Karbala International Journal of Modern Science

Volume 8 | Issue 3 Article 18

Improving Prediction of Arabic Fake News Using Fuzzy Logic and Modified Random Forest Model

Tahseen A. Wotaifi
College of Infomation Technology, University of Babylon,, tahseen.ali@uobabylon.edu.iq

Ban N. Dhannoon
Department of Computer Science, College of Science, Al-Nahrain University

Recommended Citation

Wotaifi, Tahseen A. and Dhannoon, Ban N. (2022) "Improving Prediction of Arabic Fake News Using Fuzzy Logic and Modified Random Forest Model," *Karbala International Journal of Modern Science*: Vol. 8 : Iss. 3 , Article 18.
Available at: https://doi.org/10.33640/2405-609X.3241

Improving Prediction of Arabic Fake News Using Fuzzy Logic and Modified Random Forest Model

Abstract

Throughout the last few years, the world is witnessing the so-called age of social media, as there is a complete dependence on these sites for following up on events and activities. The problem is that the misinformation or fake news is always released at the appropriate time, so this false news spreads quickly and takes a very wide resonance. Although several studies are performed to determine English fake news, the identification of Arabic misinformation remains underdeveloped. This study aims to build an improved learning model for detecting fake news in the Arabic language. Unlike previous studies that depended on analyzing the content of the tweet only, this study focuses on the text, user features, and text features. Regarding the content of the tweet, the TF-IDF method was used to convert the words into features and then determine the features that have a high rank. In contrast, a fuzzy model was used to determine the relevant features for the user. Finally, the random forest algorithm has been adapted and improved, and its results are better as compared to other machine learning methods. The accuracy of Improved Random Forest is (0.895), whereas the accuracy of Naive Bayesian and SVM techniques are found to be (0.809) and (0.848), respectively.

Keywords

Arabic Fake News, User-Based Features, Content-Based Features, Fuzzy Model, and Improved Random Forest Algorithm.

Creative Commons License

RESEARCH PAPER

Improving Prediction of Arabic Fake News Using Fuzzy Logic and Modified Random Forest Model

Tahseen A. Wotaifi [a,*], Ban N. Dhannoon [b]

[a] College of Information Technology, University of Babylon, Hillah, Babil, Iraq
[b] Department of Computer Science, College of Science, Al-Nahrain University, Baghdad, Iraq

Abstract

Throughout the last few years, the world is witnessing the so-called age of social media, as there is a complete dependence on these sites for following up on events and activities. The problem is that the misinformation or fake news is always released at the appropriate time, so this false news spreads quickly and takes a very wide resonance. Although several studies are performed to determine English fake news, the identification of Arabic misinformation remains underdeveloped. This study aims to build an improved learning model for detecting fake news in the Arabic language. Unlike previous studies that depended on analyzing the content of the tweet only, this study focuses on the text, user features, and text features.

Regarding the content of the tweet, the TF-IDF method was used to convert the words into features and then determine the features that have a high rank. In contrast, a fuzzy model was used to determine the relevant features for the user. Finally, the random forest algorithm has been adapted and improved, and its results are better as compared to other machine learning methods. The accuracy of Improved Random Forest is (0.895), whereas the accuracy of Naive Bayesian and SVM techniques are found to be (0.809) and (0.848), respectively.

Keywords: Arabic fake news, User-based features, Content-based features, Fuzzy model, Improved random forest algorithm

1. Introduction

One of the most important means of exchanging information between people or users nowadays is social media. A commonly used social media platform is Twitter, which is adopted by millions of users [1]. It is even estimated that the number of users who visit it daily reaches three million visitors. The reason behind the tendency of online users to this platform is that it provides a channel that allows users to easily create, publish and share information [2]. Tweets on Twitter often represent users' opinions, or they may also include news on a specific topic or field [3].

Not everything that is said on these platforms is true, as many people try to mislead others through their posts for their own purposes [4]. They usually spread misleading news at the right time, so it is often seen that it takes a very wide resonance and spreads rather quickly [5].

Fake news or misinformation can be described as the untrue comments or posts that are launched on social media platforms by some people. On the other hand, Fake News Detection “FND” is the prediction of the probability that an article, tweet, or story is intentionally deceptive [6]. Detecting fake news manually is slow, tedious, and impractical, so researchers in recent years have turned towards building models using machine learning techniques to discover this type of news [7]. Some other terms have a meaning similar to or close to FND such as rumor detection, rumor classification, stance classification of articles, claim verification, and misleading information detection [8]. All these terms have received wide attention from researchers to reveal the facts and not mislead users [9].

