Journal of Theoretical and Applied Information Technology

30th September 2023. Vol.101. No 18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN ENSEMBLE STACKING MODEL FOR RUMOR DETECTION BASED ON ARABIC TWEETS

THANAA MOHAMED HASSAN[1] YEHIA MOSTAFA HELMY2 DOAA S. ELZANFALY[3]

Lecturer, Faculty of Commerce & Business Administration
Helwan University, Egypt
Proffesor, Faculty of Commerce & Business Administration
Helwan University, Egypt
Professor, Faculty of Computers & Artificial Intelligence – Helwan University, Egypt

thanaahassan@hotmail.com, ymhelmy@commerce.helwan.edu.eg, doaa.saad@fci.helwan.edu.eg

ABSTRACT

Effective rumor detection within Arabic-language social networks is crucial for mitigating misinformation's impact on users. Despite extensive research on rumor detection in English, investigations into Arabic rumors are limited, despite Arabic's significance for 25 nations. This study proposes a unique hybrid ensemble model for detecting Arabic rumor tweets, addressing both Ensemble stacking-based Machine Learning and Deep Learning dimensions. The initial phase of the proposed model involves the strategic implementation of Machine Learning stacking, wherein standalone classifiers, including the Decision Tree, Random Forest, and Gaussian Naive Bayesian models, are thoughtfully employed. Impressively, the Random Forest and Gaussian Naive Bayesian models exhibit commendable accuracies, both attaining 86%, while the Decision Tree model registers a respectable accuracy of 84%. To further amplify accuracy, a subsequent stage incorporates logistic regression, culminating in an overall accuracy of 87%. In the pursuit of advancing the bounds of accuracy and performance, the study extends its exploration to the realm of deep learning stacking. This facet is manifested through the construction of four distinct neural network models, the collective accuracy of which varies between the noteworthy ranges of 77% to 89%. By judiciously integrating these neural network models with logistic regression in the ensuing stage, the accuracy remarkably ascends to an impressive 90%. This enhancement, amounting to a 3% increase over the machine learning stacking approach, substantially augments vital performance metrics encompassing precision, accuracy, recall, and F1-Score, thus significantly refining the landscape of rumor detection.

Keywords: *Rumor Detection; Online Social Media Networks; Machine Learning; Deep Learning; Ensemble Stacking Model*

1. INTRODUCTION

In this era of big data, where people's ability to acquire and exchange information has increased rapidly, the use of social media networks has developed into an integral part of our daily lives. Users on a variety of platforms can submit information in real-time, pass information to other users, and comment on any material. Therefore, microblogging systems like Sina Weibo and Twitter often provide more freedom, stronger interaction, and even complete dispersal of information. As data volumes grow exponentially, misinformation and rumors eventually follow [1].
Rapid information diffusion and the ever-evolving nature of social networks make rumor detection an exciting and difficult task. Early rumor detection refers to the process of identifying and dispelling rumors before they spread widely. [2]. In recent years, numerous methods for rumor detection have been proposed, often formulated as a binary classification (rumor or non-rumor) problem, that employs machine learning or deep neural networks [3]. These approaches simplify feature extraction and enable effective abstract representation learning. In essence, efficient and flexible models are required to identify trending rumors, capturing long-range dependencies among posts, and generating distinct representations to facilitate accurate early detection [4].
The idea of automatically spotting rumors on social media is currently attracting a lot of attention. To control rumors and lessen their negative impact on society. Research on the detection of rumors in public networks has made extensive use of machine learning, deep learning, and ensemble combined

models. The use of machine learning-based approaches (ML) has emerged as a potentially useful method for identifying online rumors. Support Vector Machine (SVM), Naive Bayes (NB), Decision Trees (DTs), and Logistic Regression (LR) are some examples of the supervised learning algorithms used in the algorithm-based approach to rumor identification. Traditional machine learning is currently the primary foundation for rumor identification. Typically, it relies on three main factors: the content of the rumor text, the communication structure, and the credibility to construct features manually [5]. Applying of different machine learning algorithms and techniques was shown in the following studies [6], [7], [8], [9].

The Deep learning (DL) paradigm is another way of automatically learning and fusing different types of information. The DL paradigm, like the ML paradigm, is founded on learning from data; however, unlike the ML, which relies on a set of carefully built features, no feature engineering is necessary for DL because the classifier learns and receives the features it needs during the training phase [10]. The DL techniques have various benefits, such as a notable performance boost and the removal of the time-consuming feature extraction process [11].There are lots of studies that have used the DL techniques as in [12]- [13].

Although prior work provides an enhanced method for rumor identification, it has been explained that an ensemble approach is preferable for statistical, computational, and representative reasons [14]. In addition, the application of the ensemble solution can reduce excessive variability, variation, and bias, as shown by empirical research in multiple prior works [15]. To take advantage of these benefits, various methods have been developed to employ ensemble solutions for rumor detection. [16].

Many of the research were successful in identifying a fake post, yet there are many additional aspects that might be examined to further enhance the models' performance, particularly when dealing with the Arabic language. Twenty-five countries have declared Arabic to be their official language. [17]. Working with Arabic data can be more challenging due to the language's higher number of morphological patterns and grammatical rules compared to English, reducing the efficiency of current natural language processing (NLP) technology. The transmission of false information through Arabic-language media is a growing and important concern. As a result, state-of-the-art methods are necessary to detect and remove fraudulent content from social media. [18].For the sake of identifying false information on Twitter [19], more research is required to create strategies for identifying rumors in Arabic.

