Future Computing and Informatics Journal

Volume 8
Issue 1 (2023) *Volume 8 Issue 1*

Article 2

News’ Credibility Detection on Social Media Using Machine Learning Algorithms

Farah Yasser
Business Information Systems, Faculty of Commerce and Business Administration, Helwan University, Egypt, farah.yasser21@commerce.helwan.edu.eg

Sayed AbdelMawgoud
Information Systems Department, Helwan University, Egypt, sgaber14@gmail.com

Amira M. Idrees AMI
Future University in Egypt, amira.mohamed@fue.edu.eg

Recommended Citation

Yasser, Farah; AbdelMawgoud, Sayed; and Idrees, Amira M. AMI () "News’ Credibility Detection on Social Media Using Machine Learning Algorithms," *Future Computing and Informatics Journal*: Vol. 8: Iss. 1, Article 2.
Available at: https://digitalcommons.aaru.edu.jo/fcij/vol8/iss1/2

Future Computing and Informatics Journal

Homepage: https://digitalcommons.aaru.edu.jo/fcij/

doi: http://Doi.org/10.54623/fue.fcij.8.1.2

News’ Credibility Detection on Social Media Using Machine Learning Algorithms

Farah Yasser
Business Information Systems,
Faculty of Commerce and Business
Administration,
Helwan University, Egypt
farah.yasser21@commerce.helwan.edu.eg

Sayed AbdelGaber AbdelMawgoud
Information Systems Department, Helwan
University, Egypt
sgaber14@gmail.com

Amira M. Idrees
Faculty of Computers and Information Technology,
Future University in Egypt, Egypt
amira.mohamed@fue.edu.eg

Abstract

Social media is essential in many aspects of our lives. Social media allows us to find news for free. anyone can access it easily at any time. However, social media may also facilitate the rapid spread of misleading news. As a result, there is a probability that low-quality news, including incorrect and fake information, will spread over social media. As well as detecting news credibility on social media becomes essential because fake news can affect society negatively, and the spread of false news has a considerable impact on personal reputation and public trust. In this research, we conducted a model that detects the credibility of Arabic news from social media; particularly Arabic tweets. The content of the tweets revolves around the COVID-19 pandemic. The proposed model applied to detect news credibility using text mining techniques and one of the well-known machine learning classifiers, Decision tree which has the best accuracy equal to 86.6%.

Keywords: social media, news credibility, text mining, and machine learning.

1.Introduction

Social networking sites has become a platform for spreading news and information between people over the world very fast, the progress of Users' interactions with others through social media have considerably risen as a result of social networking sites. it is true that the business sector now takes social networking seriously, the spreading of news via social media which has significantly impacted both people and business. [1], [2].

Within the rapid spreading of news and information, it has become difficult to differentiate between credible and non-credible news because of sharing other users' posts facilitates this and creates a cascade effect that might lead to the spread of false information. [2]. With the existence of COVID-19 pandemic, fake news increased extremely quickly, and individuals affected by it increased their fear and anxiety about this epidemic [3].

Data is a vital component in several disciplines [4]. While structured data has bottlenecks [5] [6], data expressed in a text format has additional issues[7]. This denotes the presence of a massive amount of disorganized data or information [7]. It might be useful data or information [8]. The challenge is converting unstructured to structured data and ordered data; Although, it may include information such as facts [9].

Text mining techniques and machine learning classifiers could help researchers evaluate the news credibility, Text mining is becoming more essential since it is applied in the extraction and organizing of text from unstructured data [10]. Machine learning is important because it offers organizations insights into patterns in customer behavior and firm operating procedures; it helps to categorize data and generate predictions; and it is used to make forecasts [11].

The proposed framework helps in determining the credibility of news. Based on data gathered from Twitter, the user's news is classified as true or false. The model applied decision tree algorithm and is evaluated to ensure prediction accuracy. Section 1 is an introduction to determine the credibility of news on social media via text mining techniques; Section 2 illustrated the previous study; Section 3 illustrated the proposed methodology and proposed model; Section 4 illustrated the experimental study for this paper; Section 5 envisions conclusions.

2.Related Works

M. A. Fadel (2020) [12] stated that the purpose of their study is to construct an initial dataset for determining the credibility of news in Arabic on social media applying classification model. There were 808 true tweets and 354 false tweets. The proposed technique entails creating a data collection of Arabic news on Twitter and applying several classifiers, such as Naive Bayes, random trees, and decision tables. It was discovered that the decision table classifier had a better accuracy of 81.46% using the relief algorithm. A study by [13] stated that the study's purpose is to enhance the accuracy of Arabic news using text mining techniques combined with natural language processing to enhance the quality of information and news available on social networking sites (NLP). Twitter data was gathered; the dataset included around 9000 tweets. Random Forest had the best classifier score of 76.17%.

M. Z. Sarwani, D. A. Sani and F. C. Fakhrin (2019) [14] Using a neural network algorithm, Social media can be used to measure one's identity and personality. The researchers obtained data from Facebook and examined the correlation and its bias towards Term Frequency- Inverse Document Frequency (TF-IDF). For classification, there are three phases to consider: text processing, weighting, and neural network classifiers. Text processing is classified into three types: tokenization, stopping words, and stemming. According to the findings of this research, the TF-IDF is used to assign a weighted value to each single word in text processing, neural network classification model is used to determine credibility, which had 60% accuracy.