Received 16 April 2022; revised 30 May 2022; accepted 2 June 2022.
Available online 1 August 2022

* Corresponding author.
E-mail addresses: tahseen.ali@uobabylon.edu.iq (T.A. Wotaifi), ban.n.dhannoon@nahrainuniv.edu.iq (B.N. Dhannoon).

https://doi.org/10.33640/2405-609X.3241

All of the foregoing aspects are considered the main motives behind paying attention to this work and reducing the problems of false news. Therefore, this work aims to build improved models for predicting Arab fake news with as high accuracy as possible. The main contribution is the use of fuzzy logic to build a model for determining the most important user-based features. The Random Forest model is also modified to improve the predictions.

1.1. Outline of paper

The related works are explained in Section 2. In Section 3, a review is presented of the theoretical background. The methodology of this work is illustrated in Section 4. Section 5 shows the results and discussion. Finally, Section 6 states the conclusions of this work.

2. Related work

Since fake news is one of the main issues faced at present, many researchers have contributed to developing models based on machine learning to detect this kind of news. However, through the review of previous works, it has been found that there is no sufficient number of researches conducted to identify Arabic fake news in particular. The reason for this can be traced back to several challenges and problems related to the Arabic language itself, such as 1) there is no big dataset that can be adopted to build prediction models; 2) most writers and people speak slang language, and 3) there is a lack of resources and not all libraries support the Arabic language.

In [1], the authors utilized Machine Learning (ML) to identify misinformation or unreliable tweets in Arabic based on a supervised classification model. This study is sufficient in that it included both the characteristics of the tweet (text features) and the characteristics of the user (user features), but the accuracy of the models used was not at a sufficient level. As for the work in [4], the authors introduced a technique to automatically generate Arabic manipulated news articles. They rely on real articles only to build a dataset consisting of true information and misinformation. Finally, ML algorithms have been applied to identify articles that have been manipulated using their method that generates false articles. In [5], the AraNews dataset was used by the authors to train the models. In their study, the TF-IDF technique was used to extract features. Next, three machine learning techniques were applied to predict fake news: Random Forest Classifier, Naive Bayes, and Logistic Regression. The results showed that the accuracy of the random forest model achieved the best accuracy. The authors in [7] utilized natural language processing and machine learning techniques to identify Arabic fake news. In this study, the dataset contains 1862 tweets which are collected from the Twitter platform. They relied on text features, user features, and tweets to make the prediction. The strong point is that the study was comprehensive in terms of its adoption of user and text features, but the accuracy of the results was not promising. Finally in [9], the textual content was analyzed through natural language processing. The study focused on the use of traditional methods of feature extraction and machine learning models on an English dataset. The research methodology and results were promising, however it is generally stated that the analysis of text written in English is less challenging and difficult than text written in Arabic.

3. Theoretical background

3.1. Dataset

Tweets are posts or messages that are published by individuals on the Twitter platform to exchange information with each other all over the world [10]. Usually, a set of markups are attached with each tweet that adds additional understanding and can thus be used by researchers. The most important of these are the hashtags (#) that add some basic phrases to the tweet, and the mention (@) followed by the user's name, which indicates that the tweet is a response to the user [1].

Twitter API enables users to access and download tweets, as well as download a set of user-specific features (the writer of the tweet) and other features of the tweet itself [5]. In this work, the dataset used in [1] has been expanded from 1862 tweets to 3000 tweets with the same attributes that the researcher used, as illustrated in Tables 1 and 2. After that, the tweets are labelled manually with the help of

Table 1. Content-based attributes.

has mention	mentions count	URL shorten
retweets num	is retweet	is reply
retweeted	day of week	# words
length chars	has URL	URLs count
has hashtag	hashtags count	# unique words
# unique chars	Has-?	has-? #
# symbols	has pos sent	has neg sent
pos score	neg score	

Table 2. User-based attributes.

avg retweet	followers count	tweet time spacing
avg hashtags	avg URLs	followers to friends
focused topic1	default image	has desc
listed count	status count	retweet fraction
desc len	username len.	avg tweet len.
friends to follower	is verified	registration diff.

experts. The number of not-fake and fake articles in these datasets is summarized in Fig. (1).

3.2. Text classification

Text classification, also known as text categorization, is the process or the task that classifies documents, articles, or tweets into predefined categories (target classes) [11]. Text content analysis is performed through the concept of natural language processing to extract features, after which the text is classified [12]. Fig. 2 shows the process of classifying any type of text, whether it is an article, a tweet, or a story [13].

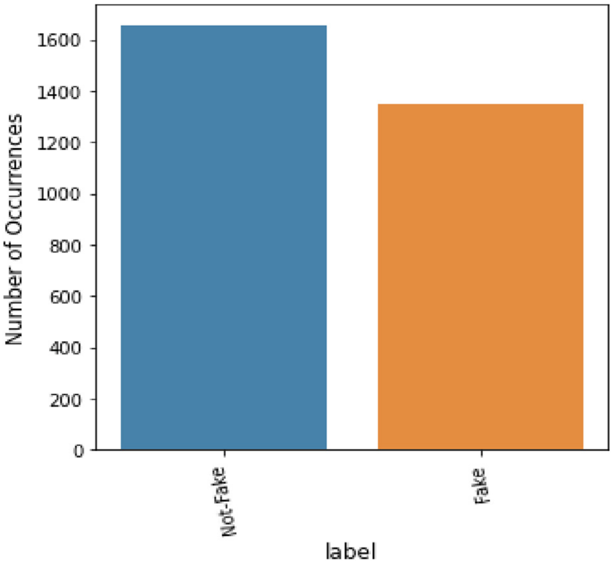


Fig. 1. Description of dataset.