In this research, we initially address the issue of parsing Arabic tweets and subsequently employ various standalone machine learning and deep learning algorithms to identify rumors. Furthermore, we assess effectiveness concerning the stacking ensemble technique by utilizing both machine learning in addition to deep learning models. Hence, this study aims to investigate the impact of the stacking model on machine and deep learning models, identify the most effective machine learning/deep learning model for detecting rumors in Arabic, and to share valuable insights from processing and working with Arabic tweets.

The remainder of the paper is structured as follows: In Section 2, we investigate the most related research work. Section 3 presents the proposed models along with the dataset and the preprocessing stage. Finally, the experimental results and the subsequent discussion summarizing the most significant findings are presented in Section 4.

2. RELATED STUDIES

The field of rumor detection has produced a variety of academic articles. Here, we group similar studies together into three groups defined by the kind of models they employ. The first group includes experts in machine learning, while the second is dedicated on deep learning. The final section discusses ensemble combined models.

a. RUMOR DETECTION USING MACHINE LEARNING APPROACHES

A huge number of papers have been studied on the topic of rumor detection using supervised machine learning techniques. It has been found that most of these papers are evaluating more than one classifier in order to decide which one gives the best results, as in [6], [8], and [9]. The Naïve Bayes technique shows great and better efficient results in more than one study, as in [7], [8], and [9], with respect to the other classifiers. Notably, most current research seek to enhance the early rumor detection process by doing the following, in addition to employing traditional classifiers: In [6], Dito et al. created an automated system that uses a number of medical keywords that are updated automatically by using the Wikipedia API to feed real-time Twitter data relevant to the health field. Wenfeng et al. in [7] applied a rumor detection model relying on Naive Bayes and NLP to detect micro-blog rumors in a large volume of data. Using

supervised machine learning classifiers, A. Habib et al. in [8] created an automated system for detecting business rumors in online business reviews. While Dubey et al. in [9], they worked in this study on algorithms like Multinomial Naive Bayes, Gradient Boosting and Random Forest with datasets imported from Kaggle to implement them and get closer to more precise results of a rumor.

b. RUMOR DETECTION USING DEEP LEARNING APPROACHES

DL has been increasingly important in recent years for rumor detection. Concerning the age of online social networks, the categorization work of rumors was already being carried out using various DL techniques. DL has become one of the most important technologies for developing effective systems that can detect and categorize rumors.

Both Cheng et al. [12] and Wang et al. [20] have trained their models in different ways by means of the pre-trained BERT model. According to the work of Cheng et al. in [21], deep neural networks can be used to reliably categorize whether or not rumors about COVID-19 are true (DNN). To extract meaningful features from textual content vectors, the authors initially employed an LSTM-based variational autoencoder (VAE) [12] and then the pre-trained BERT model. These structures are fed through a DNN classifier, which then returns a classification outcome. The average F1-score for the DNN classifier used in this investigation to determine whether or not rumors are true was 85.98%. While [20] proposed an enhanced technique using the BERT pre-training model and a CNN model constructed from FastText, Word2vec, and GloVe. Even though the models used in said research showed an F1 score of 73% on a dataset containing rumors, the method used to generate this score, extracts more semantic information by extracting BERT hidden state output.

Speaking of Recursive Neural Network model, both Shuaipu [22] and Yang et al. [23] focus on classifying COVID-19 rumors and fake news using the TextRNN model. A number of LSTM layers are integrated into the structure of the TextRNN model. In [22] , TextRNN was able to achieve higher accuracy of 98.40% in the classification results because it effectively captured the connection among the semantics in addition to contexts of the texts. Using bottom-up and top-down tree-structured neural networks, MA et al. propose two recursive models [24] for learning and classifying rumor representations that are a natural fit for the tweet propagation architecture and reach an accuracy of 72% on the twitter dataset.

Convolutional Neural Network (CNN) was also mentioned in different studies by Bian et al. [25] and Asghar et al. [26]. In [19], they created a model for graphs that can move in both directions, calling it a "Bi-Directional Graph Convolutional Network" (Bi-GCN). The model operates on both the top-down and bottom-up spread of rumors, with an 88% accuracy rate on the Twitter dataset, to investigate both characteristics. In [26], the authors propose using a convolutional neural network trained with bidirectional long-short term memory to accurately classify tweets as either rumors or facts, with an accuracy rate of 86.1%.

Neural Network models was also mentioned by Shams et al. [27] and Kar et al. [13] producing high efficiency in the early rumor detection. In [27] when it comes to public health misinformation, specifically about COVID-19, they've built an online Search Engine Misinformation (SEMiNExt) that will assess the efficacy of an ML-based method integrated with a search engine extension. Based on the data, SEMiNExt under ANN performs the best. It achieves an accuracy of 93%. In [13] to categories COVID-19 fictitious tweets composed in Indic languages other than English, such Hindi and Bengali. The authors employed a NN model with multilingual BERT (mBERT) embedding and a multi-layer perceptron (MLP) model with pre-trained BERT embedding. They were unable to get good results from their MLP model since their dataset was too small. Despite the NN model's capacity to deal with the decreased sample size problem, it was still able to achieve F1-scores of 80% or above in both monolingual (for English) and multilingual (for English, Hindi, and Bengali) scenarios.

c. RUMOR DETECTION USING ENSEMBLE COMBINED MODELS

A number of studies have improved classification performance by combining conventional ML and DL methods in novel ways. The authors of both [28] and [16] proposed a combined model of a convolutional neural network with other techniques. In [28], they performed two models for misinformation classification in COVID-19 tweets: a CNN RNN model (a CNN layer stacked on top of an RNN layer) and an RNN-CNN model (a single BiLSTM layer is used on top of a 1D-CNN layer), both of which achieve an F1 score above 70% with a word embedding in the first layer for classification purposes. While in [16], they introduced a hybrid deep learning (DL) model-based Arabic FN detection (AFND) system. Both traditional neural network and long short-term

memory (CNN-LSTM) techniques are incorporated into this model. Based on the findings, the planned CNN-LSTM achieves 81.6% accuracy.

