IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 36, No. 3, June, 2022, pp. 467-473
Journal homepage: http://iieta.org/journals/ria

Arabic Language Modeling Based on Supervised Machine Learning

Omayma Mahmoudi*, Mouncef Filali Bouami, Mustapha Badri

Laboratory of Applied Mathematics and Information Systems, Mohammed Premier University in Oujda, Multidisciplinary Faculty of Nador, Nador 60000, Morocco

Corresponding Author Email: mahmoudi.omayma@ump.ac.ma

https://doi.org/10.18280/ria.360315

Received: 31 May 2022
Accepted: 22 June 2022

Keywords:
machine learning, Arabic natural language processing, fake news, real news, COVID-19, vaccination

ABSTRACT

Misinformation and misleading actions have appeared as soon as COVID-19 vaccinations campaigns were launched, no matter what the country's alphabetization level or growing index is. In such a situation, supervised machine learning techniques for classification appears as a suitable solution to model the value & veracity of data, especially in the Arabic language as a language used by millions of people around the world. To achieve this task, we had to collect data manually from SM platforms such as Facebook, Twitter and Arabic news websites. This paper aims to classify Arabic language news into fake news and real news, by creating a Machine Learning (ML) model that will detect Arabic fake news (DAFN) about COVID-19 vaccination. To achieve our goal, we will use Natural Language Processing (NLP) techniques, which is especially challenging since NLP libraries support for Arabic is not common. We will use NLTK package of python to preprocess the data, and then we will use a ML model for the classification.

1. INTRODUCTION

The emergence of social networks facilitated communication between people and the dissemination of news in general. It allowed the generation and the exchange of ideas through publications and comments until it became one of the powerful tools that helped disseminate and transmit information, especially Fake News [1]. There are two types of Fake News [2]:

The first type is misinformation, which is unintended, so that a person can share information according to their ideas and the point of view. As for disinformation, it's role is to deliberately propagate deceptive information to mislead and harm people. In both cases, misinformation negatively affects people's experiences and decisions.

Through our work, we have exposed misinformation about coronavirus vaccine as they have a serious impact on the health of individuals. There are many studies [2] in multiple languages in this field, especially in English. This is in contrast to the Arabic language, where the use of ML to detect false news is uncommon, despite Arabic being one of the world's most widely spoken languages. That's why, we decided to do general processing of Arabic text.

Among the challenges that we have faced in this work is the lack of any data in news about the information related to coronavirus vaccines in Arabic, which we had to collect from various social networks and websites, where we were able to collect 1,000 pieces of information.

Then, comes the role of NLP, a relevant artificial intelligence (AI) subdomain [3], able to reconstruct the spoken or written human language using computational methods.

In this work, we will focus on Arabic natural language processing (ANLP). Arabic, a language we master as a mother tongue, contains some challenging specificities because of its complexity and lexical variations. ML simulates behavior through the use of learning algorithms that are adopted from vast amounts of data [4] that the computer learns and improves. As a result, the term "learning" was coined. where "learns" from data and extracts information from it. ML engines, on the other hand, are algorithms [4].

The social media interactions around COVID-19 vaccination campaigns are a representative example of a rising phenomenon on the internet: misinformation, and we believe it represents a real benchmark for the method deployed in this paper. NLP is applied for preprocessing the datasets with the aim of using ML to deduce which of the four algorithms: SVM [5, 6], Random Forest (RF) [7], Logistic Regression [8, 9], and Gradient Boosting [10] provides the best results.

Where they all excelled, RF has the highest proportion of F1-score ML classifiers with 91%.

This work is structured as follows. The related works are described in the second section and the ANLP approach is presented in the third section. The main topic of the fourth section is ML, as well as some of its algorithms. The fifth section presents our methodology and results. The final section discusses the consequences of the detection of the Arabic misinformation and the conclusion.

2. RELATED WORKS

Some studies classify misinformation detection as a hearsay resolution action with a multi-component model, such as misinformation identification, tracking, and stance categorization, which all assist in evaluating the veracity of disinformation. misinformation approaches can differ based on the data they're wanting (SM news such as comments, posts, and stories vs. large website articles), the ML algorithm used and the Language of preference. Some studies focus solely on the main tweet or post, while others look at other aspect of the

information, such as the conversation, responses, stories, and comments.