3.3. Feature extraction

Feature extraction is the task of converting a particular text (or word) into feature vectors [14]. In other words, feature extraction methods are used to extract features because text data is not accepted or comprehended by machine learning techniques [15].

One of the most common techniques is "TF-IDF". This technique takes into account the number of times the word (token) appears during all articles or tweets in the dataset [16]. At first, the Term Frequency "TF" for each word is computed, after which the Inverse Document Frequency is calculated. Once the TF and IDF values are obtained, it is possible to calculate the score of TF-IDF [17].

3.4. Feature selection

The main objective of feature selection is to select or determine the best features (a relevant subset of features) among all the features [18]. This in turn reduces time complexity (reducing dimensionality) and improves classifiers by excluding irrelevant attributes [19]. In general, there are three common approaches for feature selection:

3.4.1. Filter methods

These methods identify attributes based on the importance of features for the target class, and they are always applied before the prediction algorithm, as shown in Fig. 3 [20]. There are many techniques under the concept of filter methods such as the correlation method, the Gain ratio method, the Relief method, and others. All of these methods assign a rank to each feature based on its relevance to the target class [21].

3.4.2. Embedded methods

These methods are a special type of feature selection as they are implemented during the prediction phase [22]. Embedded methods are not implemented in all machine learning algorithms.

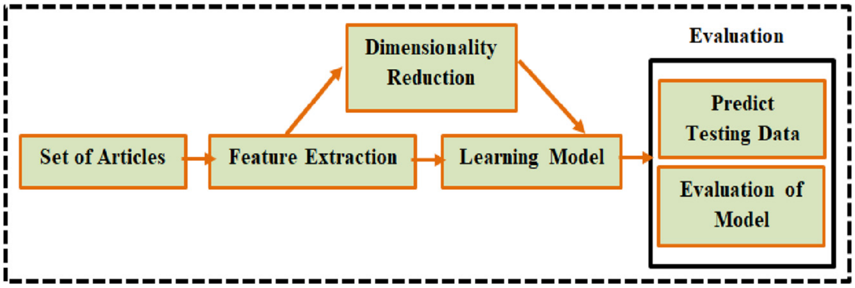


Fig. 2. A text Classification framework.

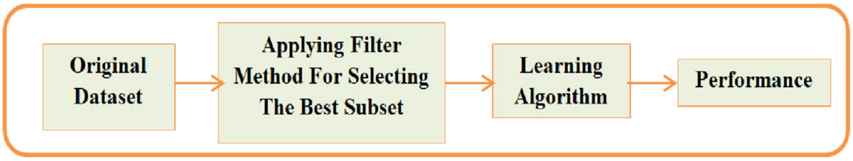


Fig. 3. Filter method framework.

They are only implemented by decision trees, where the nodes or attributes are pruned during the training process [23]. Fig. 4 explains the framework of embedded methods:

3.4.3. *Wrapper methods*

These methods differ from the two methods mentioned above, as they identify the best subset of attributes by their performance in the prediction algorithm [24]. Wrapper methods evaluate each subset of features, and the best subset is created based on a search strategy [25]. Fig. 5 explains the framework of wrapper methods:

4. Methodology

The proposed model illustrated in Fig. 6 can be explained as follows:

1 For tweets content, data preprocessing is performed to remove the noise.
2 For user features and content features, a fuzzy model is suggested for identifying relevant user features and removing irrelevant features.
3 The TF-IDF technique is used to extract the features (vector for each word) from the text and then identify the features that have a higher weight.
4 The outputs from steps 1 and 2 are concatenated.
5 Finally, the random forest algorithm is improved by controlling the selection of features in each tree.

4.1. *Data preprocessing*

One of the challenges that are faced is the noisy nature of Twitter content. Therefore, tweets that are downloaded from the Twitter API are preprocessed and then labeled manually. The labeling process was conducted by an expert. This process is verified by looking for the correctness or incorrectness of most of the tweets, so the labeling process was accurate. Regarding the content of the tweets or texts, a set of steps were performed to clean the text from the noise: Removing symbols, removing punctuations, removing numbers, removing non-Arabic letters, removing URL, removing any tweet with less than 5 words, removing stop words, and Arabic stemming.

Finally, the TF-IDF technique is applied to convert each word into a vector (feature). Depending on the score assigned to each word (feature) by TF-IDF, a total of 2000 features were selected as the most important ones, and the remainder of the features were excluded.