G. Pasi, M. De Grandis and M. Viviani (2020) [15] developed a system built on news-related standards that improves the information's worth. The CredBank dataset was applied in this research to measure the suggested model through machine learning classifiers. Support Vector Machine, Naive Bayes, Decision Tree, and Random Forest are applied. The Python programming language was used for classification and testing. Random Forest has the best accuracy (79%). L. A, S. Y and S. R. K. T, (2019) [1] proposed a model to detect news credibility using text mining techniques and machine learning classifiers: random forest, decision tree, k-nearest neighbor, and support vector machine. The experiment was applied to a fake news dataset from Kaggle. The random forest classifier had the best accuracy of 90.7%. A scholar of [16] Arabic Fake News Detection Based on textual analysis, data were collected from articles about Haj. The study applied text preprocessing steps, feature selection using NLP POS tagger, machine learning classifier SVM, random forest, and nave base detecting the news credibility. Random forest achieved the best accuracy 78%.

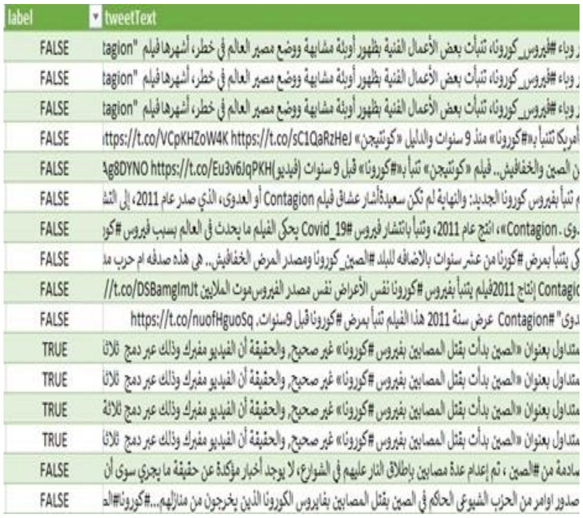
3.Proposed Methodology

In this research, the proposed methodology is to build a framework for determining the credibility of news on social media, mainly employing text mining algorithms to evaluate the credibility of tweets and applying several machine learning classifiers. It will be explained in the sections that follow. The developed framework is divided into five steps: data gathering, text processing that include two significant steps; text cleaning and preprocessing of text, extracting features, classification using machine learning, and evaluation results for detecting the credibility of news. In the following sections, the proposed model will be discussed in detail.

3.1 Dataset

Dataset downloaded from GitHub, it is a real data set, ArCOV-19 is an Arabic COVID-19 dataset that consists of Tweets that covers the months of January 27th to May 5th, 2021. ArCOV-19 is intended to support research in a wide range of areas, including Text mining, natural language processing, and social networking.

Figure 1 shows an example of the ArCOV-19 dataset.



label	tweetText
FALSE	وباء #فيروس_كورونا، تنبأت بعض الأعمال الفنية بظهور أوبئة مشابهة ووضع مصير العالم في خطر، أشهرها فيلم "tagion
FALSE	وباء #فيروس_كورونا، تنبأت بعض الأعمال الفنية بظهور أوبئة مشابهة ووضع مصير العالم في خطر، أشهرها فيلم "tagion
FALSE	وباء #فيروس_كورونا، تنبأت بعض الأعمال الفنية بظهور أوبئة مشابهة ووضع مصير العالم في خطر، أشهرها فيلم "tagion
FALSE	أمريكا تتنبأ بـ«#كورونا» منذ 9 سنوات والدليل «كونتيجن» ttps://t.co/VCpKHZoW4K https://t.co/sC1QaRzHeJ
FALSE	ن الصين والخفافيش.. فيلم «كونتيجن» تنبأ بـ«#كورونا» قبل 9 سنوات (فيديو)Ag8DYNO https://t.co/Eu3v6JqPKH
FALSE	م تنبأ بفيروس كورونا الجديد: والنهاية لم تكن سعيدةأشار عشاق فيلم Contagion أو العدوى، الذي صدر عام 2011، إلى النـ
FALSE	ـدوى . Contagion»، انتج عام 2011، وتنبأ بانتشار فيروس #Covid_19 يحكي الفيلم ما يحدث في العالم بسبب فيروس #كور
FALSE	ي يتنبأ بمرض #كورنا من عشر سنوات بالاضافه للبلد #الصين_كورونا ومصدر المرض الخفافيش.. هي هذه صدفه ام حرب مـ
FALSE	Contagio إنتاج 2011فيلم يتنبأ بفيروس #كورونا نفس الأعراض نفس مصدر الفيروس موت الملايين t.co/DSBamgImJt//
FALSE	دوى" #Contagion عرض سنة 2011 هذا الفيلم تنبأ بمرض #كورونا قبل 9سنوات. https://t.co/nuofHguoSq
TRUE	متداول بعنوان «الصين بدأت بقتل المصابين بفيروس #كورونا» غير صحيح, والحقيقة أن الفيديو مفبرك وذلك عبر دمج ثلاث
TRUE	متداول بعنوان «الصين بدأت بقتل المصابين بفيروس #كورونا» غير صحيح, والحقيقة أن الفيديو مفبرك وذلك عبر دمج ثلاث
TRUE	متداول بعنوان «الصين بدأت بقتل المصابين بفيروس #كورونا» غير صحيح, والحقيقة أن الفيديو مفبرك وذلك عبر دمج ثلاثة
TRUE	متداول بعنوان «الصين بدأت بقتل المصابين بفيروس #كورونا» غير صحيح, والحقيقة أن الفيديو مفبرك وذلك عبر دمج ثلاث
FALSE	صادمة من #الصين ، تم إعدام عدة مصابين بإطلاق النار عليهم في الشوارع، لا يوجد أخبار مؤكدة عن حقيقة ما يجري سوى أن
FALSE	صدور اوامر من الحزب الشيوعي الحاكم في الصين بقتل المصابين بفايروس الكورونا الذين يخرجون من منازلهم...#كورونا#الـ