Al-Yahya et al. [11] compared and evaluated the performance of NN and transformer-based language models for AFND. They also conducted a thorough investigation into the likely causes of the disparities in performance outcomes produced by the various techniques. The results revealed that transformer-based models beat NN, with the best transformer-based model (QARiB) receiving 0.95 F1-score and the best percentage among neural networks (GRU) receiving 0.83.

As for de Oliveira et al. [12], they investigated data preprocessing methods in natural language, routing, dimensionality reduction, ML, and assessing the quality of information retrieval. They also developed a definition to contextualize misinformation and then discussed research initiatives and opportunities.

Bangyal et al. [13] started with data preprocessing using NLP, then applied a semantic model with TF-IDF weighting to describe the data, like with most research. They used eight machine learning methods for performance evaluation, including NB, Adaboost, CN, RF, LR, DT, NN, and SVM, as well as deep learning algorithms: RNN, CNN, GRU, and LSTM. They then trained and validated the rating model based on the results, and then tested it on a batch of unclassified COVID-19 fake news to forecast the sentiment category for each one.

Research paper [2] provided a conventional AFND architecture based on textual features only. where the following deep learning models were employed to evaluate the performance: CNN, LSTM, BLSTM, CNN + LSTM, and CNN + BLSTM. The studies used three data sets, each of which contained the textual content of Arabic news stories. So, when using both basic training modes and repetition in the training process, the BLSTM model surpasses the other models in terms of accuracy rate.

Based on our review of relevant works in the field of DAFN, it is clear that there is a limited amount of work in this area, thus, more research, investigation, and diversity are required. That is why we chose this topic to develop and cover the majority of its aspects.

In Table 1, we summarize the previous works using the different models.

Table 1. Comparative table of previous works

Previous studies	Models used	Best result
Al-Yahya et al. [11]	QARiB, GRU	0.95
Bangyal et al. [13]	CN, RF, LR, DT, NN, SVM, RNN, CNN, GRU, LSTM	0.97
Khalil et al. [2]	CNN, LSTM, BLSTM, CNN+LSTM, CNN+BLSTM.	0.83

3. ARABIC NATURAL LANGUAGE PROCESSING

Today, one of the most active areas of research in data science is NLP [14]. It is a field at the intersection of ML and linguistics. Its purpose is to extract information and meaning from textual content.

NLP is used in a variety of ways in our daily lives, including: text translation [15] (DeepL, for example), spell checker [16], automatic content summary [17], speech synthesis [18], text classification [19], opinion/sentiment analysis [20], prediction of the next word on a smartphone [21], named entity extraction from the text [22], etc.

Most of the resources available today are in English, which implies that most pre-trained models are also specific to the English language. The method of pre-processing each language differs from the other, since the Arabic language is one of the rich languages. That's why, we have adopted it in this work.

Among the steps of the Arabic language pre-processing are: tokenization, punctuation and special characters removal, Stopword removal, spell checking, named entity recognition and stemming.

To understand the steps of natural language pre-processing more closely, we extracted a sentence from the dataset. Where we applied all the steps mentioned earlier.

As shown in Figure 1:

