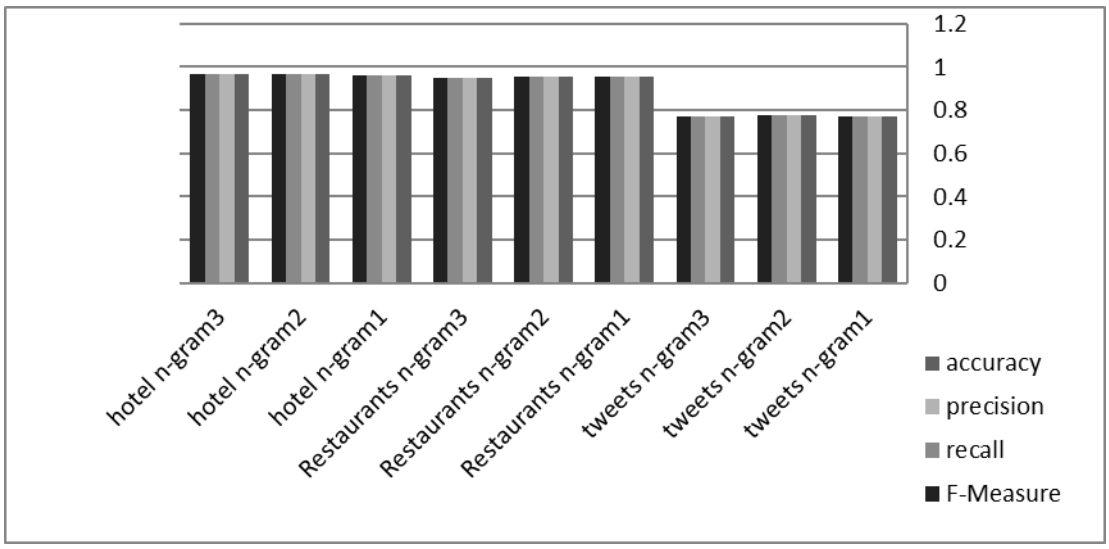


Figure 6: Results for linear SVC

F. Stochastic Gradient Descent (SGD)

Figure 7 shows the results of LSVC algorithm to the three datasets. Execution is estimated utilizing ACC, PRE, REC, and F-MES. The best outcomes accomplished by SGD were for the hotel reviews data set at n-gram 1.3 with an ACC of 0.96, a PRE of 0.962, an F-MES of 0.96 and a REC of 0.96. The lowest ACC was with 0.765 on tweets datasets; the minimum PRE was the tweets dataset with 0.77, the minimum F-MES and REC were on tweets with 0.777 and 0.765, Straight. Compared with [14], SGD is a decent learning calculation when the preparation set is enormous and gives helpful suggestions.