4.2. *Fuzzy model*

In order to identify the relevant user features and text features, a fuzzy model is proposed in this work. This model depends on the feature selection techniques: filter, embedded, and wrapper. For filtering techniques, the Correlation Method is used as it gives a weight to each attribute based on the correlation of this attribute with the target class. In the embedding methods, the Mean Decrease in Gini-index method is

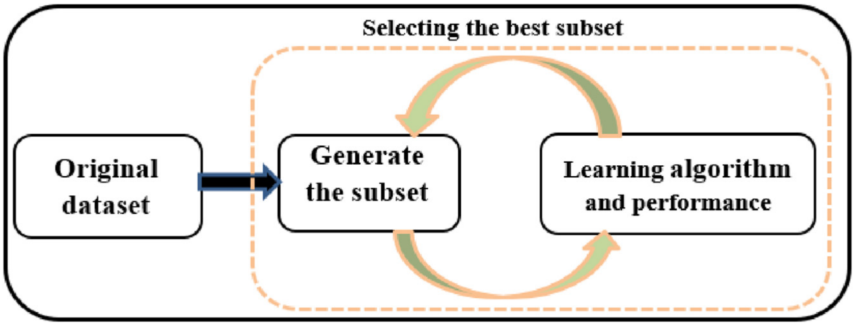


Fig. 4. Embedded method framework.

proposed for evaluating features based on the random forest algorithm. This method monitors the feature in each tree and calculates its contribution in reducing of Gini-Index criteria. It then calculates the average for the contributions of the feature in all trees as a weight of this feature. In addition, the Drop-Feature Importance is suggested to represent the wrapper approach. This method evaluates features based on their contribution to increased accuracy. In this method, first, the accuracy of the model is calculated for all attributes. Then the first feature is removed and a re-evaluation of the model is performed. The difference between accuracy with all attributes and accuracy after removing the first attribute is the weight of that attribute. The same goes for all the other features.

Each of the above methods removes features that have zero importance to the target class. Therefore, a Majority Voting Method is conducted in this work. Through this method, any attribute that is omitted by two of the techniques is discarded. In other words, a feature is relied upon in the next stage if it is important in the view of two or more feature selection methods. This method reduced the features from (46) to (33).

The fuzzy logic is deployed in this work because each of these (33) features has three different weights from the feature selection methods described above: Correlation Method, Mean Decrease in Gini-Index, and Drop-Feature Importance. In fuzzification, the Membership Function is calculated for each value (each weight) through the Triangle Membership Function, as shown in Fig. 7.

The Triangle Membership Function is calculated according to Equation (1) [26].

$$MF\,(x; a, b) = \frac{x - a}{b - a} \tag{1}$$

where: MF is Membership Function, x is the Value (weight), a is the minimum value, and b is the maximum value.

In defuzzification, the Center of Gravity method (COG) has been applied to assign one weight for each feature. Through this method, the final rank is obtained by relying on three weights and three values of the Membership Function for each feature, as in Equation (2) [27].

$$\mu_0 = \frac{\sum_{i=1}^{n} \mu(x_i) * x_i}{\sum_{i=1}^{n} \mu(x_i)} \tag{2}$$

Finally, the relevant attributes are selected according to a predefined threshold. The number of features is further reduced from 33 to 10, whereby any attribute whose importance or weight is less than 0.25 is removed. The top 10 features are listed in Table 3.

All of the above steps and procedures can be summarized through Algorithm 1:

```
Algorithm 1: Fuzzy Model
Input: Matrix content three ranks for each
Output: One rank for each feature
Begin
Apply Correlation Method, Mean Decrease of Gini-Index, and DFI
Apply Majority Voting Method
Save the remained feature with their three weights in the matrix: IMP[f]
// Compute MF for weights (Fuzzification)
for k = 1 to n                // n=33 (it is number of features)
  m-f is MF, R[j] is the array of ranks, and M-F[j] is the array of Membership Function
  // m = 3 (three ranks of each feature)
  for j = 1 to m
      $m - f = \frac{X_i - a}{b - a}$
      M-F[j]= m-f
  end-for
end-for
//Compute the COG Method (Defuzzification)
    for k = 1 to n
      Calculate COG  // Equation 1
        if COG > 0.25
          S-F[k] = COG
        end-if
    end-for
Return S-F[k]
End
```

4.3. Improved random forest

Random Forest is an ensemble algorithm according to the decision tree technique, whereby the features in the original dataset are randomly selected in each sample to build a decision tree model [16]. In this study, both user- and tweet-specific features have been used which were eventually reduced through the fuzzy model to 10 features. In addition, text-specific features have been used that represent unique words generated by the TF-IDF technique. Since these features are very large, max-features are determined in the TF-IDF model to be 2000. Next, the random forest algorithm

