ISSN: 0067-2904

Data Mining Methods for Extracting Rumors Using Social Analysis Tools

Manahil Zayno *, Abdulkareem Merhej Radhi
Department of Computer Science, College of Science, Al-Nahrain University, Jadriya, Baghdad, Iraq

Received: 31/7/2021 Accepted: 11/10/2021 Published: 30/8/2022

Abstract

Rumors are typically described as remarks whose true value is unknown. A rumor on social media has the potential to spread erroneous information to a large group of individuals. Those false facts will influence decision-making in a variety of societies. In online social media, where enormous amounts of information are simply distributed over a large network of sources with unverified authority, detecting rumors is critical. This research proposes that rumor detection be done using Natural Language Processing (NLP) tools as well as six distinct Machine Learning (ML) methods (Nave Bayes (NB), random forest (RF), K-nearest neighbor (KNN), Logistic Regression (LR), Stochastic Gradient Descent (SGD) and Decision Tree (DT)). The data set size for the suggested experiment was 16,865 samples. For pre-processing tokenization was used to separates each one of the tokens from the others. Normalization that removes all non-word tokens, deleting stop words was utilized to remove all unnecessary words, and stemming was used to obtain the stem of the tokens. Prior to using the six classification algorithms, the major feature extraction approach Term Frequency -Inverse Document Frequency (TF-IDF) was applied. The RF classifier performed better compared to all other classifiers with an accuracy of 99%, according to the data.

Keywords: Machine learning, Text classification, Naïve Byes, RF, KNN, DT, Natural language processing, SGD).

طرق التنقيب عن البيانات لاستخراج الشائعات باستخدام أدوات التحليل الاجتماعي

مناهل زينو محمد* ، عبد الكريم مرهج راضي

قسم علوم الحاسوب، كلية العلوم، جامعة النهرين , بغداد , العراق

الخلاصه

تُعرَّف الشائعات بأنها عبارة لا يمكن التحقق من قيمتها الحقيقية. قد تنشر الشائعات معلومات خاطئة (معلومات كاذبة) على شبكة من الناس. يعد تحديد الشائعات أمرًا بالغ الأهمية في وسائل التواصل الاجتماعي عبر الإنترنت حيث تنتشر كميات كبيرة من المعلومات بسهولة عبر شبكة كبيرة من المصادر ذات سلطة غير مؤكدة. اقترح هذا البحث استخدام أدوات معالجة اللغة الطبيعية (NLP) و ستة خوارزميات مختلفة للتعلم الآلي (NB RF LR,KNN,DT,SGD,) لاكتشاف الشائعات. في هذا البحث ، الحجم الكامل لمجموعة البيانات يساوي 16،865 عينة، وتبدأ خطوات المعالجة المسبقة tokenization لكسر كل رمز عن الآخر، normalization لحذف جميع الرموز المميزة التي لا تحتوي على كلمات ، وإزالة stop words لحذف جميع الكلمات غير المهمة ، و Stemming للحصول على الجذع من الرموز. بعد ذلك، يتم استخدام طريقة

*Email: manahilzayno@gmail.com

استخراج الميزات الأكثر شيوعًا التردد– تردد المستند العكسي ا(TF–IDF) قبل تطبيق خوارزميات التصنيف الستة المقترحة. أظهرت النتائج أن مصنف Random Forest تفوق في الأداء على جميع المصنفات الأخرى بدقة تصل إلى 99٪. لمجموعات بيانات الشائعات.

1. INTRODUCTION

Data mining is the process of extracting useful information, and patterns from vast amounts of data by using various techniques [1]. There are many different Data mining techniques that are used based on the purpose of the mining process. Generally, data mining tasks are separated into two types: prediction and description[2,3]. The prediction uses supervised learning techniques to forecast the value of a specific characteristic based on the values of other attributes. Predictive modeling tasks include classification and regression clustering, mining associations, sequence discovery, and summarizing are examples of description tasks. Unsupervised learning techniques are used in these methods to uncover clear patterns in data[4].

It has been noted in the field of data mining that data grows rapidly. With the rapid growth of data and the rising availability of electronic documents, classification has become a critical task [5]. With the significant expansion of online social media in recent years, applying data mining techniques to social media data has attracted increasing interest [6]. Social networks are without a doubt the most widespread sources of information currently. This is Because of the massive volume of data and extensive social network connections. When scientists investigate social networks, they aim to tackle a few difficulties that arise from the complex relations which social networks contain[7]**.** The identification of rumors is one such issues.