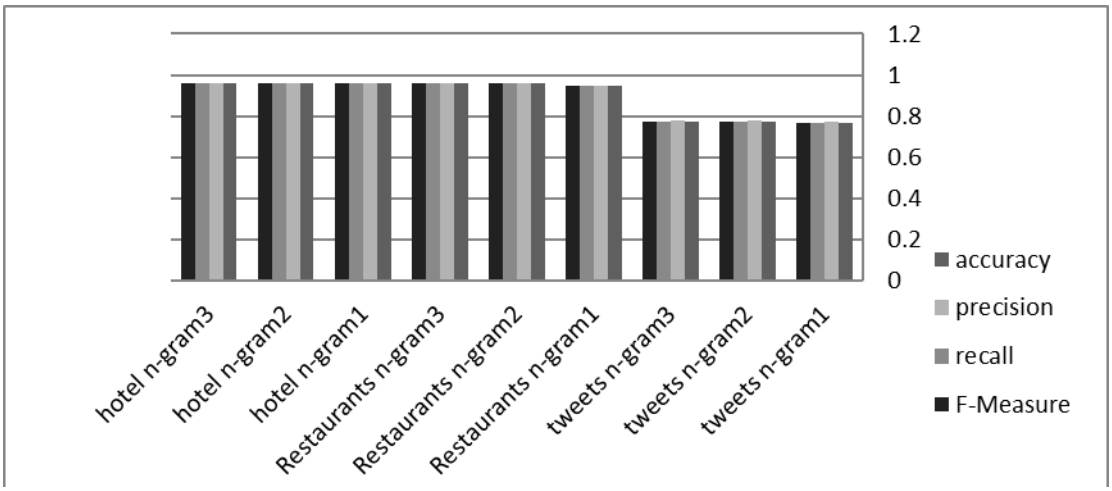


Figure 7: Results for SGD Classifier

G. Random Forest Classifier

Figure 8 shows the results of Random Forest Classifier algorithm to the three datasets. Execution is estimated utilizing ACC, PRE, REC, and F-MES. The best outcomes accomplished by Random Forest Classifier were for restaurants reviews data set when n-gram 1 with an ACC of 0.538, a PRE of 0.693, an F-MES of 0.605 and a REC of 0.538. The minimum

ACC was with 0.504 on hotel datasets; the minimum PRE was the tweets dataset with 0.62, the minimum F-MES and REC were on tweets with 0.556 and 0.504 Straight. Compared to the results in [15], Random Forest Classifier perhaps the most productive arrangement strategies when used in image, but it gives low results in text classification.

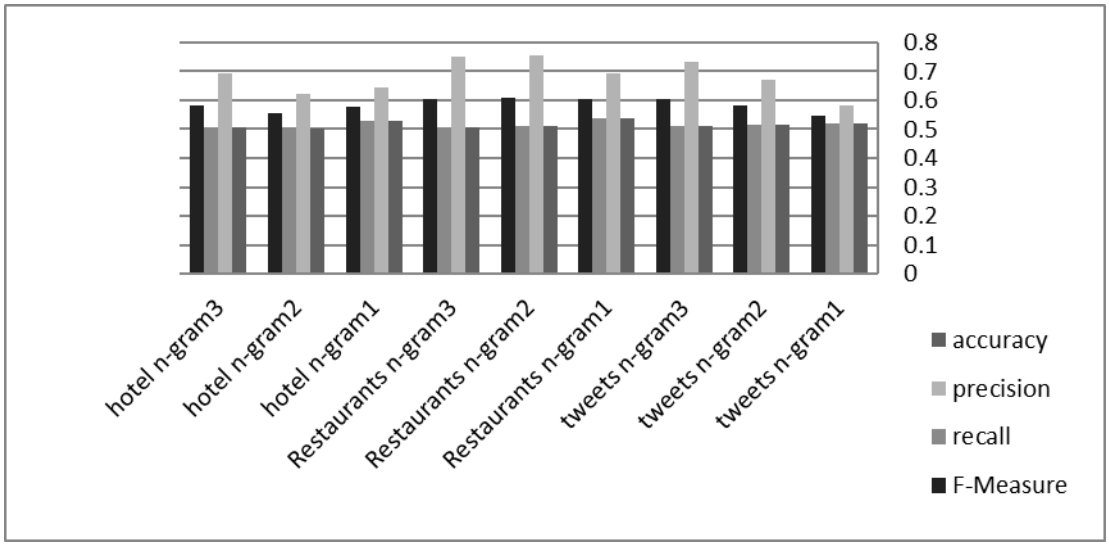


Figure 8: Results for Random Forest Classifier

H. Decision Tree (DT)

Figure 9 show the results when applying Decision Tree Classifier algorithm on the three datasets. The best results achieved by Decision Tree Classifier were for restaurants reviews data set when n-gram 1 with an ACC of 0.853, a PRE of 0.88, an F-MES of 0.866 and REC of 0.853. The lowest ACC was unified with 0.572 on tweets datasets; the lowest PRE was the tweets dataset with 0.694, the minimum F-MES and recall were the tweets with 0.627 and 0.572, straight. Compared to the results in [3],[17], the results of the DT are also low.