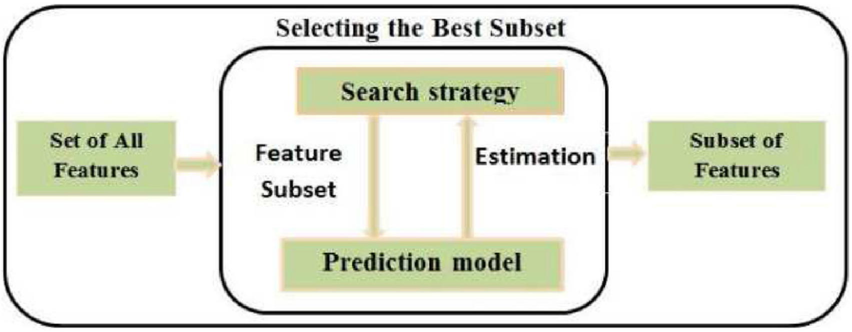


Fig. 5. Wrapper method framework.

is used to control the selection process for features in each sample, which implies that the selection process is not random. The number of trees in the random forest model is set to 100, and for each tree, 20 features are selected. Finally, the algorithm chooses all the user's features and text features (the 10 features selected by the fuzzy model) in each tree and randomly chooses the other 10 features resulting from the TF-IDF technique. The Improved Random Forest is summarized in Algorithm 2:

Algorithm 2: Improved Random Forest

```
Input: Array represents user features and TF-IDF features.
Output: Accuracy model.
Begin
$n_{tree}$ = 100 and it is a number of trees. $m_{try}$ = 20 and it is a number of features. Set the weight of user features based on the fuzzy model. Set TF-IDF features to zero to be selected at random.
Set samble_1, sample_2, and bootstrap to null
 for tr=1 to $n_{tree}$
     for fe=1 to $m_{try}$
         if (rank of feature(F) >Ɵ)
             sample_1 = F
          end_if
         else
     smple_2 = select feature at random with replacement
                 end_else
          end_for
      bootstrap = merge sample_1 with sample_2
      build regression tree on (bootstrap)
      compute accuracy model
return accuracy
End
```

5. Results and discussion

This work aims to improve the predictions of fake news by identifying important features of the writer as well as analyzing the text itself. Initially, user features and text features are evaluated in three ways: Correlation Method, Mean Decrease of Gini Index, and Drop-Feature Importance. Each method refines an attribute that has zero significance. Because the outputs are different for each method, the Majority Voting Method has been applied, which reduced the features from (46) to (33) features. These features have been reduced to 10 only through the fuzzy model, as shown in Table 3.

In the process of analyzing the tweet, data-preprocessing is performed to remove noise from the text. Then the TF-IDF method is used to convert the words into vector features. The features resulting from this method represent the unique words in the dataset. Because the number of features generated is very large and some of them are considered less important, there are 2000 features taken into account which have a high score. Finally, the user-specific features are combined with the features generated by the TF-IDF method into a single dataset.

In order to build a reliable model on this dataset, the random forest model is improved by controlling the process of selecting features in each tree instead of random selection. A random selection is made for TF-IDF features only, meanwhile the 10 user features were required to be present in all trees. In other words, each tree of the random forest model is constructed from 10 user-specific features and 10 randomly selected features from the outputs of the TF-IDF technique.

The quality of the proposed model was evaluated using accuracy measures. The results showed that the accuracy of the Improved Random Forest model is better as compared with the work presented in [1], which used the same dataset. Then, the dataset used