Figure 1: A sample of the Dataset

3.2 Text Preprocessing

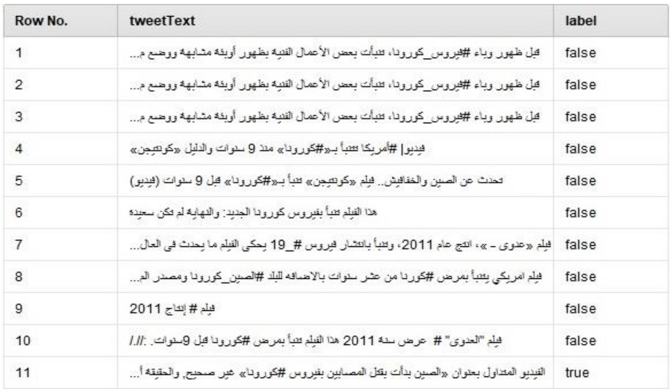
The second step consists of two stages: text cleaning and text preprocessing as shown in figure 1.

3.2.1 Text Cleaning

Text cleaning is a basic method that prepares data for analysis by changing or deleting data that is corrupted, incomplete, duplicate, or poorly structured [17]. It enhances accuracy after collecting text for detecting. also, data can be cleaned by deleting redundant samples and characteristics and removing missing values. The procedure of deleting outliers involves removing columns in the data that have the same values or will not affect the results. The process of minimizing redundancy data involves deleting duplicated rows from data [18].

Three steps were required to clean the tweets' text, the following will display these steps by using Rapid miner.

Step1: Remove English text and remove https links such as (https://t.co/a1sAWHhChA). Figure 2 shows the dataset after removing English text and HTTP links.

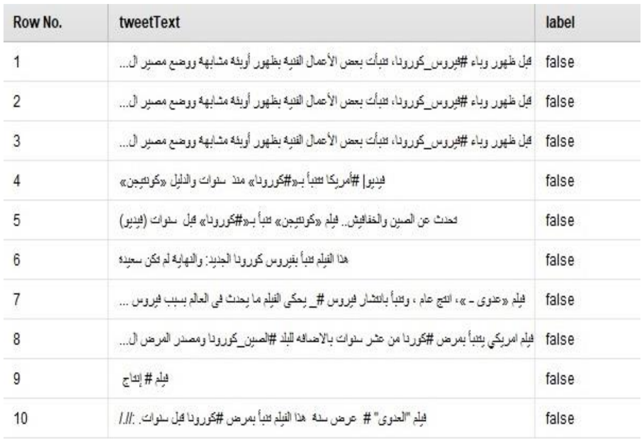


Row No.	tweetText	label
1	قبل ظهور وباء #فيروس_كورونا، تنبأت بعض الأعمال الفنية بظهور أوبئة مشابهة ووضع م...	false
2	قبل ظهور وباء #فيروس_كورونا، تنبأت بعض الأعمال الفنية بظهور أوبئة مشابهة ووضع م...	false
3	قبل ظهور وباء #فيروس_كورونا، تنبأت بعض الأعمال الفنية بظهور أوبئة مشابهة ووضع م...	false
4	فيديو| #أمريكا تتنبأ بـ«#كورونا» منذ 9 سنوات والدليل «كونتيجن»	false
5	تحدث عن الصين والخفافيش.. فيلم «كونتيجن» تنبأ بـ«#كورونا» قبل 9 سنوات (فيديو)	false
6	هذا الفيلم تنبأ بفيروس كورونا الجديد: والنهاية لم تكن سعيدة	false
7	فيلم «عدوى ـ »، انتج عام 2011، وتنبأ بانتشار فيروس #_19 يحكي الفيلم ما يحدث في العال...	false
8	فيلم امريكي يتنبأ بمرض #كورنا من عشر سنوات بالاضافه للبلد #الصين_كورونا ومصدر الم...	false
9	فيلم # إنتاج 2011	false
10	فيلم "العدوى" # عرض سنة 2011 هذا الفيلم تنبأ بمرض #كورونا قبل 9سنوات. ://./	false
11	الفيديو المتداول بعنوان «الصين بدأت بقتل المصابين بفيروس #كورونا» غير صحيح, والحقيقة أ...	true

Figure 2: A sample of dataset after removing English text

Figure 3 shows a sample of data from dataset after removing numbers from text.

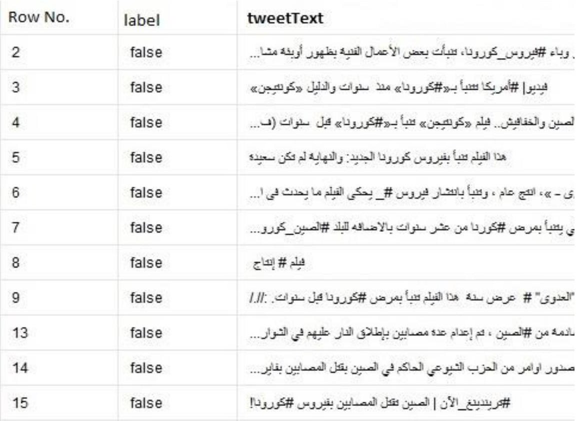
Step 2: Remove numbers such as [0-9]