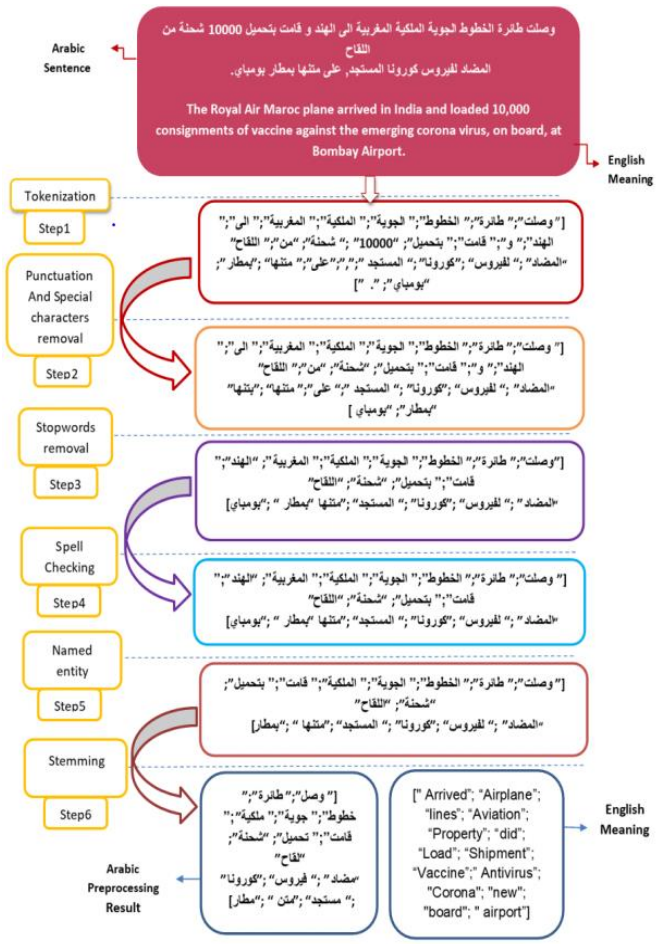

Figure 1. The Implementation of the different steps of ANLP

- Tokenization: used to divide the sentence into terms [23].
- Removal of punctuation and special characters: Used to delete special characters and punctuation (The number 10000 has disappeared).
- Stopword removal [24]: we also remove the empty words (we no longer see the words "و" (and), "الى" (to)).
- Spell-checking: its role is to check spelling.
- Named entity recognition: name entities are also removed ((Bombay)" بومباي" has disappeared).
- Stemming [25]: consists in cutting the end of the words in order to keep only the root of the word.

So far, we have only standardized the terms, which is insufficient for implementing ML techniques for classification, so it must first be converted into data that can be mathematically operated on. in this step, we finally manage to transform our words, now in the form of tokens, into numbers. There are several encoding methods, and each method has a different principle, such as binary, Bag-of-Words (BoW), and term frequency-inverse document frequency (TF-IDF) [26, 27].

To know how each vectorization pattern works, we have got Table 2, it consists of 3 sentences in the Arabic language. Each sentence is considered news [12].

3.1 Binary vector space model

Is the simplest vectorization model, in which a binary value has been allocated to each vector (With a value of 1 indicating that the phrase appeared in the document and a value of 0 indicating that it did not). As shown in Table 3, this representation method is inadequate in terms of semantics because it provides no insights into the role of a word for the set of texts. However, because it facilitates the generation of binary comparison masks, this representation paradigm is very beneficial for algorithms that processing techniques are applied to textual data. Furthermore, only a few computational resources are needed for this paradigm [12].

The *BoW* method consists of using the entire corpus of data to encode the sentences. As we know that our corpus of data will never change, we will use it as a reference base to create encodings for the sentences. The idea is to count the number of times each word in the corpus appears in each of the documents, as shown in Table 4.

Mathematically expressed as follows.

$$Vd=[W1, W2,....., Wn-1, Wn] \quad (1)$$

where, *Vd* is the weight vector *w* for each sentence in the document *d* up to the *n*-the term.

Each phrase entry relates to the count of the phrase entry that corresponds to it. For example, in the first sentence (which represents sentence T1), the first two inputs are "1,0":

The first input corresponds to the word "نصائح" (Tips) which is the first word of the sentence, and its value is "1" because "نصائح" (Tips) appears once in the first sentence of T1.

The second input matches the word "مرضى" (Patients), which is the second word in the list, and its value is "1" because "مرضى" (Patients) occurs once in the third sentence of T3.

The TF-IDF is a measurement that allows, from a set of texts, to know the relative importance of each word. This statistical assessment allows you to assess the significance of a term in a document in comparison to a collection or a corpus. The weight grows respect to the amount of times the term appears in the document. It differs as well depending on how frequently the word appears in the corpus.

Mathematically, the equation is:

$$\boldsymbol{Wi,j}=\boldsymbol{tfi,j}\times\log(\boldsymbol{N}/\boldsymbol{dfi}) \quad (2)$$

where, tfi=Number of occurrences of *i* in *j*; dfi=Number of documents containing i; *N*=Total number of documents.

The TF-IDF can be used to assess a document term's semantic relevance in relation to the entire collection. As shown in Table 5, the number of rows and columns is the same as in the BoW model. The TF-ISF is a variant of the original TF-IDF that is commonly used in summarizing texts at the sentence level rather than the document level.

Table 2. A toy example for vectorization techniques.