Rumors are events about something which hasn't happened but is spreading from person to person as if it had. Social networks that have been considered excellent news gathering platforms have evolved to be a rumor tool for all topics and a powerful weapon to manipulate individuals [8]. Rumors have numerous detrimental consequences, which is a social issue. Peoples' and companies' reputations could be harmed, and good relationships could deteriorate. Families, people, businesses, and even governments can lose a lot of money. The deadly Nipah virus was spread through broiler chicken, according to a rumor that circulated on social media on May 30, 2018. According to the message, the Nipah virus is spread through chicken, and as a result, several dealers in Tamil Nadu have experienced significant losses [9]. This case of a false rumor emphasizes the need of predicting the veracity of content on social media in an automatic way. Various research on developing systems that automatically detect rumors rely heavily on Artificial Intelligence (AI) techniques like NLP and ML tools. This paper, therefore, aims to build a model using supervised learning techniques using (NB, RF, DT, LR, KNN, SGD) methods to classify rumors within the social network.

2. Contribution:

The following are the main contributions of this paper

- This paper engaged rumor detection in a methodical process, which offers a way of classifying the features of rumors and seeing their relationships.
- Pruning and extracting valuable features using proposed methods.
- Improve the performance of the rumor detection to obtain better results compared to previous works

3. Related works

A few relevant papers on rumor detection are presented in this section:

In 2015, Qiao Zhang et al.[10] suggested the focus on detecting rumors on social networks. To distinguish rumors from normal messages, they proposed a rumor detection method based on implicit features of contents and users. They published the findings of the detection approach of rumors by means SVM classification based on implicit characteristics regarding

the content and users obtain higher performance through a comparison according to basic features. They believe user credibility is a significant aspect that influences information credibility, analyzing user credibility can aid in the detection of message credibility. So some work should be done on user credibility in the future, and leverage that to improve rumor detection performance

In 2015, Gang Liang et al.[11] identified rumors on the Sina Weibo platform based on user's behavior by means of ML. The behavior of users was used as a clue for showing who tends to be a rumor maker. They observed that the rumor publishers' behaviors may diverge from normal users' and a rumor post may have different responses from a normal post. They propose rumor identification schemes, based on user behavior, in which users' behaviors are treated as hidden clues to indicate who is likely to be rumormongers or what posts are potential rumor microblogs.

In 2017 Ma, B., Lin, D., and Cao, D[12] employed 2 text representations to construct text vectors from the rumor content: the NN language and bag of words (BoW) model. Using advanced classification algorithms, they compared the performance of 2 text representations in rumor detection. The best classification accuracy of the BoW model has been more than 90%, while the optimal classification accuracy related to the NN language model has been more than 60%, according to 10,000 Sina Weibo posts. They conclude that words of posts are more beneficial than semantic context vector representation in a small data set. Finally, they propose to study how to integrate the context information to promote the performance of rumor detection in the future.

In 2018 Vijeev, A., Mahapatra, A., Shyamkrishna, A., and Murthy, S.[13] suggested the PHEME dataset, which contains non-rumors and rumors about 5 big events, and created a rumor identification algorithm that classifies tweets. They began by analyzing and ranking a variety of user- and content-based features. NLP methods are used to generate certain content-based attributes. After that, they used various combinations of features for the training of multiple ML models (SVM, RF, and NB). Lastly, they compared the models' performance. The models' performance on one such event had led to 78% accuracy. Finally, they believe that further improvements are needed both in accuracy and in processing before a fully automated rumor detection solution can be integrated into a microblogging site.

Because natural language processing libraries for Indian languages do not yet exist, establishing them and then using them to extract content-based characteristics like sentiment analysis could help classify tweets in Indian languages with greater accuracy. Also, finding out the source of a rumor by building a social network graph can be considered as a worthy addition to the existing rumor detection system.

In 2020, Pratiwi, A. R. D., and Setiawan, E. B.[14] suggested a system for detecting rumors in Indonesian-language. According to SVM classification and feature selection using TF-IDF weighting. With the greatest accuracy score of 78.71%, the system performs well while employing 10% of the testing data and unigram characteristics. They believe their System performance was influenced by preprocessing and labeling. There are still a lot of non-standard words missing from the normalization dictionary. This results in words that cannot be processed in preprocessing and becomes a separate feature so that the number of features used is increasing. Labeling was also done manually using human intelligence so that the error rate in labeling the data can be quite high. They propose to use other classification models with the additional features to find out which model has the best performance and influential features in detecting rumors on Twitter in future work.