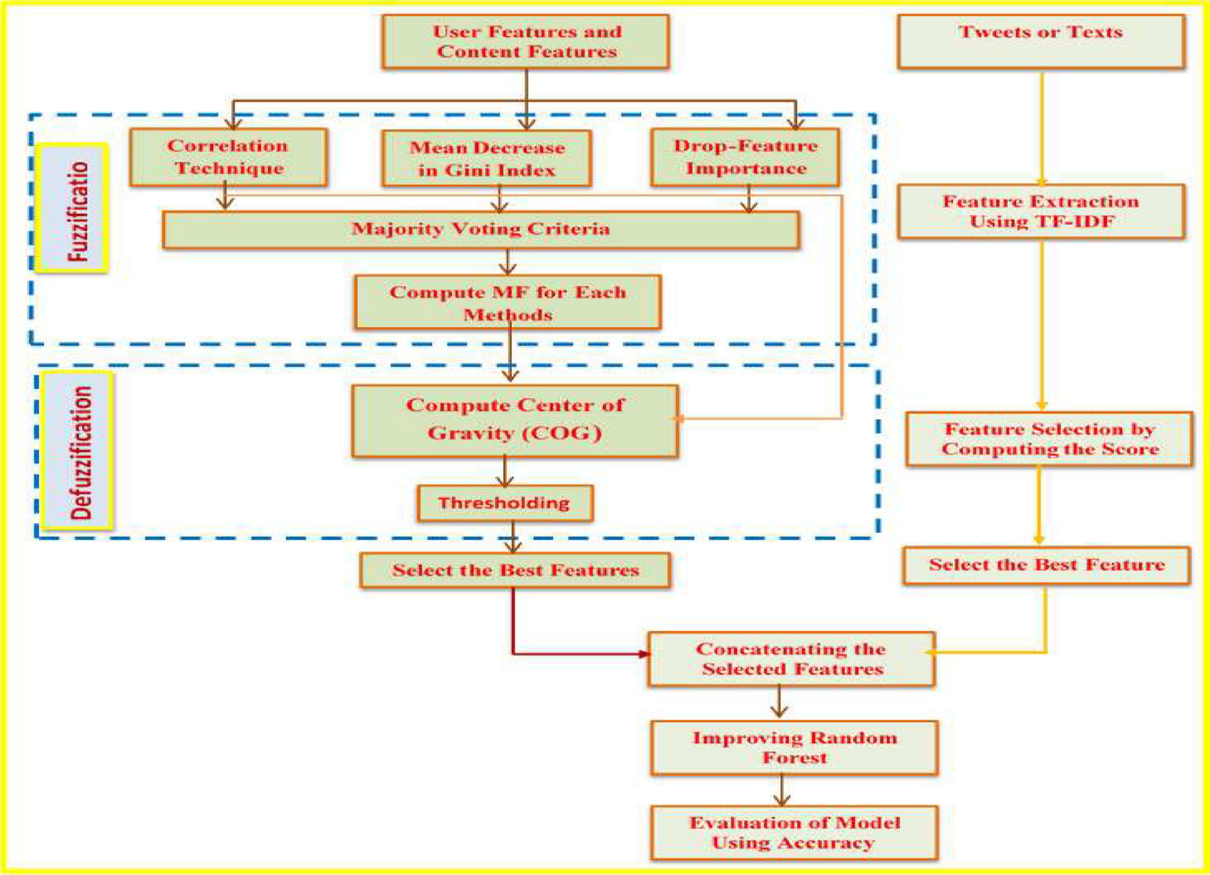

User Features and Content Features
Tweets or Texts
Fuzzificatio
Correlation Technique
Mean Decrease in Gini Index
Drop-Feature Importance
Majority Voting Criteria
Compute MF for Each Methods
Feature Extraction Using TF-IDF
Defuzzification
Compute Center of Gravity (COG)
Thresholding
Feature Selection by Computing the Score
Select the Best Features
Select the Best Feature
Concatenating the Selected Features
Improving Random Forest
Evaluation of Model Using Accuracy

Fig. 6. Proposed methodology.

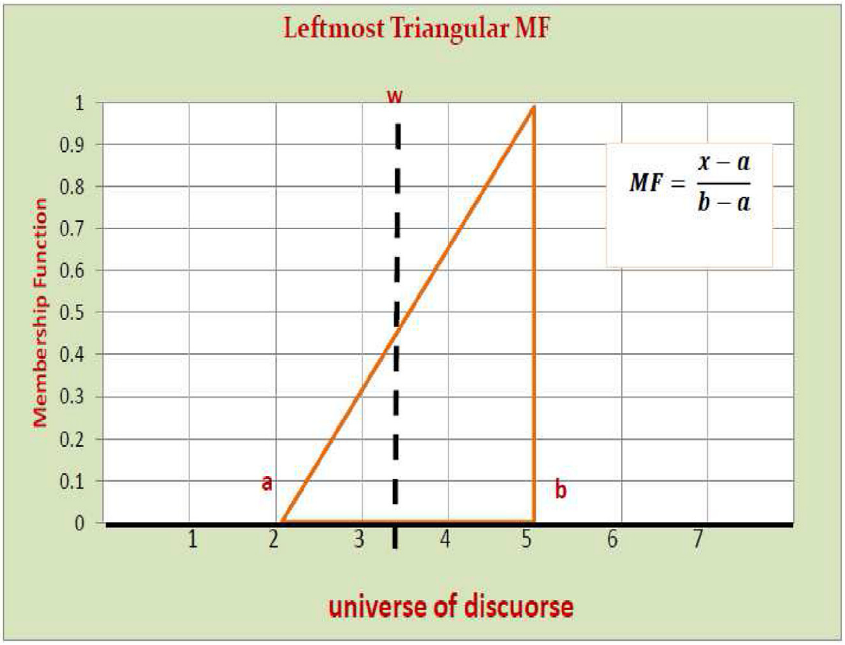


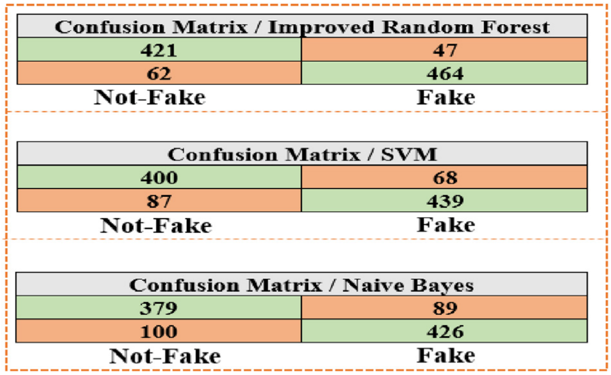
Fig. 7. Triangle membership function.