	Arabic example	**English meaning**
Text 1 (T1)	'نصائح', 'تعزيز', 'مناعة','ضد','كورونا','تجنب','الشائعة'	'Tips', 'boost', 'immunity', 'against', 'corona', 'avoid', 'rumour'
Text 2 (T2)	'الشائعة', 'حصل', 'لقاح', 'كورونا'	Rumours', 'happened', 'vaccine', 'corona'
Text 3 (T3)	'مرضى', 'كورونا', 'حصل', 'لقاح'	'Patients', 'Corona', 'Happened', 'Vaccine'

Table 3. The binary model vector representation of the sample corpus displayed in Table 2

Terms in Ar.	نصائح	مرضى	مناعة	ضد	كورونا	تجنب	شائعة	لقاح	حصل
Corr. Trans. to Eng.	Tips	Patients	immunity	against	corona	avoid	rumor	vaccine	Happened
T1	1	0	1	1	1	1	1	0	0
T2	0	0	0	0	1	0	1	1	1
T3	0	1	0	0	1	0	0	1	1

Table 4. *BoW* model representation of the collection apparent in Table 2

Terms in Ar.	نصائح	مرضى	مناعة	ضد	كورونا	تجنب	شائعة	لقاح	حصل
Corr. Trans. to Eng.	Tips	Patients	immunity	against	corona	avoid	rumor	vaccine	Happened
T1	1	0	1	1	1	1	1	0	0
T2	0	0	0	0	1	0	1	1	1
T3	0	1	0	0	1	0	0	1	1

Table 5. The TF-IDF model was used to represent the sample corpus in Table 2

Terms in Ar.	نصائح	مرضى	مناعة	ضد	كورونا	تجنب	شائعة	لقاح	حصل
Corr. Trans. to Eng.	Tips	Patients	immunity	against	corona	avoid	rumor	vaccine	happened
T1	0,60	0	0.73	0,90	0,90	0,73	0,60	0	0
T2	0	0	0	0	0,70	0	0,60	0,89	0,67
T3	0	0,80	0	0	0,93	0	0	0,78	0,81

4. MACHINE LEARNING

ML is a vital part of the rapidly expanding field of data science [4]. Using statistical methods [28], algorithms are taught to classify or predict data and identify key insights in data mining projects [29]. These findings are then used to inform decisions made within applications and businesses, with the goal of influencing key growth metrics.

There are two principal types of ML classifiers:

(1) *Supervised learning* is an automatic learning technique (ALT) where the aim is to automatically produce rules from a learning database containing "examples" (generally cases that have already been verified and processed).

(2) *Unsupervised Learning* identifies clusters or groups based on unlabeled data, with very little human intervention. Its capacity to find difference and similarity in data makes it an excellent categorization tool.

4.1 Random forest classifier

RF is a ML, deep learning, and artificial intelligence learning model that is used to make predictions. It is made up of several decision trees, each of which is focused on a different aspect of the problem. Each produces an estimate, and the overall estimate is determined by the assembly of the decision trees, where the analysis is performed. Subsets of decision trees are randomly assigned to each model [30].

A RF operates on the bagging principle. The first step is to split of a dataset into subgroups (decision trees), after which a training model is proposed for each group. Finally, the outcomes of these trees are combined to produce the most reliable prediction. The final result can be determined in one of two ways. The mean of the forecasts obtained is calculated in a regression RF. As a result, all of the decision tree predictions are taken into account. It can also make a reasonable prediction without hyper-parameter tuning, as we will see in the GridSearchCV section.

4.2 GridSearchCV [31]

When creating a RF, the number of decision trees and variables must be previously established. The grid search is used to test several parameters and find the one that is most useful. As a result, it's an optimization tool that's particularly useful for determining the best parameters for a forest of decision trees. So, the Grid search is a hyper-parameter optimization method that allows us to test a series of parameters and compare performance to determine the best setting.

The Grid Search is one of the simplest ways to test the parameters of a model among several options. For each parameter, we select a set of values to investigate.

5. METHODOLOGY & RESULT

As mentioned earlier, the goal of this research is to develop an automated classification model to classify Arabic fake news (AFN) based on an analysis of the Arabic text using NLP and supervised ML techniques. News articles containing any incorrect information are classified as "fake," while articles containing all verified (correct) information are classified as "real."

Because there are no AFN records about COVID-19 vaccination, we collected the data using hashtags and applied the ANLP method.