Other attempts in [29], [30] and [31] have proposed different combinations of ensemble models in ML and DL. In ML, [29] proposed a framework for defending the truth of tweets that is based on ensemble learning and uses features from both the tweet and the user level. With the use of stacking-based ensemble learning, they applied six classic ML algorithms to improve accuracy. There were many experiments done to build the ensemble model. For (level 0), they employed SVM+RF models, and for r (level 1), they employed the C4.5 model as a meta-model. Several other combinations were utilized in the experiment, including C4.5+RF, C4.5+kNN, SVM+kNN, SVM+BN+kNN, and C4.5+BN+kNN.

For DL ensemble models, [30] planned a data pre-processing method in addition to ensemble model using several deep learning techniques for easier detection of rumors on social media. Our data pre-processing method transforms Twitter conversations into time-series vectors using the tweet creation timestamps achieving an F1 score of 64.3%.In order to increase the performance of rumor tracking, they [31] presented a deep RL-based ensemble model (RL-ERT) that combines many components using a weight-tuning policy network and takes advantage of unique social traits. They concluded their work with experimental results on public datasets that demonstrate RL-ERT's superiority in terms of efficiency and effectiveness.

According to the stated studies and the realm of misinformation mitigation within the context of Arabic-language social networks, the challenge of timely detecting rumors emerges as a pivotal concern. While considerable strides have been made in rumor detection within the English language, the exploration of Arabic rumors remains notably limited, despite the language's significance across 25 nations. This discrepancy highlights an unaddressed need for robust methods tailored to the intricacies of Arabic rumor propagation. As explained below this study aims as to bridge this gap by proposing a novel hybrid ensemble model for the early detection of Arabic rumor tweets. Through the integration of Ensemble stacking-based Deep Learning and Machine Learning dimensions, the research endeavors to elevate the accuracy and efficacy of rumor detection strategies in the Arabic linguistic landscape, thereby advancing the understanding and management of misinformation dissemination.

3. PROPOSED ENSEMBLE MODELS FOR RUMOR DETECTION

To build the model for rumor detection, all of the gathered ground-truth must be first prepared, and after that, it must go through feature engineering stages in addition to model construction outputting evaluation indicators.

a. DATASET SELECTION AND PREPROCESSING

Researchers, especially those working with Arabic, are left to create their own benchmark dataset. Arabic is difficult for a variety of reasons. It has a great deal of inflectional and derivational complexity and a very extensive morphological system. In addition, many Arabic words have many meanings and/or are interchangeable with each other. This ambiguity and uncertainty complicate of the Arabic topic models [32].

In this research, the ArCOV19-Rumors dataset [19] will be utilized. ArCOV19 being the first Arabic dataset for misinformation detection on Twitter from January 27, 2020, to April 30, 2020. The Qatar National Research Fund assisted in the collection of this dataset. This dataset consists of 138 verified claims, primarily from prominent fact-checking websites, and 9,400 tweets that are germane to those claims. Veracity manually annotated Tweets in support of study on misinformation detection, being one of the most significant challenges during the pandemic. This dataset also includes social, political, sports, entertainment, and religious categories that were affected by COVID-19.

- **Data Preprocessing**

During data preprocessing, the tweets were distributed in IDs, we used the Hydrator to convert those IDs into a csv file. The Hydrator is an electron-based desktop application for hydrating Twitter ID datasets. Twitter's Terms of Service didn't permit full JSON for datasets of tweets to be classified to third parties. However, they do permit datasets of tweet IDs to be shared.

Concerning the extracted texts, containing rumor in addition to non-rumor tweets, were then moved to the next stage, where several preprocessing techniques were applied, including some common cleaning steps and a variety of approaches, such as the removal of punctuation, emoji's, duplication, stop words, URLs, special symbols, and non-Arabic terms.

- **Data and Feature Engineering**

In this paper, 321092 total tweets were gathered and worked on. The tweets were classified as true

or false rumors. The proportion of records containing true rumors to those containing false rumors was roughly 23:77 as shown in Table 1.

- *Table 1: Dataset Statistics and Features* [19]

Features	Tweets Statistics
Data Frame	27 Jan 2020-April 2020
Language	Arabic
Dataset size	365500
% Missing data	12.15%
Obtained Data	321093
• Rumor	249216 (77%)
• Non-Rumor	71878 (23%)

The feature set that was used in this paper was derived from the feature sets that were proposed in [33] and [34]. The feature set is typically broken down into three distinct categories: content features, user features, and propagation features.
The length of a message, whether or not it was verified, and whether or not it contained potentially sensitive material were some of the characteristics of messages that were taken into consideration when working on the content feature. These elements were taken into account when working on this research. Regarding the user features, we did some work on the follower's section of the user interface, taking into consideration the characteristics of users who post messages. As part of the propagation feature, we investigated the properties associated with the propagation tree that may be constructed from a message's retweets. Among these are activities such as working on the counts of retweets as one of the crucial features for our early rumor detection task.