Row No.	tweetText	label
1	قبل ظهور وباء #فيروس_كورونا، تنبأت بعض الأعمال الفنية بظهور أوبئة مشابهة ووضع مصير ال...	false
2	قبل ظهور وباء #فيروس_كورونا، تنبأت بعض الأعمال الفنية بظهور أوبئة مشابهة ووضع مصير ال...	false
3	قبل ظهور وباء #فيروس_كورونا، تنبأت بعض الأعمال الفنية بظهور أوبئة مشابهة ووضع مصير ال...	false
4	فيديو| #أمريكا تتنبأ بـ«#كورونا» منذ سنوات والدليل «كونتيجن»	false
5	تحدث عن الصين والخفافيش.. فيلم «كونتيجن» تنبأ بـ«#كورونا» قبل سنوات (فيديو)	false
6	هذا الفيلم تنبأ بفيروس كورونا الجديد: والنهاية لم تكن سعيدة	false
7	فيلم «عدوى ـ »، انتج عام ، وتنبأ بانتشار فيروس #_ يحكي الفيلم ما يحدث في العالم بسبب فيروس ...	false
8	فيلم امريكي يتنبأ بمرض #كورنا من عشر سنوات بالاضافه للبلد #الصين_كورونا ومصدر المرض ال...	false
9	فيلم # إنتاج	false
10	فيلم "العدوى" # عرض سنة هذا الفيلم تنبأ بمرض #كورونا قبل سنوات. ://./	false

Figure 3: A sample of data after removing numbers

Step 3: Duplicate tweets should be removed to ensure that the results are as accurate as possible. Figure 4 shows a sample from dataset after removing duplicates.



Row No.	label	tweetText
2	false	وباء #فيروس_كورونا، تنبأت بعض الأعمال الفنية بظهور أوبئة مشا...
3	false	فيديو| #أمريكا تتنبأ بـ«#كورونا» منذ سنوات والدليل «كونتيجن»
4	false	لصين والخفافيش.. فيلم «كونتيجن» تنبأ بـ«#كورونا» قبل سنوات (ف...
5	false	هذا الفيلم تنبأ بفيروس كورونا الجديد: والنهاية لم تكن سعيده
6	false	ى – »، انتج عام ، وتنبأ بانتشار فيروس #_ يحكى الفيلم ما يحدث فى ا...
7	false	ي يتنبأ بمرض #كورنا من عشر سنوات بالاضافه للبلد #الصين_كورو...
8	false	فيلم # إنتاج
9	false	"العدوى" # عرض سنة هذا الفيلم تنبأ بمرض #كورونا قبل سنوات. :/.//
13	false	ادمه من #الصين ، تم إعدام عده مصابين بإطلاق النار عليهم في الشوار...
14	false	صدور اوامر من الحزب الشيوعي الحاكم في الصين بقتل المصابين بفاير...
15	false	#تريندينغ_الآن | الصين تقتل المصابين بفيروس #كورونا!

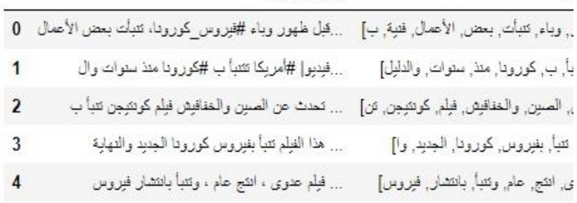
Figure 4: A sample of dataset after removing duplicates

3.2.2 Text Preprocessing

The text obtained from social media sites, such as Tweets, is unstructured. It contains unusual text and symbols that must be cleaned before a machine learning model can comprehend it. Text preprocessing is just as important as building a complex machine learning model. The trustworthiness of your model is highly dependent on the quality of your text [19], [20].

Five steps were required to this dataset for text processing.

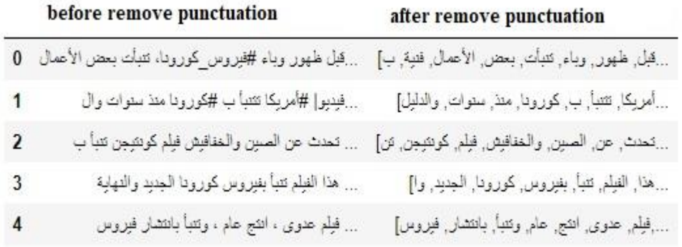
Step 1: Tokenization is a fundamental step for working with text-based data. Tokenization is the act of dividing a sentence, phrase, and essay, such as single words or phrases. These smaller pieces are referred to as tokens [21]. Figure 5 shows an example of tokenized tweets.



	tweetText	t
0	...قبل ظهور وباء #فيروس_كورونا، تنبأت بعض الأعمال	، وباء, تنبأت, بعض, الأعمال, فنية, ب]
1	...فيديو| #أمريكا تتنبأ ب #كورونا منذ سنوات وال	أ, ب, كورونا, منذ, سنوات, والدليل]
2	... تحدث عن الصين والخفافيش فيلم كونتيجن تنبأ ب	، الصين, والخفافيش, فيلم, كونتيجن, تن]
3	... هذا الفيلم تنبأ بفيروس كورونا الجديد والنهاية	تنبأ, بفيروس, كورونا, الجديد, وا]
4	... فيلم عدوى ، انتج عام ، وتنبأ بانتشار فيروس	ى, انتج, عام, وتنبأ, بانتشار, فيروس]

Figure 5: A sample of tokenized data

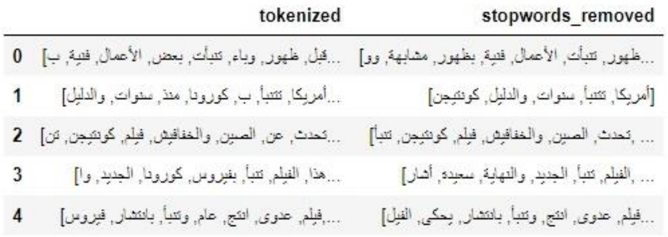
Step 2: remove punctuation such as **[!"#$%&'() *+, /:;<=>?@\[\\\]_`{|}~]** because of concentration on the text itself, not the punctuation or emotions. Figure 6 displays the difference between before and after removing punctuation.