Methodological approach is composed of four basic steps. Like any ML model, we need a dataset to train a model to predict the class of data in the test database. The raw AFN dataset is the fruit of a collection and preparation work. Each piece of information was transformed into a feature vector and saved to an xlsx file. To our knowledge, and based on our research, a public dataset for AFN containing both fake and real news with enough characteristics does not exist. We then merged two collected databases "Getting Real about AFN " containing false and real news about the vaccination of COVID-19 and "All the news" containing the real news. These databases were obtained after collecting them from several sites and SM. Getting Real about AFN: It contains text and metadata extracted from SM (Facebook, Instagram,) and websites. To collect the dataset, we used the following hashtags present in Table 6:

Table 6. List of hashtags used to collect the dataset

#	Hashtags	English Translation
1	#لقاح_كورونا	Corona vaccine
2	#لقاح_فيروس_كورونا	Corona virus vaccine
3	#لقاح_كورونا_المستجد	New Corona vaccine
4	#لقاح_كوفيد19_	COVID 19 vaccine
5	#لقاح_فايزر	Pfizer vaccine

5.1 Data pre-processing

In order to create a NLP model, the first step is cleaning the data. The dataset can have unnecessary characters and missing values.

5.2 Data cleaning

- Deleting unnecessary characters: delete the following characters from text column: #.][! ()"".
- Deleting punctuations.
- Text correction.
- Deleting repeating characters [13].

5.3 Tokenizing

The purpose of tokenization is to break up text into smaller entities called tokens. The definition of what a token is, varies depending on the tokenizer used. A token can be a word, a character, or a sub-word. Tokenization is a fundamental step in every ANLP operation. Given the different existing linguistic structures, tokenization is different in each language.

5.4 Stopwords

It's common to remove Stopwords. Stopwords are considered common words in the language, but they do not have important information. Arabic Stopwords contain « ,"لكن" ل","ف ","و","من " " » in English ("But ", "of", "and", "q", "for").

Some words like « و », « كيف » in English ("How", "and") (have extremely high recurrence in all Arabic texts and provide no meaning as our model will use to make predictions. Removing them will reduce noise and let our model focus only on the relevant words. To do this, we will use a list and close

on all articles by deleting all the words that appear in the list [32].

5.5 Stemming

Stemming consists of reducing a word to its “root” form. The purpose of stemming is to group many variations of a word together as a single word. For example, once we apply a stemming to " يكتبون " (they write) or "كتب" (Wrote), the resulting word is the same. This makes it possible in particular to reduce the size of the vocabulary in BoW or TF-IdF type approaches.

One of the best-known stemmers is the Snowball Stemmer. This stemmer is available in Arabic language.

5.6 Data vectorization

The majority of ALT do not use text directly, but instead a digital vector converts the text to a digital vector based. We use the "bag-of-words" method to calculate the frequency of each term [13].

5.7 Model evaluation and testing

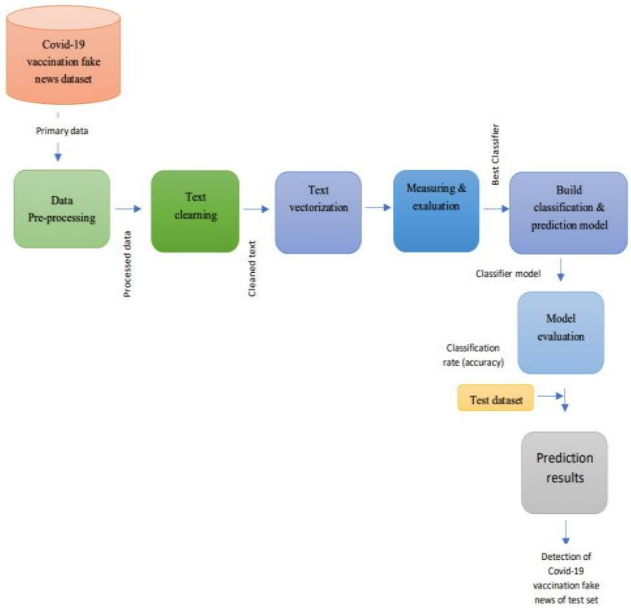

Figure 2. Research methodology

For training and evaluating our model, we use performance measures. Then, we test the model with a set of test data representing a set of unclassified AFN for COVID-19 vaccination, to predict the misinformation class of each dataset.