- **Feature Extraction**

In order to properly extract features, special preprocessing of the text data that will be used for rumor analysis and predictive modelling is required. Tokenization of a sample text is performed by first dividing a huge text into words, then each of these words is encoded as an integer or floating-point value using vectorization so that feature extraction can take place. This process is done in order to get text data ready for predictive modelling. The work being done now makes use of Python scikit-learn libraries called Count Vectorizer for converting text to word count vectors. Additionally, TF-IDF vectorizer is utilized in order to conduct an analysis of the word frequency and delete the majority of the words in order to illuminate the number of calculations requested.

$$TF - IDF\,(t,d) = TF(t,d) \times IDF(t,D)$$

[35]

Where, TF (t, d) is the frequency of 't' in'd' and IDF (t, D) is how 't' is common or rare across 'D

Two Models were used at this stage to analyze the tweets' texts, including ensemble-based machine learning and ensemble-based deep learning models that analyze attributes from both users and tweets.

b. MACHINE LEARNING ENSEMBLE STACKING MODEL

There was a rise in the use of ensemble learning methods in the sector of machine learning. These methods combine the greatest features discovered during model training from numerous models [36]. In comparison to bagging and boosting, the stacking approach is slightly unique. In contrast to bagging and boosting, it employs a third model ("meta learner") to combine the outputs of the individual "base" models. Secondly, stacking-based models are typically heterogeneous because they train distinct forms of algorithmically distinct base models. The meta-learner makes predictions about the final outcome using the results of the base models as input [36].

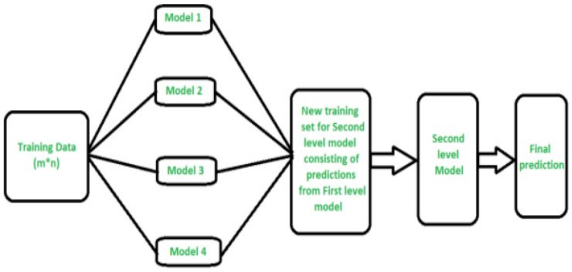
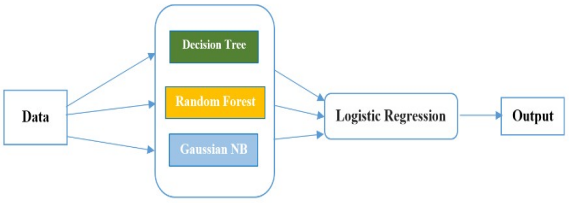


Figure 1 Stacking Technique Obtained from [37]

In this model, we started by loading the base learning algorithms (Decision Tree, Random Forest and Gaussian NB) in order to perform the cross validation and stacking recording the score at the end as shown in Figure 3.



- *Figure 2 Ensemble Stacking Machine Learning Model*

Decision Tree, Random Forest, and Gaussian NB were the three machine learning models employed. These models served as the foundational classifiers in later iterations of ensemble techniques.

These base classifiers are chosen for their performance and adaptability to high-dimensional data, DT one of its main benefits that it can easily interpret the predictions of target variables through learning feature data in addition to splitting the area to sub-areas [38], the Naïve Bayes technique expects the class of unseen data (testing data), as it uses the training data to predict the probability of features belonging to a specific class and one of its main advantage that it can work on any type of data and any size with extremely fast speed compared to more sophisticated methods [8] and in comparison to other models, the error rate in RF is less mainly backed by the lack of correlation between the trees [39] as it is considered as an ensemble approach that boosts accuracy by using various decision trees [40]. Table 2 displays the specific model configurations and hyper-parameter settings. The texts of tweets can be used as input to these models.

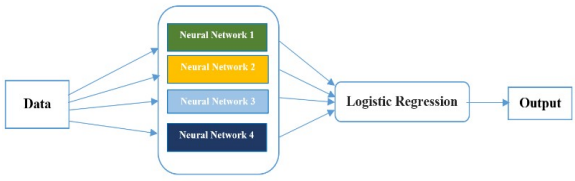
Table 2 Classifier Names Used and their Hyper-parameter

Classifier Name	Hyper-parameter Used
Decision Tree	The maximum depth of the tree = 2, random_state=0, criterion='entropy', splitter='random'
Random Forest	maximum_depth=2, random_state=0, number_estimators=2, min_samples_split=3, minimum_samples_leaf=3
Gaussian NB	variance_smoothing=1e-02

Concerning the meta-model is then trained on the predictions made by the base models on out-of-sample data. The Logistic Regression is used as one of the generalized linear models, as well as a classical classification method that has been proven its efficiency in solving the optimization problem with the likelihood function as the objective function [41].

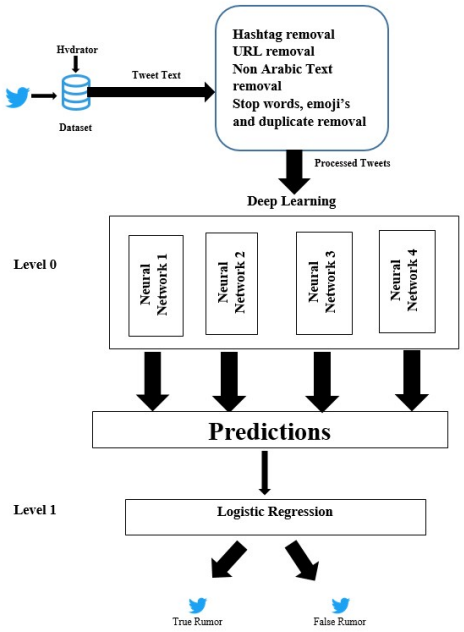
c. DEEP LEARNING ENSEMBLE STACKING MODEL

Concerning the context of deep learning stacking models, our initial focus is on constructing and developing neural network models, as illustrated in Figure 4. Our goal is to construct four unique architecture models and train them all using the same set of principles. Choosing the right activation function is a crucial step in building a neural network. Predicting whether a tweet is a rumor or not is a binary classification problem that we face in our scenario. As a result, the sigmoid function is an excellent tool to use. To further aid in the reduction of deviations between the predicted and actual outputs, we have opted to use binary cross-entropy as the loss function. We have decided to use the Adam optimizer, an adaptive learning rate optimization algorithm, to perform weight updates in the neural network during training. To guarantee deep learning, we will cycle through the training dataset 10-100 times (an epoch) and use the F1-score as the error metric to assess how well the model performed.