	before remove punctuation	after remove punctuation
0	...قبل ظهور وباء #فيروس_كورونا، تنبأت بعض الأعمال	...قبل, ظهور, وباء, تنبأت, بعض, الأعمال, فنية, ب]
1	...فيديو| #أمريكا تتنبأ ب #كورونا منذ سنوات وال	...أمريكا, تتنبأ, ب, كورونا, منذ, سنوات, والدليل]
2	... تحدث عن الصين والخفافيش فيلم كونتيجن تنبأ ب	...تحدث, عن, الصين, والخفافيش, فيلم, كونتيجن, تن]
3	... هذا الفيلم تنبأ بفيروس كورونا الجديد والنهاية	...هذا, الفيلم, تنبأ, بفيروس, كورونا, الجديد, وا]
4	... فيلم عدوى ، انتج عام ، وتنبأ بانتشار فيروس	...,فيلم, عدوى, انتج, عام, وتنبأ, بانتشار, فيروس]

Figure 6: a sample before and after remove punctuation

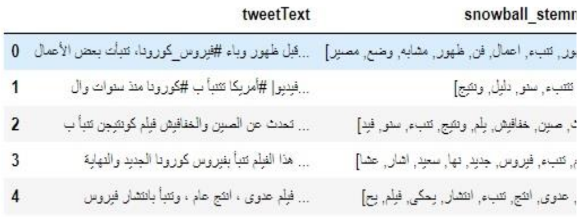
Step 3: filter stop words allows you to enable or disable stop word filtering. Stop words are words that are rarely used as classification features. Stop words are typically high frequency words [22] [23]. The Arabic stop words were filtered and removing custom stop words as ["وباء,"عام" ,كورونا" "كوفيد" ,"فيروس",""]. Figure 7 shows an example after filtering stop words



	tokenized	stopwords_removed
0	...قبل, ظهور, وباء, تنبأت, بعض, الأعمال, فنية, ب]	...ظهور, تنبأت, الأعمال, فنية, بظهور, مشابهة, وو]
1	...أمريكا, تتنبأ, ب, كورونا, منذ, سنوات, والدليل]	[أمريكا, تتنبأ, سنوات, والدليل, كونتيجن]
2	...تحدث, عن, الصين, والخفافيش, فيلم, كونتيجن, تن]	..., تحدث, الصين, والخفافيش, فيلم, كونتيجن, تنبأ]
3	...هذا, الفيلم, تنبأ, بفيروس, كورونا, الجديد, وا]	..., الفيلم, تنبأ, الجديد, والنهاية, سعيده, أشار]
4	...,فيلم, عدوى, انتج, عام, وتنبأ, بانتشار, فيروس]	...فيلم, عدوى, انتج, وتنبأ, بانتشار, يحكى, الفيل]

Figure 7: A sample after filtering stop words

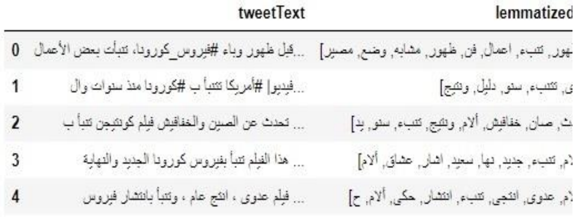
Step 4: stemming (Arabic) is the process of remove any kind of suffix from a word and return it to its initial form, which is the root word, we applied on Arabic dictionary [22] [24]. Figure 8 shows text after stemming process.



	tweetText	snowball_stemm
0	...قبل ظهور وباء #فيروس_كورونا، تنبأت بعض الأعمال	ور, تنبء, اعمال, فن, ظهور, مشابه, وضع, مصير]
1	...فيديو| #أمريكا تتنبأ ب #كورونا منذ سنوات وال	تتنبء, سنو, دليل, ونتج]
2	... تحدث عن الصين والخفافيش فيلم كونتيجن تنبأ ب	ت, صين, خفافيش, يلم, ونتج, تنبء, سنو, فيد]
3	... هذا الفيلم تنبأ بفيروس كورونا الجديد والنهاية	م, تنبء, فيروس, جديد, نها, سعيد, اشار, عشا]
4	... فيلم عدوى ، انتج عام ، وتنبأ بانتشار فيروس	, عدوى, انتج, تنبء, انتشار, يحكى, فيلم, يج]

Figure 8: A sample of stemmed text

Step 5: lemmatization (Arabic): In natural language processing, lemmatization entails organizing words based on their root lexical components. It is used in computer programming and artificial intelligence for natural language processing and understanding [25] [26]. Figure 9 shows the text after using stemming and lemmatization.



	tweetText	lemmatized
0	...قبل ظهور وباء #فيروس_كورونا، تنبأت بعض الأعمال	هور, تنبء, اعمال, فن, ظهور, مشابه, وضع, مصير]
1	...فيديو| #أمريكا تتنبأ ب #كورونا منذ سنوات وال	ى, تتنبء, سنو, دليل, ونتج]
2	... تحدث عن الصين والخفافيش فيلم كونتيجن تنبأ ب	ث, صان, خفافيش, ألام, ونتج, تنبء, سنو, يد]
3	... هذا الفيلم تنبأ بفيروس كورونا الجديد والنهاية	لام, تنبء, جديد, نها, سعيد, اشار, عشاق, ألام]
4	... فيلم عدوى ، انتج عام ، وتنبأ بانتشار فيروس	لام, عدوى, انتجى, تنبء, انتشار, حكى, ألام, ح]

Figure 9: A sample of stemmed and lemmatized text

The preprocessing step is applied twice, the first-time applying stemming, and the second time applying stemming and lemmatization.