We calculated five metrics to evaluate performance. Precision is the percentage of related cases among all recovered occurrences, and recall basically divides the sum of the recovered relative documents. The F1 counts the average recall and precision score. Also, the confusion matrix which is a measurement of various parameters used to evaluate the performance of the classification algorithm [13].

Four different classifiers were used to analyze false information about COVID-19 vaccination. Which are: RF, SVM, LR, and GB. According to the results of this search, we find RF classifiers and logistic regression performance better than other learning methods as they achieved a very high accuracy of 91%.

The Figure 2, presents the structure of the classification system [13].

To evaluate the performance of the model, we will run the tests on a different set of tests to estimate the performance of the generalized model.

5.8 Evaluation metrics for ML algorithms

Evaluating ML algorithms is capital in the sense that a model may produce satisfactory results when tested against a metric such as accuracy score, but it may produce poor results when tested against other metrics. For example, logarithmic loss or any other metric of this type. The majority of the time, we utilize classification accuracy to evaluate the performance of our model; nevertheless, this is insufficient to fully analyze our model. This paragraph will go over the various types of evaluation metrics that are available [33].

Classification Accuracy: Classification when we use the term "accuracy," we usually mean "accuracy." It is the proportion of correct predictions to total sample count. It only works properly if each class has an equal number of samples.

Logarithmic Loss: Logarithmic Loss, also known as Log Loss, works by penalizing incorrect classifications. It is effective for multi-class classification. When employing Log Loss, the classifier must give a probability to each class for each sample.

There is no upper bound to Log Loss, and it exists in the range [0, ∞). A smaller Log Loss suggests more precision, while a higher Log Loss indicates lesser accuracy.

In general, minimizing Log Loss improves the classifier's performance.

Confusion Matrix: is produces a matrix as an output, which describes the model's overall performance. Let's pretend we're dealing with a binary classification problem. We have some samples that fall into one of two categories: yes or no. We also have our own classifier that predicts the class of an input sample.

There are 4 important terms:

(1) TP: (true positive) Correctly predicted the value as positive.

(2) TN: (true negative) Correctly predicted the value as negative.

(3) FP: (false positive) I mistakenly classified the value as positive, but they are actually negative.

(4) FN: (false negative) I mistakenly classified the value as negative, but they are actually positive. Various evaluation metrics are displayed in Table 7 [33].

The following Table 8 shows the four classification models selected for this work, representing a variety of machine learning techniques and have been confirmed utilizing metrics F1-score, we have obtained the following results:

From the previous Table 8, we can see that the results are too close, so actually any choice would be great. We have decided to use the RF Classifier since it got the highest cross validation F1 score.

Table 7. Evaluation metrics

Metric	Equation	Explanation
Accuracy (acc)	$\frac{TP+TN}{TP+TN+FP+FN}$	The percentage of correctly classified observations.
Precision (P)	$\frac{TP}{TP+FP}$	The percentage of positive classes correctly classified from a set of observations predicted to be positive.
Recall (R)	$\frac{TP}{TP+TN}$	The percentage of positive classes that were correctly classified from the total number of observations.
F1 Score	$\frac{2*P*R}{P+R}$	The harmonic mean of precision and recall metrics

Table 8. Model selection

Algorith.	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean
RF Classifier	0.9085	0.9085	0.9085	0.9085	0.9085	0.9185
LR	0.9111	0.9005	0.9247	0.9155	0.9159	0.9135
Gradient Boosting Classifier	0.9000	0.8795	0.9209	0.9234	0.9214	0.9090
Linear SVM	0.9039	0.8907	0.9180	0.9080	0.9000	0.9041

After tuning our model using GridSearchCV, we have tested multiple values for each parameter, and finally we have obtained the following optimized parameter. As shown in Figure 3:

```
RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                       criterion='gini', max_depth=11, max_features=14,
                       max_leaf_nodes=None, max_samples=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=3, min_samples_split=10,
                       min_weight_fraction_leaf=0.0, n_estimators=3000,
                       n_jobs=None, oob_score=False, random_state=0, verbose=0,
                       warm_start=False)
```

Figure 3. RF Parameters

Figure 4 present the following findings after training the model:

```
Model Report
Accuracy : 0.86
AUC Score (Train): 0.9413773815173911
cv Score:  0.6559616862403944
```

Figure 4. Model report

The average mean cross validation score is not enough to decide whether the model is good or not. In order to decide that, we have to use evaluation metrics: The F1-score obtained on our test set is 0.9. The recall obtained on our test set is 1 and the accuracy obtained: 0.86.