- *Figure 3 Ensemble Stacking Deep Learning Model*

After training the base capable learners, which consist of the four neural networks (Figure 3) on the data from the test set, we collect the base learners' predictions and utilize them for training the Meta learner. Concerning this case, each model will then produce a single prediction, which will either be a "rumor" value of 1 or a "not rumor" value of 0. In order to put the stacking phase into action, the first thing we will do is produce a stacked model input dataset using the outputs from the ensemble. Second, once the stacked dataset has been created, we will proceed to train the meta-learner on it using the data that has been provided.



- *Figure 4 The Proposed Ensemble Deep Learning Stacking model*

After working with both models, it has been revealed that the best approach for Arabic rumors detection was achieved by applying the said proposed ensemble-based deep learning model as shown in Figure 5.

4. PROPOSED MODELS EVALUATION

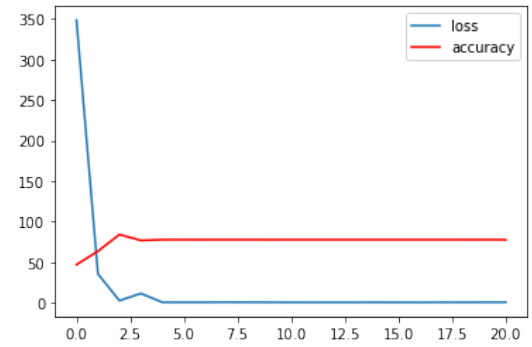
In light of what was covered before in this part, the outcomes of both of the rumor detection models that were proposed are addressed. The experiments that were carried out as a part of this research were carried out on a platform that consisted of Python 3.8 running on Windows 10. Windows 10 was the platform for the tests that were carried out as a part of this research, and Sklearn version 0.22.2 was the key Python library that we depended on for the majority of the implementation of the classifiers. Sklearn version 0.22.2 was the primary Python package that we depended on for the majority of the implementation of the classifiers. This was the case for the majority of the work. After converting the dataset's Tweet IDs into a csv file using the Hydrator , we started importing our ArCOV19-Rumors dataset in Visual Studio Code, which included a total of 321,093 tweets and was divided into 249,216 rumor tweets and 71,878 non-rumor tweets. We used 70% of our data set for training, while the remaining 30% was used for testing. The preparation processes of all of the classifiers were standardized to ensure that a fair comparison could be made. After that, the results obtained by the two proposed detection models were compared using the accuracy, recall, precision, and F1-score.

a. RESULTS OF MODEL ONE: ENSEMBLE STACKING MACHINE LEARNING MODEL

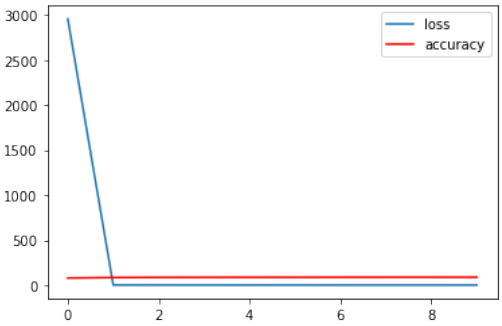
First of all, we begin the experiments by passing the cleaned and tokenized tweets' texts to three standalone machine learning classifiers. The decision tree classifier was the first standalone model passed in the model and gave an accuracy of 84%. The second model applied was the Random Forest classifier and gave and accuracy of 86%. The Third model applied was the Gaussian Naïve Bayesian classifier with an accuracy of 86%.
The Random Forest classifier and Gaussian NB are the highest accuracy given by with an accuracy of 86% and the decision tree with an accuracy of 84%. Hence, the standalone classifiers presented in this section showing good performance, they can be used as base/weak classifier for the ensemble models. The stacking-based classifier with LR (Stacking-LR) base gives a high accuracy of 87% outperforming the other standalone classifiers as shown in table.

b. RESULTS OF MODEL TWO: ENSEMBLE STACKING DEEP LEARNING MODEL

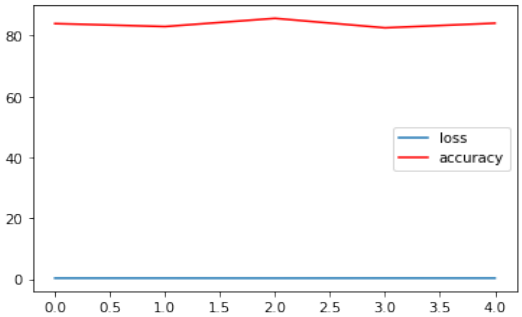
The experiments were started by passing the cleaned and tokenized tweets' texts to four standalone Deep Learning Neural Network classifiers. After training, the four neural networks model and recording the performance for each model. We found that the first neural network model results to an accuracy of 77%, The second neural network model results an accuracy of 87%, the third neural network model gave an accuracy of 89%, and the fourth neural network model results an accuracy of 77%. Loss and accuracy graphs are displayed separately for each model, Figure 5 to 9. The diagram displayed red and blue lines. The red value represents the proportion of true predictions made by the model based on the training data. The blue line, which represents the loss, shows how well the model can predict the correct output for each input.