3.4 Feature Selection

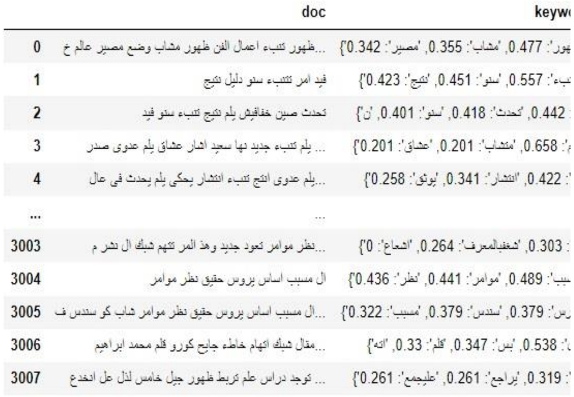
Feature selection is used in the machine learning process to enhance accuracy. Also, it improves the algorithms' prediction power by choosing the most important variables and removing the redundant and useless ones [27]. In this model we applied TF-IDF Vectorizer.

The TF-IDF technique is designed to compute word frequency, The TF-IDF score is subsequently applied to each document. Word frequency is used to find keywords that are more significant (occur more frequently) in a document, use TF-IDF. The TF-IDF Vectorizer converts documents into tokens, learns vocabulary, and reverses the frequency weightings of texts[28].

The TF-IDF is a weighing matrix that is used to determine the importance of a phrase (count + weight) to a document in a dataset. Tokens retrieved from text data that use the TF-IDF and count vector procedures are identical; although, the term frequency (TF) and inverse document frequency (IDF) metrics are combined in TF-IDF (IDF)[29]. Equation represents the TF-IDF equation (1).

$$TF/IDF = tf(t,d) \times idf(t,d) \qquad (1)$$

Studying how TF-IDF works will help obtain a better understanding of how machine learning algorithms operate. While machine learning algorithms are consistently better at operating with numbers, TF-IDF approaches assist them in deciphering words by assigning them a numerical value or vector. it improves the performance of machine learning classifiers [30], [31]. The machine learning classifiers use the keywords depend on their weights, Figure 10 displays an example of the text weight of the applied dataset using TF-IDF.



```
                                                    doc    keyw
0     ظهور تتبء اعمال الفن ظهور مشاب وضع مصير عالم خ...    هور': 0.477, 'مشاب': 0.355, 'مصير': 0.342}
1                        فيد امر تتبء ستو دليل نتيج    تبء': 0.557, 'ستو': 0.451, 'نتيج': 0.423}
2               تحدث صين خفافيش يلم نتيج تتبء ستو فيد    : 0.442, 'تحدث': 0.418, 'ستو': 0.401, 'ن'}
3     ... يلم تتبء جديد دها سعيد اشار عشاق يلم عدوى صدر    م': 0.658, 'متشاب': 0.201, 'عشاق': 0.201}
4       يلم عدوى انتج تتبء انتشار يحكي يلم يحدث في عال...    : 0.422, 'انتشار': 0.341, 'يوثق': 0.258}
...                                                 ...
3003  نظر موامر تعود جديد وهذ المر تتهم شبك ال نشر م...    : 0.303, 'شغفبالمعرف': 0.264, 'اشعاع': 0}
3004                  ال مسبب اساس يروس حقيق نظر موامر    سبب': 0.489, 'موامر': 0.441, 'نظر': 0.436}
3005  ال مسبب اساس يروس حقيق نظر موامر شاب كو سندس ف...    رس': 0.379, 'سندس': 0.379, 'مسبب': 0.322}
3006  مقال شبك اتهام خاطء جايح كورو قلم محمد ابراهيم...    : 0.538, 'بس': 0.347, 'قلم': 0.33, 'اته'}
3007  ... توجد دراس علم تربط ظهور جيل خامس لذل عل انخدع    : 0.319, 'يراجع': 0.261, 'عليجمع': 0.261}
```

Figure 10: A sample of text weight

3.5 Credibility Detection using Decision Tree Algorithm

In this paper, Decision Tree algorithm is applied to detect the credibility of news as will be explained in the following.

A study by [33] stated that: "DTs are a non-parametric supervised learning method used for classification; it predicts the value of a target variable by learning simple decision rules inferred from the data features."

Decision trees, which are termed white box ML algorithms, employ internal decision-making logic that obtained information from a data set may be simply retrieved in a comprehensible manner, DT demands very little work from its users for data preparation and analysis.

The decision tree employs a non-parametric technique, which means it is not dependent on probability distribution assumptions and is distribution-free. It has remarkable accuracy while working with high-dimensional data [35]

4. Experimental study

In the experiment, Precision, recall, Accuracy and F1-score were the methods used to measure the performance of the applied classifiers in this model using Python; 30% of the dataset was randomly selected for training, and the accuracy of each classifier method will be measured. The results will be illustrated. Decision Tree classifier is used twice in this experiment, once before and once after applying lemmatization, and the results are shown in the tables below: table1, table 2, table 3 and table 4.

A-Results before using lemmatization

Table 1: illustrates the result of decision tree classifier before using lemmatization for credible and non-credible news.

Table 1: DT results of credible and non-credible news before using Lemmatization

Decision Tree Classifier using Stemming only				
matrices	Precision	Recall	F1-Score	Support
Credible	0.89	0.86	0.88	463
Non-Credible	0.87	0.90	0.89	496

Table 2: illustrates the result of the decision tree classifier before using lemmatization for average results and measuring accuracy.

Table 2: DT average results before using Lemmatization

Decision Tree Classifier using Stemming only				
matrices	Precision	Recall	F1-Score	Accuracy
Average	0.874	0.901	0.887	0.882

The confusion matrix of the applied decision tree classifier before using lemmatization in the previous figure 11.