Table 3. Feature and the rank.

The Best Features	COG (the rank)
Has user mention	0.391
Count of retweets	0.339
Has hashtags	0.329
Count of hashtags	0.326
Count of unique words	0.323
Has hashtag	0.320
Has URL	0.277
Count of friends	0.273
Retweeted	0.266
Has default image	0.258
Registration age	0.256

Table 4. The results of a models.

The Model	The Accuracy
Improved Random Forest Model	0.895
SVM Model	0.848
Naïve Bayesian Model	0.809



Confusion Matrix / Improved Random Forest	
421	47
62	464
Not-Fake	Fake

Confusion Matrix / SVM	
400	68
87	439
Not-Fake	Fake

Confusion Matrix / Naive Bayes	
379	89
100	426
Not-Fake	Fake

Fig. 8. Confusion matrixes of models.

in [1] is expanded in this study by adding about 1000 other tweets and manually labeling them by an expert. The results of the proposed method are compared to two other methods of machine learning, namely the Naive Bayesian theorem and SVM. The results listed in Table 4 show that this proposed method outperformed the other methods in terms of accuracy.

In [1], the accuracy of Random Forest is 0.76 whereas, in the proposed model the accuracy is 0.895.

Finally, the confusion matrix for Improved Random Forest, SVM, and Naive Bayesian are shown in Fig. 8.

6. Conclusions

Twitter is one of the leading social networks in the process of spreading news, whereby Twitter users tend to spread false news for different political and financial purposes. Given the serious problems that fake news causes to societies, it is necessary to establish reliable models that are based on the principle of machine learning to discover this type of news. In order to improve prediction accuracy, different aspects of the tweet are taken into account, including the analysis of the text itself as well as the features of the writer of this text. In this study, the main focus is on identifying important user features and text features by building a fuzzy model. Next, the random forest algorithm was improved by controlling the process of selecting features in each tree. The results showed that the model proposed in this work is better than the other models in terms of prediction accuracy. There are some challenges faced that have a significant impact on the prediction models, which is the lack of a large and comprehensive dataset for Arabic news. Therefore, it is aimed to contribute to the creation of a large dataset in the near future. It is also aimed that this study is complementary to the rest of the research in providing a contribution to solving the problem of fake news.

Conflict of interest

There is no conflict of interest.

References

[1] G. Jardaneh, H. Abdelhaq, M. Buzz, D. Johnson, Classifying Arabic tweets based on credibility using content and user features, in: IEEE Jordan international joint conference on electrical engineering and information technology vol. 1, 2019: pp. 596–601, https://doi.org/10.1109/jeeit.2019.8717386.

[2] H. Bicen, R. Haidov, A content analysis on articles using Twitter in education, Postmod Openings 12 (2021) 19–34, https://doi.org/10.18662/po/12.1Sup1/269.

[3] E. D'Andrea, P. Ducange, A. Bechini, A. Renda, F. Marcelloni, Monitoring the public opinion about the vaccination topic from tweets analysis, Expert Syst Appl. 116 (2019) 209–226, https://doi.org/10.1016/j.eswa.2018.09.009.

[4] E.B. Nagoudi, A. Elmadany, M. Abdul-Mageed, T. Alhindi, H. Cavusoglu, Machine generation and detection of Arabic manipulated and fake news, arXiv preprint arXiv:20110309 vol. 2, 2020: pp. 69–84. https://doi.org/10.48550/arXiv.2011.03092.

[5] F. Aljwari, W. Alkaberi, A. Alshutayri, E. Aldhahri, N. Aljojo, O. Abouola, Multi-scale machine learning prediction of the spread of Arabic online fake news, Postmod Openings 13 (2022) 1–14, https://doi.org/10.18662/po/13.1Sup1/411.

[6] J. Kim, B. Tabibian, A. Oh, B. Schölkopf, M. Gomez-Rodriguez, Leveraging the crowd to detect and reduce the spread of fake news and misinformation, in: Proceedings of the eleventh ACM international conference on web search and data mining vol. 4, 2018: pp. 324–332, https://doi.org/10.1145/3159652.3159734.

[7] T. Thaher, M. Saheb, H. Turabieh, H. Chantar, Intelligent detection of false information in Arabic tweets utilizing hybrid harris hawks based feature selection and machine learning models, Symmetry 13 (2021) 1–24, https://doi.org/10.3390/sym13040556.

[8] M. Al-Yahya, H. Al-Khalifa, H. Al-Baity, D. AlSaeed, A. Essam, Arabic fake news detection, comparative study of neural networks and transformer-based approaches, Complexity 2021 (2021) 1—10, https://doi.org/10.1155/2021/5516945.