6. COMPARATIVE DISCUSSION

The results of all of the ML models that were tested can be found in the Table 7. On a vast set of data for COVID-19 vaccine AFND, we tested various models using the cross-validation in the training-test process. Although certain classifiers outperformed other ML classifiers in terms of F1-score, they all performed admirably. RF has the most F1-score ML classifiers, with 91%. 90% was attained using logistic regression and support vector machines.

7. CONCLUSIONS

The pace of information sharing has increased exponentially Throughout the previous decade thanks to the rapid adoption of SM platforms such as Facebook, Twitter and Instagram. Besides, this fast information growth, comes with side effects such as sharing AFN. Users are creating, consuming and sharing more information every day. The information could be real or not. AFN classification is a challenging research area, especially for languages that have few or no support at all of NLP libraries. In this paper, we have explored the methods and explained the pipeline we have used to create a ML model that would classify textual data to fake or real ones. We have trained the model using data that we collected manually from SM. In order to create the model, we have used python programming language and its libraries, and we have used four supervised learning models: Gradient Boosting and Support Virtual Machine. Finally, we have tuned its parameters using GridSearchCV. The model could be further improved by training it on more data, since our dataset is imbalanced. Getting more misinformation would help the model by getting more accurate results. We have talked about ANLP and its methods. Then we explained how we would apply those methods on our Arabic dataset. Finally, we have gone through the implementation and the results of the application.

We hope to use a bigger and more complicated dataset in the future, and It's also possible to increase the number of labels. Other languages can be included by using special characters and numeric values. Emoticons, which are frequently used in SM largely to represent expressions, would be useful to include.

REFERENCES

[1] Bondielli, A., Marcelloni, F. (2019). A survey on fake news and rumour detection techniques. Information Sciences, 497: 38-55. https://doi.org/10.1016/j.ins.2019.05.035

[2] Khalil, A., Jarrah, M., Aldwairi, M., Jararweh, Y. (2021). Detecting Arabic fake news using machine learning. In 2021 Second International Conference on Intelligent Data Science Technologies and Applications (IDSTA), pp. 171-177. https://doi.org/10.32604/cmc.2022.021449

[3] Chopra, A., Prashar, A., Sain, C. (2013). Natural language processing. International Journal of Technology Enhancements and Emerging Engineering research, 1(4): 131-134.