In 2020 Dubey, A. K., Singhal, A., and Gupta, S [15] suggested work on Gradient Boosting, Multinomial Naive Bayes, and RF with specific datasets for implementing them and move closer to more adequate rumor conclusions. In this case, when using Multinomial Naive Bayes, the accuracy was approximately 90.4% when using RF, it is about 86.5%; and in the

case of using Gradient Boosting, it's about 88.3%. Finally, they believe working on granular aspects of the data and using techniques that are more advanced could lead to far better results in the future.

4. Methodology

The proposed system was performed in five steps. Figure (1) shows a diagram that depicts a general view of these steps. The initial step is to choose and preprocess the relevant Rumor dataset from kaggel.com. After that, after separating the dataset into training 70 % and testing 30 % data, TF-IDF is used to extract word features. The next step is to categorize the data using classifiers, such as Naive Bayes, Decision Tree, Random Forest, K-nearest Neighbor, Logistic Regression, and Stochastic Gradient Descent. Finally, evaluate the model performance using various metrics, such as (accuracy, recall, and precision).

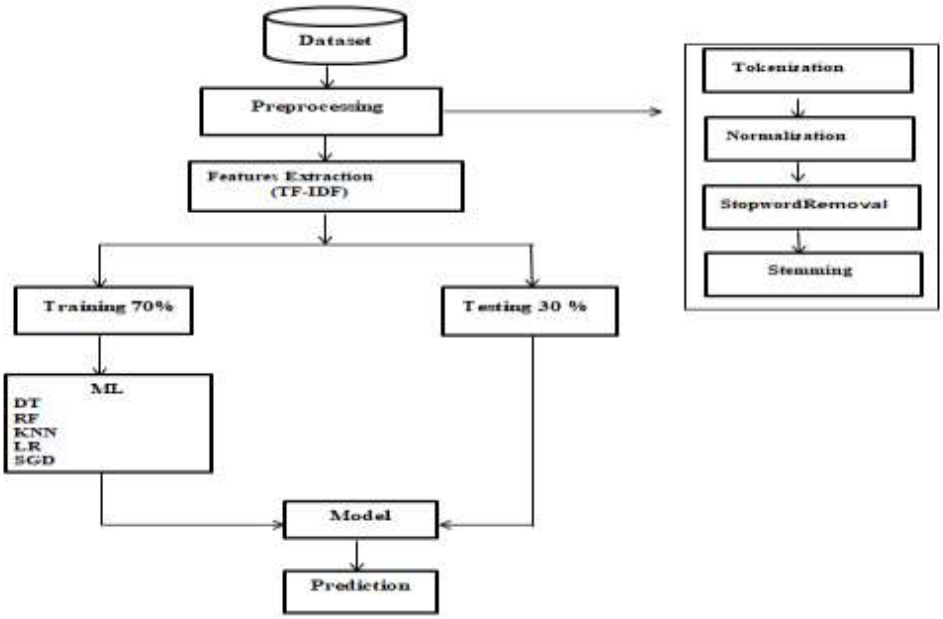

Figure 1-Block Diagram of the proposed system

4.1 Dataset

The datasets are crucial in the process of classification. To train and test the classifier, the collection Rumor Dataset was used. In this paper. The dataset (snopes.CSV file) was collected from https://www.kaggle.com/arminehn/rumor-citation[16] This dataset has about 16,865 records. It contains seven labels of rumors (true, false, mfalse (mostly false), mtrue (mostly true), mixture, legend, or undetermined.

4.2 Preprocessing Step

Because the words, characters, and sentences detected at such a level are the fundamental units transmitted to all subsequent processing stages, it is a crucial aspect of any NLP system [17]. The presented section depicts the execution related to preprocessing processes that are identical in testing and training phases. This stage has a major benefit in that it organizes data to make the rumor detection work easier. Tokenization, normalizing, removing stop words, and stemming are the four processes of preprocessing

4.2.1 Tokenization

In NLP, tokenization is crucial. Tokenization is the process of separating a written document into tokens by utilizing space to divide one word from the next. Words, symbols, and numbers are examples of tokens. The token outputs become inputs for the following steps of pre-processing [18]

4.2.2 Normalization

This is the process of unifying diverse forms of the same letter (by converting all letters to uppercase or lowercase), as well as the deletion of all non-letters (digits and symbols) before

using the stop words and stemming method[19, 20]

4.2.3 Stop Word

Stop words are often occurring words that might be characterized as any word that is insignificant in the classification process and has no obvious value. Stop words like (at, a, by, be, is, in, was what, on, where, when, will, who, etc.). These terms were eliminated from each one of the documents, and the processed documents were saved and forwarded to the next phase [21].