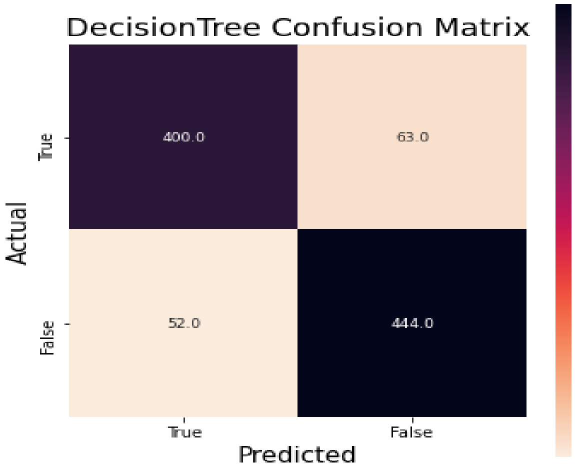


Figure11: CM of DT before using lemmatization

B-Results after using lemmatization

Table 3: illustrates the result of decision tree classifier after using lemmatization for credible and non-credible news.

Table 3: DT results of credible and non-credible news after using Lemmatization

Decision Tree Classifier using lemma				
matrices	Precision	Recall	F1-Score	Support
Credible	0.86	0.86	0.86	463
Non-Credible	0.87	0.87	0.87	496

Table 4: illustrates the result of the decision tree classifier after using lemmatization for average results and measuring accuracy.

Table 4: DT average results after using Lemmatization

Decision Tree Classifier using lemma				
matrices	Precision	Recall	F1-Score	Accuracy
Average	0.88	0.868	0.87	0.866

The confusion matrix of the applied decision tree classifier after using lemmatization in the previous figure12.

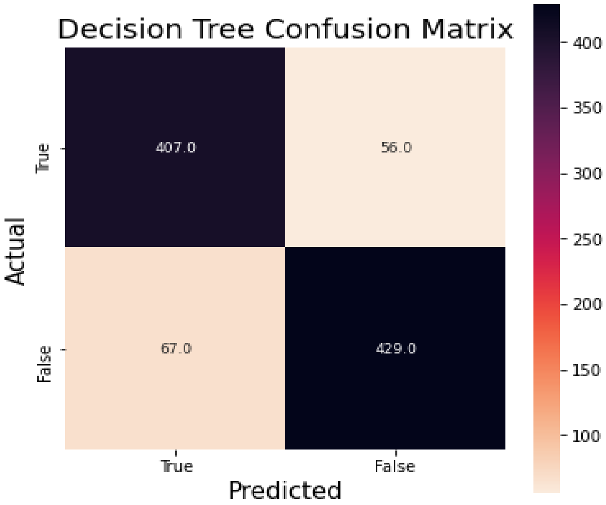


Figure 12: CM of DT after using lemmatization

5. Conclusion and future works

This paper aims to detect news credibility on social media from Twitter by using decision tree algorithm. The model is applied twice: before applying lemmatization and after in the preprocessing stage. Lemmatization produced better results equal 86.6%.

Future work can be conducted to classify Arabic by using different classifiers machine learning algorithms and compare between results. Using franco Arabic dataset to determine their credibility of news. Using alternative methods in feature selection, as information gain and information gain ratio, and comparing outcomes, applying data sets from multiple social media platforms, such as Facebook and Instagram, rather than Twitter.

References:

[1] L. A, S. Y and S. R. K. T, "An Effecient Fake News Detection System," *International Journal of Innovative Technology and Exploring Engineering, 2019,* vol. 10, no. 8, p. 5.

[2] K. Ali and C. Li, "Fake news on Facebook: examining the impact of heuristic cues on perceived credibility and sharing intention," *Internet Research, 2020,* vol. 22, no. 1, p. 19.

[3] O. D. Apuke and B. Omar, "Fake news and COVID-19: modelling the predictors of fake news," *Telematics and Informatics, 2021,* vol. 56, p. 16.

[4] A. Mostafa, A. E. Khedr and A. Abdo, "Advising Approach to Enhance Students' Performance Level in Higher Education Environments," *Journal of Computer Science, 2017,* vol. 13, no. 5, pp. 130-139.

[5] A. Khedr, S. Kholeif and S. Hossam, "Enhanced Cloud Computing Framework to Improve the Educational Process in Higher Education: A case study of Helwan University in Egypt," *INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY, 2015,* vol. 14, no. 6.

[6] A. E. Khedr, S. A. Kholeif and S. H. Hessen, "Adoption of cloud computing framework in higher education to enhance educational process," *International Journal of Innovative Research in Computer Science and Technology (IJIRCST), 2015,* vol. 3, no. 3, pp. Pp. 150 -156.

[7] M. Othman, H. Hassan, R. Moawad and A. M. Idrees, "A linguistic approach for opinionated documents summary," *Future Computing and Informatics Journal, 2018,* vol. 3, no. 2, pp. 152-158.

[8] H. A. Hassan, M. Y. Dahab,, . K. Bahnassy, A. M. Idrees and F. Gamal, "Arabic Documents classification method a Step towards Efficient Documents Summarization," *International Journal on Recent and Innovation Trends in Computing and Communication, 2015,* p. 351-359.

[9] H. A. Hassan, M. Y. Dahab, K. Bahnasy, A. M. Idrees and F. Gamal, "Query answering approach based on document summarization," *International Open Access Journal of Modern Engineering Research, 2014,* vol. 4, no. 12.

[10] A. Abed, J. Yuan and L. Li, "A Review of Towered Big-Data Service Model for Biomedical Text-Mining Databases," *International Journal of Advanced Computer Science and Applications ,* vol. 8, no. 8, 2017.