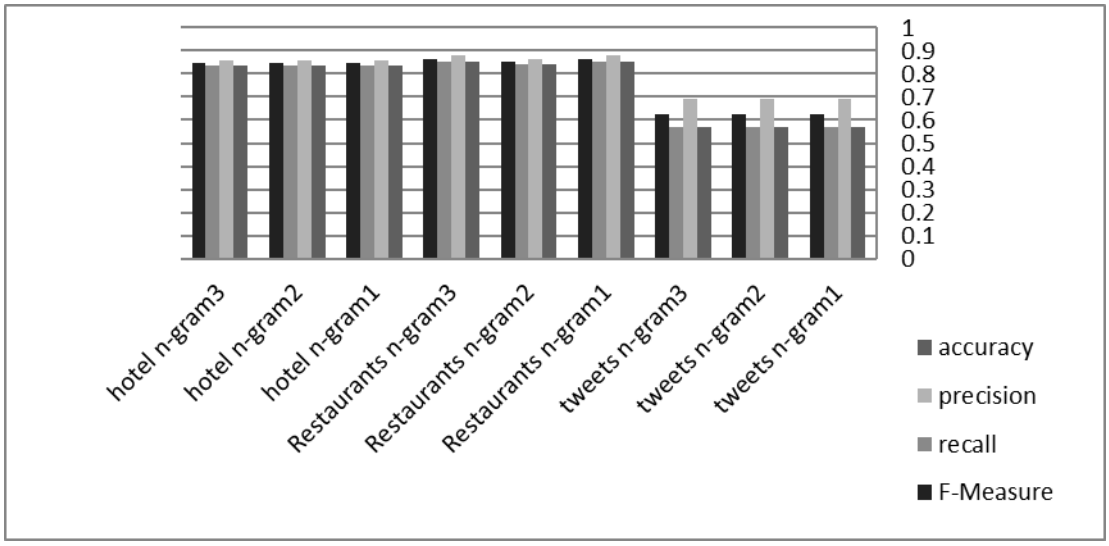


Figure 9: Results for Decision Tree (DT)

5 CONCLUSION

This study evaluated the performance of eight Machine Learning (ML) algorithms: Multinomial Naïve Bayes (MNB), Bernoulli Naïve Bayes (BNB), Decision Tree (DT), K-nearest neighbor classifier (KNN), Support Vector Machines (SVM), Linear Support Vector Classifier (LSVC), Random Forest Classifier (RFC) and Stochastic Gradient Descent Classifier (

SGD). We used three cases of n-gram 1.2 and 3 for every algorithm. Three different datasets (hotel reviews, Restaurants reviews, tweets) have different number of reviews were applied to compare the performance of the algorithms. The assessment was done dependent on four assessment measurements, namely ACC, PRE, REC, and F-MES. The outcomes acquired showed that the best accuracy for the hotel reviews data set was 96.6% by LSVC when n-gram =2.3 lowest accuracy was achieved by RFC was 0.507 when n-gram =3. When using Restaurants reviews data set the best accuracy obtain by SVC when n-gram =2, it was 96.2% whereas, the lowest accuracy was achieved by RFC was 0.507 when n-gram =3. When using tweets data set, the best accuracy obtained by SVC when n-gram =1,2 was 80%, and the lowest accuracy achieved by RFC was 0.512 when n-gram =3. All algorithms were not affected by the size of the datasets except KNN, which gave low results with large datasets.

REFERENCES

[1] A. Alamoodi, B. Zaidan, A. Zaidan, O. Albahri, K. Mohammed, R. Malik and M. Alaa, "Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review", *Expert systems with applications*, 114155, 2020.

[2] I.A. Farha, and W. Magdy, "A comparative study of effective approaches for Arabic sentiment analysis". *Information Processing & Management*, vol. *58, no.* 2, 102438, 2021.