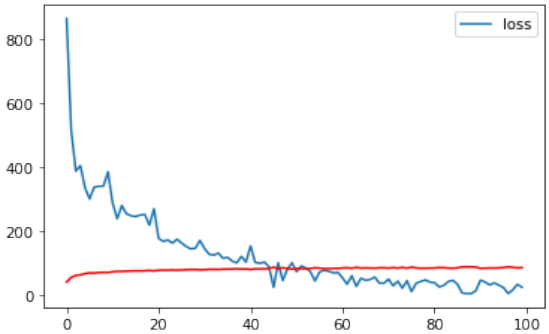
- *Figure 5 Loss and Accuracy Deep Leaning Model with depth 1*



- *Figure 6 Loss and Accuracy Deep Leaning Model with depth 2*



- *Figure 7 Loss and Accuracy Deep Leaning Model with depth 3*



- *Figure 8 Loss and Accuracy Deep Leaning Model with depth 4*

Since the standalone classifiers presented showing good performance, they were used as base/weak classifier for the ensemble deep learning model with LR. Concerning the proposed model, it syndicates the results acquired by the four neural network models and the machine learning algorithm (Stacking with Logistic Regression). The classification results show the overall performance of the proposed model and how our proposed model outperforms with an accuracy of 90%.

As a summary, Table 3 and Table 4 show the all the applied models performance (the standalone machine learning model, the ensemble machine learning model (stacking), and the neural network standalone model) compared with the proposed DL staking model.

In Table 3, the Random Forest outperformed the other classifiers concerning accuracy and f1score of 86% and 77%, respectively, due to its ability to increase the model performance due to its ability to integrate multiple classifiers to solve difficult issues [42]. Speaking of the ML stacking model, which gave much better results than the standalone ML models concerning accuracy, f1 score, precision, in addition to recall, with 87%, 78%, 89%, and 74%, respectively, due to its technique of combining multiple models that results in a better outcome and enhances the performance of Arabic rumor detection.

- *Table 3 Summary on the ML classifiers performance*

Models	Accuracy	F1score	Precision	Recall
Decision Tree	83%	68%	88%	65%
Random Forest	86%	77%	83%	73%
Gaussian NB	86%	74%	89%	69%
Machine Learning (stacked model)	87%	78%	89%	74%

In Table 4, the deep learning stacking model showed better and higher results in accuracy, f1score, precision, and recall with 90%, 84%, 86%, and 83%. Both the deep learning (DL) stacking ensemble model and the machine learning (ML) ensemble stacking model are ensemble methods that combine multiple models to improve the accuracy of predictions. However, DL models are more complex and can handle more complex data than ML models [31]. This was proven when our proposed ensemble stacking DL model showed

greater significance than the ML ensemble stacking model with 3 percentage points.

- *Table 4 Summary on the DL classifiers performance*

Models	Accuracy	F1score	Precision	Recall
Neural Network Model 1	77%	54%	49%	60%
Neural Network Model 2	87%	83%	87%	81%
Neural Network Model 3	89%	84%	86%	82%
Neural Network Model 4	77%	54%	49%	60%
Deep Learning (stacked model)	90%	84%	86%	83%

c. COMPARISON WITH THE STATE-OF THE-ART MODELS

A higher-level experiment was conducted to evaluate the performance of the latest paper using hybrid deep learning model and another paper using different language deep learning model on English tweets.

In the first comparison, our model surpassed the performance of the Arabic Fake News Detection (AFND) system, which relies on a hybrid deep learning approach encompassing both conventional neural networks and long short-term memory (CNN-LSTM) modalities [16]. By comparing accuracy, precision, F1 score, and recall, our proposed stacking deep learning model demonstrated superior results across all metrics.

Table 5 Comparison between the Proposed Model and the work done by [16]

Ref	Accuracy	F1score	Precision	Recall
[16]	81%	81%	81%	81%
The proposed model	90%	84%	86%	83%

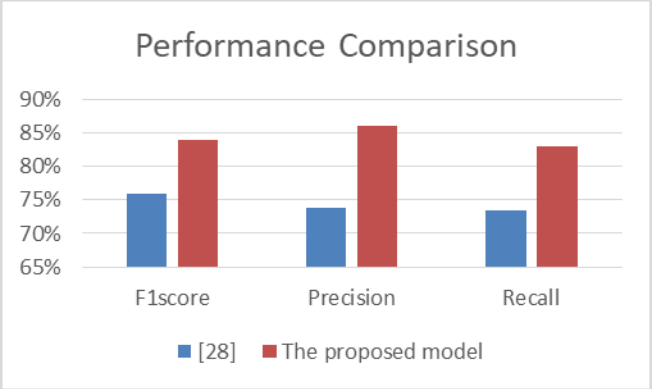


Figure 9 Performance Comparison with Arabic ensemble model presented by [16]

The other comparison held between the performance of our proposed model and another paper conducting number of experiments on different various models using manually annotated English tweets. In this paper they utilized various deep learning language models such as RNNs, CNN, BERT, RoBERTa, ALBERT, and performed a comparative analysis of these models for the misinformation type classification task. In their study, they found that the larger pretrained model RoBERTa performed better than the other models with an F1 score of 76% [28].

Table 6 Comparison between the Proposed Model and the work done by [28].

Ref	Accuracy	F1score	Precision	Recall
[28]	-	76%	73.75%	73.5%
The proposed model	90%	84%	86%	83%

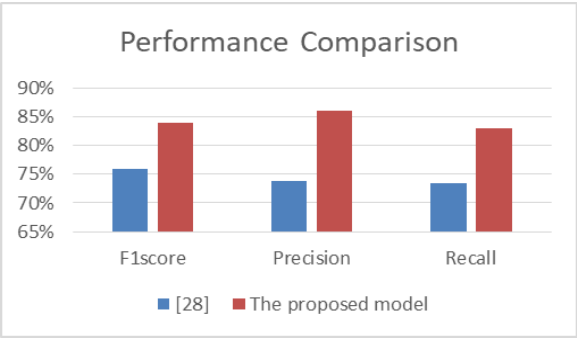


Figure 10 Performance Comparison with RoBERTa presented by [28]

5. CONCLUSION AND FUTURE WORK

The said study highlights the use of ensemble stacking through a new deep learning model, to investigate the complex world of Arabic tweets.