4.2.4 Stemming

The root/stem of a word is determined using stemming procedures. Stemming reduces words to their stems, using a significant amount of language-specific linguistic knowledge. The words users, user, used, and using, for instance, can all be derived from the word 'USE' [22].

4.3 Feature Extraction

Feature extraction can be defined as a crucial stage in rumor categorization, as it extracts the features from the text input after it has been preprocessed. In-text processing, the challenge of transforming a specific text into a vector-based on space is crucial. The TF-IDF is a major approach. This method considers the number of times the word appears in all documents in the document set. The other term has been Inverse Document Frequency (IDF) that has been assessed as a logarithm of the number of documents in the dataset, divided by the number of documents where a specific term appeared. It will be feasible to evaluate the TF-IDF once the TF and IDF values have "been received [23]." as shown in equations (1,2,3).

$$\mathrm{TF}\ (\mathrm{t},\mathrm{d})\ = \frac{\text{term t count in d}}{\text{count of term } T_d} \quad (1)$$

Where TF represents Term Frequency, t denotes the term, and d represents the document. TF (t, d) depicts the (times term t appears in a document) / (Total number of terms in the document) through the calculation of the following equation

$$\mathrm{IDF(i)}\ = \mathrm{Log}_2(\tfrac{\mathrm{N}}{\mathrm{N_i}}) \quad (2)$$

N represents the total number of documents in a group of documents. Ni represents the number of documents where the word i had arisen in a set of documents.

$$\mathrm{TF}-\mathrm{IDF} = TF_i * (\mathrm{Log}_2(\tfrac{\mathrm{N}}{\mathrm{N_i}})) \quad (3)$$

TF: number of times where the word i appeared in a document.

N: total number of the documents in a group of the documents.

Ni: number of the documents where the word i had occurred in a group of the documents[24]

4.4. Classification Techniques:

In the context of data mining, information or potentially useful patterns are typically hidden and unknown, so automatic techniques are required to facilitate the extraction of this data. The information in text mining is obvious, but the problem is that this information is not represented in a way suitable for processing by a computer. The goal of text mining is to represent data in texts in a way that can be processed automatically[25, 26]. Text mining can be defined as applying algorithms and methods from machine learning and statistics to natural language texts to extract nontrivial information for further use [25]. There are numerous applications for machine learning, but data mining is the most important.

Machine learning can be divided into two categories: supervised machine learning and unsupervised machine learning [27]. In unsupervised learning, the training data are unlabeled and the algorithm must learn without prior assistance. On the other hand, in supervised learning, the training set used to feed the algorithm includes the desired solutions, called labels.[28]. Supervised techniques can be further classified into two main categories: classification and regression. The output variable in regression accepts continuous values, whereas the output variable in classification takes class labels[29].

Classification is a data mining (machine learning) technique that is used to forecast group

membership for data instances[30]. Although there are a variety of machine learning techniques available, classification is the most often utilized. In machine learning, classification is an important activity, especially in future planning and information discovery [31]. Classification is categorized as one of the most studied problems by researchers in the machine learning and data mining fields[32]. There are several classification techniques that can be used for classification purposes. In this work, we will focus only on six classifiers that were used to build the rumor detection model.

4.4.1 Naïve Bayes (NB) Algorithm

The Naïve Bayes classifier can be defined as a supervised learning algorithm based upon the Bayesian theorem and a group of conditional independence assumptions on attributes. It can be calculated using Bayes' rule:

$$P(C/X) = \frac{P(X/C).P(C)}{P(X)} \quad (4)$$

P (C|X) represents the posterior probability of class (target) given predictor (attribute).
P (X|C) represents likelihood, which is the probability of the predictor of a given class.
P (C) represents the prior probability of a class.
P(X) represents the predictor's prior probability[33]