[3] M. A. Algburi, A. Mustapha, S.A. Mostafa and M.Z. Saringatb," Comparative Analysis for Arabic Sentiment Classification ", In International Conference on Applied Computing to Support Industry: Innovation and Technology (pp. 271-285). Springer, Cham, Ramadi, Iraq, September 2019.

[4] M. Nabil, M. A. Aly and A. F. Atiya, " Labr: A large scale Arabic book reviews dataset." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, vol. 2*, pp. 494-498. 2013.

[5] A. Alayba, V. Palade, M. England, and R. Iqbal. "A combined CNN and LSTM model for Arabic sentiment analysis." In International cross-domain conference for machine learning and knowledge extraction, pp. 179-191. Springer, Cham, 2018.

[6] Almutiry, Samar, and Mohamed Abdel Fattah. "Arabic CyberBullying Detection Using Arabic Sentiment Analysis." *The Egyptian Journal of Language Engineering*, vol.8, no. 1, pp. 39-50, 2021

[7] https://www.kaggle.com/mksaad/arabic-sentiment-twitter-corpus. (Accessed 1 August 2021)

[8] R. M. Duwairi, and I. Qarqaz, "Arabic sentiment analysis using supervised classification". In *2014 International Conference on Future Internet of Things and Cloud*, pp. 579-583. IEEE, Spain,August 2014.

[9] G. Singh, B. Kumar, L. Gaur and A. Tyagi, "Comparison between multinomial and Bernoulli naïve Bayes for text classification." *2019 International Conference on Automation, Computational and Technology Management (ICACTM)*. London, IEEE, 2019.

[10] Chan, Wint Nyein, and Thandar Thein. "A comparative study of machine learning techniques for real-time multi-tier sentiment analysis." *2018 1st IEEE International Conference on Knowledge Innovation and Invention (ICKII)*. IEEE, South Korea, 2018.

[11] N. Al-Twairesh, H. Al-Khalifa, A. Al-Salman," Subjectivity and sentiment analysis of Arabic: trends and challenges" In: 2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA), pp. 148–155. Doha. Qatar, IEEE 2014.

[12] A. Shoukry,A. Rafea, " Sentence-level Arabic sentiment analysis", In: 2012 International Conference on Collaboration Technologies and Systems (CTS), USA, pp. 546–550, IEEE 2012

[13] *L. Bottou, "Stochastic gradient descent tricks. In Neural networks: Tricks of the trade" , Springer, Berlin, Heidelberg. 2012.*

[14] Akar, Özlem, and Oguz Güngör. "Classification of multispectral images using Random Forest algorithm." *Journal of Geodesy and Geoinformation* vol.1, no.2, pp.105-112, 2012

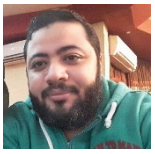
[15] M. Y. H. Setyawan, R. M. Awangga, and S.R. Efendi, "Comparison of multinomial naive bayes algorithm and logistic regression for intent classification in chatbot". In *2018 International Conference on Applied Engineering (ICAE)* (pp. 1-5). Indonisea, IEEE, October 2018

[16] G. Gautam, D. Yadav " Sentiment analysis of Twitter data using machine learning approaches and semantic analysis". In: 2014 Seventh International Conference on Contemporary Computing (IC3), pp. 437–442. India, IEEE 2014

[17] R, L. Jerlin, and E Perumal. "Hybrid kernel support vector machine classifier and grey wolf optimization algorithm based intelligent classification algorithm for chronic kidney disease." *Journal of Medical Imaging and Health Informatics* 10.10 (2020): 2297-2307.