The study takes care with dataset selection, preprocessing, feature engineering, model creation, and evaluation. Empirical studies test a machine learning-based ensemble stacking model using a benchmark Twitter dataset. The two-tier DL ensemble stacking model launches the study. The first tier creates and evaluates self-learning neural network models. This prepares for the second tier, a logistic regression predictor that classifies tweets as rumor or not. Due to these significant empirical results, the DL ensemble stacking model is crucial. A 90% accuracy level shows the model outperforms its competitors. This greatly affects rumor detection and suppression. This enhancement, amounting to a 3% increase over the machine learning stacking approach, substantially augments vital performance metrics encompassing precision, accuracy, recall, and F1-Score, thus significantly refining the landscape of rumor detection.

In summary, this study underscores the compelling exigency of robust rumor detection mechanisms specifically tailored for Arabic-language social networks. By unveiling a cutting-edge hybrid model that synergistically combines Ensemble stacking-based Machine Learning and Deep Learning, the study not only advocates for heightened attention to the challenges of Arabic rumor detection but also substantiates the potential for significantly elevated accuracy and performance standards in the domain. Through these multifaceted contributions, the study forges a path toward a more effective and insightful approach to addressing the intricate landscape of rumor dissemination and veracity within the digital age.

Future research should compare more classifiers and ensemble methodologies to advance this field. Adding current tweets to the model's training data can also increase its accuracy. This strategic addition makes the study's conclusions more full and dependable and may illuminate previously unexplored areas. Careful research and more data may provide new views and improve rumor-finding. This increased understanding may make these strategies more beneficial for navigating the ever-changing social media landscape.

REFERENCES

[1] Xu, Shouzhi, X. Liu, K. Ma, F. Dong, B. Riskhan, S. Xiang and a. C. Bing, "Rumor detection on social media using hierarchically aggregated feature via graph neural networks.," *Springer Applied Intelligence,,* vol. 53, no. 3, pp. 3136-3149, 2023.

[2] Gao, Jie, S. Han, X. Song and F. Ciravegna, "A tweet level propagation context based deep neural networks for early rumor detection in social media," *arXiv preprint arXiv:2002.12683 ,* 2020.

[3] Cao, Juan, J. Guo, X. Li, Z. Jin, H. Guo and a. J. Li., "Automatic rumor detection on microblogs: A survey.," *arXiv preprint arXiv:1807.03505,* 2018.

[4] Zhou, Honghao, T. Ma, H. Rong, Y. Qian, Y. Tian and a. N. Al-Nabhan., "MDMN: Multi-task and Domain Adaptation based Multi-modal Network for early rumor detection," *Expert Systems with Applications ,* vol. 195, p. 116517, 2022.

[5] Gongane, V. U., M. V. Munot and a. A. Anuse, "Machine Learning Approaches for Rumor Detection on Social Media Platforms: A Comprehensive Survey.," *Advanced Machine Intelligence and Signal Processing,* pp. 649-663, 2022.

[6] Dito, F. Mohammed, H. A. Alqadhi and A. Alasaadi, "Detecting medical rumors on twitter using machine learning," *In 2020 International Conference on Innovation and Intelligence for Informatics, Computing and Technologies (3ICT), IEEE.,* pp. pp. 1-7., 2020.

[7] Wenfeng, Z. H. Gan and C. Ruoyi., "A Study on Online Detection of micro-blog Rumors Based on Naive Bayes Algorithm," *In 2020 Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC),* pp. p 22-25, 2020.

[8] A. Habib, S. Akbar, M. Z. Asghar, A. M. Khattak, R. Ali and U. Batool, "Rumor detection in business reviews using supervised machine learning.," *In 2018 5th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC), IEEE,* pp. pp. 233-237, 2018.

[9] Dubey, A. Kr, A. Singhal and a. S. Gupta, "Rumor detection system using machine learning.," *Int Res J Eng Technol (IRJET),* vol. 7, no. 5, pp. 2395-0056, 2020.

[10] G. Liang, W. He, C. Xu and L. Chen and J. Zeng, "Rumor identification in microblogging systems based on users' behavior," *IEEE Trans. Comput. Social Syst.,* vol. 2, no. 3, pp. 99-108, Sept.2015.

[11] Al-Sarem, Mohammed, W. Boulila, M. Al-Harby, J. Qadir and a. A. Alsaeedi., "Deep

learning-based rumor detection on microblogging platforms: a systematic review.," *IEEE access 7,* pp. 152788-152812., 2019.

[12] Cheng, Mingxi, S. Nazarian and P. Bogdan, "Vroc: Variational autoencoder-aided multi-task rumor classifier based on text," *In Proceedings of the web conference 2020.,* pp. 2892-2898, 2020.

[13] Kar, Debanjana, M. Bhardwaj, S. Samanta and A. P. Azad., "No rumours please! a multi-indic-lingual approach for covid fake-tweet detection.," *2021 Grace Hopper Celebration India (GHCI),* pp. PP 1-5. IEEE, 2021.

[14] T. G. Dietterich, " Ensemble methods in machine learning" in Multiple Classifier Systems.," *Berlin, Germany:Springer,* pp. pp. 1-15, 2000.

[15] D. O. a. R. Maclin, "Popular ensemble methods: An empirical study," *J. Artif. Intell. Res,* vol. 11, pp. pp. 169-198, Aug. 1999.

[16] Sorour, S. E. and H. E. Abdelkader., "AFND: Arabic fake news detection with an ensemble deep CNN-LSTM model," *J. Theor. Appl. Inf. Technol. 100, no. 14,* pp. 5072-5086, 2022.

[17] N. N. A. C.-R. K. I.-K. W. a. K. N. D. Wimalasena, "What makes a healthy home? A study in Auckland, New Zealand," *Building Research & Information,,* vol. 50, no. 7, pp. 738-754, 2022.