[9] Z. Khanam, B. Alwasel, H. Sirafi, M. Rashid, Fake news detection using machine learning approaches, Mater Sci Eng. 1099 (2021) 1—13, https://doi.org/10.1088/1757-899x/1099/1/012040.

[10] Z.T. Osakwe, I. Ikhapoh, B.K. Arora, O.M. Bubu, Identifying public concerns and reactions during the COVID-19 pandemic on Twitter: a text-mining analysis, Health Nurs. 38 (2021) 145—151, https://doi.org/10.1111/phn.12843.

[11] X. Luo, Efficient English text classification using selected machine learning techniques, Alex Eng J. 60 (2021) 3401—3409, https://doi.org/10.1016/j.aej.2021.02.009.

[12] A. Kadhim, Survey on supervised machine learning techniques for automatic text classification, Artif Intell Rev. 52 (2019) 273—292, https://doi .org/10.1007/s1 0462-018-09677-1.

[13] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, D. Brown, Text classification algorithms: a survey, Information 10 (2019) 1—68, https://doi.org/10.3390/info10040150.

[14] M.G. Huddar, S.S. Sannakki, V.S. Rajpurohit, Attention-based word-level contextual feature extraction and cross-modality fusion for sentiment analysis and emotion classification, Int J Intell Eng Inf. 8 (2020) 1—18, https://doi.org/10.1504/ijiei.2020.10543 0.

[15] D. Wu, R. Yang, C. Shen, Sentiment word co-occurrence and knowledge pair feature extraction based LDA short text clustering algorithm, Intell Inform Syst. 56 (2021) 1—23, https://doi.org/10.1007/s10844-020-00597-7.

[16] X. Chen, Y. Xue, H. Zhao, X. Lu, X. Hu, Z. Ma, A novel feature extraction methodology for sentiment analysis of product reviews, Neural Comput Appl. 31 (2019) 6625—6642, https://doi.org/10.1007/s00521-018-3477-2.

[17] R. Ahuja, A. Chug, S. Kohli, S. Gupta, P. Ahuja, The impact of features extraction on the sentiment analysis, Procedia Comput Sci. 152 (2018) 341—348, https://doi.org/10.1016/j.procs.2019.05.008.

[18] D. Jain, V. Singh, Feature selection and classification systems for chronic disease prediction: a review, Egypt Inform J. 19 (2018) 179—189, https://doi.org/10.1016/j.eij.2018.03.002.

[19] B. Remeseiro, V. Bolon-Canedo, A review of feature selection methods in medical applications, Comput Biol Med. 112 (2019) 1—35, https://doi.org/10.1016/j.compbiomed.2019.103375.

[20] V. Galiano, J. Luque-Espinar, M. Chica-Olmo, M.P. Mendes, Feature selection approaches for predictive modelling of groundwater nitrate pollution: an evaluation of filters, embedded and wrapper methods, Sci Total Environ. 624 (2018) 661—672, https://doi.org/10.1016/j.scitotenv.2017.12.152.

[21] P. Saqib, U. Qamar, A. Aslam, A. Ahmad, Hybrid of filters and genetic algorithm-random forests based wrapper approach for feature selection and prediction, Intell Comput Proc. 998 (2019) 190—199, https://doi.org/10.1007/978-3-030-22868-215.

[22] B. Venkatesh, J. Anuradha, A review of feature selection and its methods, Cybern Inf Technol. 19 (2019) 3—26, https://doi.org/10.2478/cait-2019-0001.

[23] M. Lu, Embedded feature selection accounting for unknown data heterogeneity, Expert Syst Appl. 119 (2019) 350—361, https://doi.org/10.1016/j.eswa.2018.11.00 6.

[24] G. Ansari, T. Ahmad, M.N. Doja, Hybrid Filter—Wrapper feature selection method for sentiment classification, Arabian J Sci Eng. 44 (2019) 9191—9208, https://doi.org/10.1007/s13369-019-040 64-6.

[25] M. Mafarja, S. Mirjalili, Whale optimization approaches for wrapper feature selection, Appl Soft Comput. 62 (2018) 441—453, https://doi.org/10.1016/j.asoc.2017.11.006.

[26] M.K. Mandal, A.P. Burnwal, B. Mahatha, A. Kumar, J. Ghosh, Fuzzy rule-based system for route selection in WSN using quadratic programming, in: Architectural wireless networks solutions and security, Springer, 2021: pp. 81—98, https://doi.org/10.18662/po/12.1Sup1/269.

[27] M.K. Mandal, B. Mahatha, A.P. Burnwal, S.K. Das, A. Sharma, Fuzzy-based optimal solution for minimization of loss of company based on uncertain environment, in: Nature-inspired computing for smart application design, Springer, 2021: pp. 71—83, https://doi.org/10.1007/978-981-33-6195-9_5.