BIOGRAPHY

Ahmed S. Sakr received the B.Sc. degree (Hons.) in mathematics and computer science and the M.Sc. degree in computer science from Menoufia University, Egypt, in 2008 and 2013, Straight, and the Ph.D. degree in computer science from Menoufia University, Egypt, in 2016. He is currently a Lecturer at faculty of computers and information Menoufia University. He has authored or co-authored many publications, including refereed IEEE, conference papers.His areas of interests are multimedia content encryption, big data security, secret media sharing, information hiding, cloud computing, bioinformatics and quantum information.



Tamer S. EL Grwany received a B.Sc. degree in Science – Mathematics and Computer Science Department with good degree in 2002. Postgraduate Diploma in Computer Science With excellent degree in 2014 from the Faculty of Science, Menoufia University His areas of interest include sentiment analysis, machine learning.

Mohamed Amin received his B.Sc. in Mathematics, Faculty of Science, Menoufia University 1983, He studied computer science from 1986 to 1989 at Ain Shams University in Cairo and received the M.Sc. degree in 1990 and the Ph.D. degree in computer science in 1997 at the University of Gdansk, Poland. He is a professor of computer science at the Faculty of Science, Menoufia University. He is an author and co-author of many international repute journals such as Information Sciences, Scientific reports, Physica A, optics communications, Communications in Nonlinear Science and Numerical Simulation and others. He is a reviewer for many prestigious journals in the field of his interest. His research interests include Grammar systems as a link between AI & Compiler Design, Metaheuristic optimization algorithms, ets, reasoning dynamic fuzzy systems, Cryptography, Quantum information processing, image ing, and biometrics.

تقييم الأداء في تحليل المشاعر العربية أثناء جائحة كوفيد -19

أحمد صابر صقر[*1] , تامر سامي الجرواني [**2], محمد أمين عبدالواحد [**3]
[*]قسم نظم المعلومات – كلية الحاسبات و المعلومات – جامعة المنوفية – المنوفية – مصر
[1]a.ssakr@yahoo.com
[**]قسم الرياضيات وعلوم الحاسب -كلية العلوم – جامعة المنوفية – المنوفية – مصر
[2] vipcompgroup55@gmail.com
[3]mohamed_amin110@yahoo.com

ملخص

تصنف هذه الورقة تحليل المشاعر في اللغة العربية ومشاعر التعدين فيما يتعلق بوباء COVID-19 في الفترة (2019 - 2021). تم جمع ثلاث مجموعات بيانات كبيرة من التغريدات ومراجعات الفنادق والمطاعم لبناء نموذج تحليل المشاعر المقترح. تم مقارنة ثماني خوارزميات للتعلم الآلي ، و Multinomial Naïve Bayes (MNB)، و Bernoulli Naïve Bayes (BNB) ، وشجرة القرار (DT) ، ومصنف الجار الأقرب (KNN) ، وآلات متجه الدعم (SVM)، ومصنف ناقل الدعم الخطي (LSVC) ، Random Forest Classifier (RFC)و Stochastic Gradient Descent Classifier (SGD)في ثلاث حالات n-gram unigram :، bigram، و trigram لكل خوارزمية. تم مقارنة تقييمات الأداء وفقًا للدقة والاستدعاء و(القياس F) تم تحقيق نتائج التنبؤ بالقطبية في نماذج تحليل المشاعر بواسطة SVC الخطي باستخدام مجموعة بيانات الفندق مع حالة bigram ، بدقة 0.966 ، ودقة 0.967 ، واستدعاء 0.966 ، و(القياس F) 0.966. تعطي الخوارزميات الباقية أداءً متوسطًا في جميع مجموعات البيانات. يمكن أن نستنتج أن خوارزميات التعلم الآلي تحتاج إلى السمات المورفولوجية الصحيحة لتعزيز دقة التصنيف عند التعامل مع كلمات مختلفة تلعب أدوارًا مختلفة في الجملة ذات الأحرف نفسها.

الكلمات المفتاحية

تحليل المشاعر ، اللغة العربية ، التغريدات ، جائحة كوفيد -19