[18] S. Tasnim and M. M. H. a. H. Mazumder, "Impact of rumors and misinformation on COVID-19," *Journal of Preventive Medicine and Public Health,* vol. 53, no. 3, pp. 171-174, 2020.

[19] F. Haouari, M. Hasanain and R. S. a. T. Elsayed, "ArCOV19-rumors: Arabic COVID-19 twitter dataset for misinformation detection," *ArXiv Preprint ArXiv:2010.08768,* 2020.

[20] H. Wang, J. Gan, J. Chen and Z. OUYANG., "Automatic detecting for covid-19-related rumors data on internet.," *In 2021 9th International Conference on Communications and Broadband Networking,* pp. 22-26, 2021.

[21] M. Cheng, S. Wang, X. Yan, T. Yang, W. Wang, Z. Huang, X. Xiao, S. Nazarian and P. Bogdan., "A COVID-19 rumor dataset.," *Frontiers in Psychology. https://doi.org/10.3389/fpsyg.2021.644801,* 2021.

[22] Chen and Shuaipu., "Research on Fine-Grained Classification of Rumors in Public Crisis——Take the COVID-19 incident as an example," *In E3S Web of conferences,2020.,* vol. 179, p. 02027, 2020.

[23] Yang, Chen, X. Zhou and R. Zafarani, "CHECKED: Chinese COVID-19 fake news dataset.," *Social Network Analysis and Mining 11, no. 1 ,* 2021.

[24] J. MA, W. GAO and K.-F. WONG, "Rumor detection on twitter with tree-structured recursive neural networks," *Association for Computational Linguistics.,* 2018.

[25] T. Bian, X. Xiao, T. Xu, P. Zhao, W. Huang, Y. Rong and J. Huang2, "Rumor detection on social media with bi-directional graph convolutional networks," *In Proceedings of the AAAI conference on artificial intelligence,* vol. 34, pp. 549-556, 2020.

[26] M. Z. Asghar, A. Habib, A. Habib, A. Khan, R. Ali and A. Khattak, "Exploring deep neural networks for rumor detection," *Journal of Ambient Intelligence and Humanized Computing ,* vol. 12, pp. 4315-4333, 2021.

[27] Shams, A. Bin, E. H. Apu, A. Rahman, M. M. S. Raihan, N. Siddika, R. B. Preo, M. R. Hussein, S. Mostari and R. Kabir, "Web search engine misinformation notifier extension (SEMiNExt): A machine learning based approach during COVID-19 Pandemic," *In Healthcare MDPI,* vol. 9, p. 156, 2021.

[28] S. Kumar, R. R. Pranesh and K. M. Carley, "A Fine-Grained Analysis of Misinformation in Covid-19 tweets," *Research Square,* 2021.

[29] Al-Rakhami, M. S. and A. M. Al-Amri., "Lies kill, facts save: Detecting COVID-19 misinformation in Twitter.," *Ieee Access 8,* pp. 155961-155970., 2020.

[30] Kotteti, C. M. Madhav, X. Dong and L. Qian, "Ensemble deep learning on time-series representation of tweets for rumor detection in social media.," *Applied Sciences 10, no. 21,* 2020.

[31] Li, Guohui, M. Dong, L. Ming, C. Luo, H. Yu, X. Hu and B. Zheng., "Deep reinforcement learning based ensemble model for rumor tracking.," *Information Systems 103: 101772,* 2022.

[32] Abdelrazek, Aly, W. Medhat, E. Gawish and a. A. Hassan., Topic Modeling on Arabic Language Dataset: Comparative Study, Cairo, Egypt: Springer, November 21–24, 2022, pp. pp. 61-71.

[33] Castillo, Carlos, M. Mendoza and B. Poblete, "Information credibility on twitter," *In Proceedings of the 20th international conference on World wide web,* pp. 675-684, 2011.

[34] Qazvinian, Vahed, E. Rosengren, D. Radev and Q. Mei., "Rumor has it: Identifying misinformation in microblogs," *In Proceedings of the 2011 conference on empirical methods in natural language processing,* pp. 1589-1599, 2011.

[35] Ramos and Juan., "Using tf-idf to determine word relevance in document queries.," *In Proceedings of the first instructional conference on machine learning,* vol. vol. 242, pp. 29-48, 2003.

[36] K. Chaudhary, "Bagging, Boosting, and Stacking in Machine Learning," *https://dropsofai.com/bagging-boosting-and-stacking-in-machine-learning/,* 24 August 2020.

[37] A. Dutta, "Geeks For Geeks," 20 May 2019. [Online]. Available: https://www.geeksforgeeks.org/stacking-in-machine-learning/. [Accessed 15 May 2023].

[38] Zayno, Manahil and a. A. M. Radhi., "Data Mining Methods for Extracting Rumors Using Social Analysis Tools:," *Iraqi Journal of Science,* pp. 3618-3627., 2022.

[39] Bharadwaj, Pranav and a. Z. Shao., " Fake news detection with semantic features and text mining," *International Journal on Natural Language Computing (IJNLC) ,* vol. 8, 2019.

[40] Ahmad, Iftikhar, M. Yousaf, S. Yousaf and a. M. O. Ahmad., "Fake News Detection Using Machine Learning Ensemble Methods," *Complexity,* pp. 1-11, 2020.

[41] H. Li, "Statistical Learning Methods," *, Tsinghua University Press, Beijing, China,* 2016.

[42] M. Chaudhary, "Random Forest Algorithm - How It Works & Why It's So Effective," TURING, 2023. [Online]. Available: https://www.turing.com/kb/random-forest-algorithm. [Accessed 16 May 2023].