[11] Z. Wang, X. Peng, A. Xia, A. Shah, Y. Huang, X. Zhu and Q. Liao, "The role of machine learning to boost the bioenergy and biofuels conversion," *Bioresource Technology,* 2022.

[12] M. A. Fadel, "Evaluating the Credibility of Arabic News in Social Media through the use of Advanced Classifier Algorithms," *International Journal of Advanced Trends in Computer Science and Engineering, 2020,* vol. 9, no. 4, p. 15.

[13] R. Mouty and A. Gazdar, "The Effect of the Similarity Between the Two Names of Twitter Users on the Credibility of Their Publications," in *Joint 2019 8th International Conference on Informatics, Electronics & Vision (ICIEV) & 3rd International Conference on Imaging, Vision & Pattern Recognition (IVPR). Institute of Electrical and Electronics Engineers.Inc,* 2019.

[14] M. Z. Sarwani, D. A. Sani and F. C. Fakhrin, "Personality Classification through Social Media Using Probabilistic Neural Network Algorithms," *International Journal of Artificial Intelligence & Robotics (IJAIR), 2019*, vol. 1, no. 1, p. 7.

[15] G. Pasi, M. De Grandis and M. Viviani, "Decision Making over Multiple Criteria to Assess News Credibility in Microblogging Sites," in *In IEEE Conference on Fuzzy Systems. Institute of Electrical and Electronics Engineers Inc.*, 2020.

[16] H. Himdi, G. Weir, F. Assir and H. Al-Barhamtoshy, "Arabic Fake News Detection Based on Textual Analysis," *Arabian Journal for Science and Engineering, 2022*, p. 17.

[17] K. Sharma and N. Garg, "Text pre-processing of multilingual for sentiment analysis based on social network data.," *International Journal of Electrical and Computer Engineering, 2022*, vol. 12, no. 1, pp. 776–784.

[18] T.-Y. Ou, G.-Y. Lin, H.-P. Fu, S.-C. Wei and W.-L. Tsai, "An Intelligent Recommendation System for Real Estate Commodity," *Tech Science Press, 2022*, vol. 42, no. 3, p. 17.

[19] G. Singhal, "https://www.pluralsight.com/," 5 oct 2020. [Online].

[20] A. Rusli, N. M. S. Iswari and R. Setiabudi, "Enhancing text classification performance by preprocessing," *Telecommunication, Computing, Electronics and Control, 2021*, vol. 19, no. 4, p. 8.

[21] S.Singh,"https://www.analyticsvidhya.com/," 18 July 2019. [Online].

[22] . V. Baradad and A. Mugabushaka, "Corpus specific stop words to improve the textual analysis in scientometrics," in *In: International Conference on Science in Information*, 2015.

[23] R. Garreta, "https://help.monkeylearn.com/," August 2022. [Online].

[24] N.Arora,"https://www.analyticsvidhya.com/," 15 June 2021. [Online].

[25] S. Gotovac, A. Doko and I. Boban, "Sentence retrieval using Stemming and Lemmatization with Different Length of the Queries," *Advances in Science, Technology and Engineering Systems Journal*, vol. 5, no. 3, p. 6, 2020.

[26] "What Does Lemmatization Mean?," [Online]. Available: https://www.techopedia.com/.

[27] S. Kaushik , "Introduction to Feature Selection methods with an example (or how to select the right variables?)," 2020. [Online]. Available: https://www.analyticsvidhya.com/.

[28] d. L. Salge, C. Alves and N. Berente, ""Is that social bot behaving unethically?."," *Communications of the ACM, 2017*, vol. 60, p. 5.

[29] S. Kaur, P. Kumar and P. Kumaraguru, "Automating fake news detection system using multi-level voting," *Soft Computing, November 2019*, p. 21.

[30] B. Stecanella, "Understanding TF-ID: A Simple Introduction," 10 may 2019. [Online]. Available: https://monkeylearn.com/blog/what-is-tf-idf/.

[31] M. D. R. Wahyudi, "Evaluation of TF-IDF Algorithm Weighting Scheme in The Qur'an Translation Clustering with K-Means Algorithm," *Journal of Information Technology and Computer Science, 2021*, vol.6, p. 13.

[32] e. a. Pedregosa, "Scikit-learn: Machine Learning in Python," 2011. [Online]. Available:https://scikit-learn.org/stable/modules/svm.html.

[33] S. Almatarneh, P. Gamallo, F. R. Pena and A. Alexeev, "Supervised Classifiers to Identify Hate Speech on English and," in *In: Jatowt A., Maeda A.,Syn S. (eds) Digital Libraries at the Crossroads of Digital Information for the ICADL 2019*, 2019.

[34] O. Mbaabu, "Introduction to Random Forest in Machine Learning," December 2020. [Online].Available: https://www.section.io/engineering-education.

[35] N. Duggal, "Advantages of Decision Trees," 2022september. [Online]. Available: https://www.simplilearn.com/advantages-of-decision-tree-article.

[36] "Neural Networks," 2020. [Online]. Available: https://www.ibm.com/.

[37] J. Mahanta, "Introduction to Neural Networks, Advantages and Applications," *Towards Data Science*, 2017.

[38] N. Sultan, A. E. Khedr, A. M. Idrees and S. Kholeif, "Data Mining Approach for Detecting Key Performance Indicators," *Journal of Artificial Intelligence, 2017,* vol. 10, no. 2, pp. 59-65.

[39] A. M. Mohsen, A. M. Idrees and H. A. Hassan, "Emotion Analysis for Opinion Mining From Text: A Comparative Study," *International Journal of e-Collaboration, 2019,* vol. 15, no. 1.