[4] Wei, J., Chu, X., Sun, X. Y., Xu, K., Deng, H.X., Chen, J., Wei, Z., Lei, M. (2019). Machine learning in materials science. InfoMat, 1(3): 338-358. https://doi.org/10.1002/inf2.12028
[5] Cortes, C., Vapnik, V. (1995). Support-vector networks. Machine Learning, 20: 273-297. https://doi.org/10.1007/BF00994018
[6] Burges, C.J.C. (1998). A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery volume, 2: 121-167. https://doi.org/10.1023/A:1009715923555
[7] Rigatti, S.J. (2017). Random forest. Journal of Insurance Medicine, 47(1): 31-39.
[8] Wright, R.E. (1995). Logistic regression. In L. G. Grimm, & P. R. Yarnold (Eds.), Reading and Understanding Multivariate Statistics (pp. 217-244). Washington DC: American Psychological Association.
[9] LaValley, M.P. (2008). Logistic regression. Circulation 2008; 117: 2395-2399. https://doi.org/10.1161/CIRCULATIONAHA.106.682658
[10] Natekin, A., Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in neurorobotics, 7: 21.
[11] Al-Yahya, M., Al-Khalifa, H., Al-Baity, H., AlSaeed, D., Essam, A. (2021). Arabic fake news detection: comparative study of neural networks and transformer-based approaches. Complexity, 2021: 5516945. https://doi.org/10.1155/2021/5516945
[12] de Oliveira, N.R., Pisa, P.S., Lopez, M.A., de Medeiros, D.S.V., Mattos, D.M. (2021). Identifying fake news on social networks based on natural language processing: trends and challenges. Information, 12(1): 38. https://doi.org/10.3390/info12010038
[13] Bangyal, W.H., Qasim, R., Ahmad, Z., Dar, H., Rukhsar, L., Aman, Z., Ahmad, J. (2021). Detection of fake news text classification on COVID-19 using deep learning approaches. Computational and Mathematical Methods in Medicine, 2021: 5514220. https://doi.org/10.1155/2021/5514220
[14] Chong, M., Specia, L., Mitkov, R. (2010). Using natural language processing for automatic detection of plagiarism. In Proceedings of the 4th International Plagiarism Conference (IPC-2010).
[15] Evans, J.A., Aceves, P. (2016). Machine translation: Mining text for social theory. Annual Review of Sociology, 42: 21-50. https://doi.org/10.1146/annurev-soc-081715-074206
[16] Altarawneh, R. (2017). Spelling detection errors techniques in NLP: A survey. Jordan – Alsalt, 172: 1-5.
[17] Garbade, M.J. (2018). A quick introduction to text summarization in machine learning. Towards Data Science.
[18] Ning, Y., He, S., Wu, Z.Y., Xing, C.X., Zhang, L.J. (2019). A review of deep learning based speech synthesis. Applied Sciences, 9(19): 4050. https://doi.org/10.3390/app9194050
[19] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., Brown, D. (2019). Text classification algorithms: A survey. Information, 10(4): 150. https://doi.org/10.3390/info10040150
[20] Kastrati, Z., Dalipi, F., Imran, A. S., Pireva Nuci, K., Wani, M.A. (2021). Sentiment analysis of students' feedback with NLP and deep learning: A systematic mapping study. Applied Sciences, 11(9): 3986. https://doi.org/10.3390/app11093986
[21] Hard, A., Rao, K., Mathews, R., et al. (2018). Federated learning for mobile keyboard prediction. arXiv preprint arXiv:1811.03604. https://arxiv.org/abs/1811.03604.
[22] Al-Smadi, M., Al-Zboon, S., Jararweh, Y., Juola, P. (2020). Transfer learning for Arabic named entity recognition with deep neural networks. IEEE Access, 8: 37736-37745. https://doi.org/10.1109/ACCESS.2020.2973319
[23] Webster, J.J., Kit, C. (1992). Tokenization as the initial phase in NLP. In COLING 1992 volume 4: The 14th international conference on computational linguistics.
[24] Al-Shalabi, R., Kanaan, G., Jaam, J. M., Hasnah, A., Hilat, E. (2004). Stop-word removal algorithm for Arabic language. In Proceedings. 2004 International Conference on Information and Communication Technologies: From Theory to Applications. https://doi.org/10.1109/ICTTA.2004.1307875
[25] Pande, B.A., Dhami, H.S. (2011). Application of natural language processing tools in stemming. International Journal of Computer Applications, 27(6): 14-19.
[26] Shahmirzadi, O., Lugowski, A., Younge, K. (2019). Text similarity in vector space models: a comparative study. In 2019 18th IEEE international conference on machine learning and applications (ICMLA), pp. 659-666. https://doi.org/10.1109/ICMLA.2019.00120
[27] Walkowiak, T., Datko, S., Maciejewski, H. (2018). Bag-of-words, bag-of-topics and word-to-vec based subject classification of text documents in polish-a comparative study. In International Conference on Dependability and Complex Systems, pp. 526-535. https://doi.org/10.1007/978-3-319-91446-6_49
[28] Boulesteix, A.L., Schmid, M. (2014). Machine learning versus statistical modeling. Biometrical Journal, 56(4): 588-593. https://doi.org/10.1002/bimj.201300226
[29] Raval, K.M. (2012). Data mining techniques. Gujarat – India.
[30] Mbaabu, O. (2020). Introduction to random forest in machine learning. https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning.
[31] Kartini, D., Nugrahadi, D. T., Farmadi, A. (2021). Hyperparameter tuning using GRIDSEARCHCV on the comparison of the activation function of the ELM method to the classification of pneumonia in toddlers. In 2021 4th International Conference of Computer and Informatics Engineering (IC2IE), pp. 390-395. https://doi.org/10.1109/IC2IE53219.2021.9649207
[32] Khanna, C. (2021). Text pre-processing: Stop words removal using different libraries. https://towardsdatascience.com/text-pre-processing-stop-words-removal-using-different-libraries-f20bac19929a/, accessed on Jul. 15, 2022.
[33] Brownlee, J. (2020). Tour of evaluation metrics for imbalanced classification. Machine Learning Mastery. https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/.