4.4.2 Decision tree (DT) Algorithm

The DT classifiers may be utilized in classification as well as regression. The classifier can predict the target variables through learning feature data and dividing the area to sub-areas. Based on two criteria, multiple features have been divided, one of them is an Entropy measure and the other is the information gain[34]. considering a binary (2-class) classification C, and a set of examples, S represents the class distribution at any node, which may be expressed in the form of (p0, p1), where p1 = 1- p0, and entropy, H (S) represents the information sum:

$$\boldsymbol{H(p) = -\sum_{i=1}^{c} P_i \log_2 P_i} \quad \textbf{(5)}$$

For the determination of the optimal attribute to select for every one of the decision nodes of a tree. The optimal attribute is the attribute best for separating them into homogeneous sub-sets. More particularly, the Gain (S, A), of the attribute A, relative to a collection of samples S, can be characterized as.

$$\textbf{Gain (S, A) =} \text{ Impurity } \textbf{(S)-} \sum_{i=1}^{k} \frac{|S_i|}{|S|} \text{ Impurity } (\boldsymbol{Sv_i}) \quad (6)$$

4.4.3 Random Forest (RF) Algorithm

RF classifier is considered as an ensemble approach that boosts accuracy by using various decision trees[35]. RF includes a large number of decision trees that collaborate to forecast the outcome of a class, with the final prediction based upon the class with the most votes. When compared to other models, the error rate in RF is low due to the lack of correlation among the trees[36].

4.4.4 K-nearest Neighbor (KNN) Algorithm

A common example-based classifier is k-NN. Because of its accuracy and simplicity, it is one of the most widely used classification approaches. Euclidean distance was used since it is the most common method where humans interpret distance in the real world [37, 38] as illustrates in the equation.

$$\textbf{D} = \sqrt{\sum_{i=1}^{k} (X_i - Y_i)^2} \quad (7)$$

Where:
d= Euclidean distance
X= data point from the dataset
k= number of dimensions
Y=new data point to be predicated.

4.4.5 Logistic Regression

The advanced Linear regression method of logistic regression is applied to categorize both non-linear and linear data.[39]. It's a classifier that figures out which properties from the input are the most significant in distinguishing between the various classes [40]. The logistic regression equation is shown in Equations (8,9):

$$Z = \sum_{i=1}^{N} w_i f_i\ (Y, X) \quad (8)$$

$$p = \frac{1}{1+e^{-Z}} \quad (9)$$

Where:
P represents the likelihood of (Y|X)
X represents a feature-vector
Y represents a class
w represents the weight of the word
f represents frequency.

4.4.6 Stochastic Gradient Descent Learning

Gradient descent is a commonly utilized algorithm that might provide a fresh viewpoint on issue resolution.[41]. Stochastic gradient descent is a classifier that optimizes an objective function using an iterative process. Because it evaluates gradients using randomly picked samples, this classifier is known as stochastic [42].

4.5 Evaluating performance measures

A variety of evaluation measures were utilized to evaluate the algorithm's classification accuracy in detecting Rumors. The performance metrics that are utilized for evaluating the results of the classification are precision, recall, and F-measure. In this work, the most frequently utilized measure metric (Confusion Matrix) to detect Rumor was used. Through the formulation of this as a task of classification, it is possible to define the measures that the confusion matrix has as below[43, 44].

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (11)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (12)$$

$$\text{F1-score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision+Recall}} \quad (13)$$

Were
• TP: true positives: samples predicted positive that are actually positive.
• TN: true negatives: samples predicted negative that are actually negative.
• FP: false positives: samples predicted positive that are actually negative.
• FN: false negatives: samples predicted negative that are actually positive [45, 46].

4. Results and Discussion

In this section, all results are shown in tables (1) and figure (2) of used evaluation metrics applied to classify the rumors accurately.

Table -1 Results metrics for each classifier about Snopes dataset

Snopes dataset			
Classifier Name	**Precision%**	**Recall%**	**F1-measure %**
RF	0.99	0.99	0.99
NB	0.91	0.86	0.87
LR	0.94	0.91	0.92
KNN	0.82	0.75	0.65
DT	0.98	0.98	0.98
SGD	0.96	0.96	0.96

In this paper, rumors detection in online social media was modeled as a classification problem. The Snopes dataset (of the seven labels) was split into 70% for training and the rest

30% as testing data. The meaningful 'Content Based' features of the data was extracted, with the use of the Natural Language Processing approaches.

We have compared 6 different machine learning models. Table (1) and Figure (2) summarizes the highest precision that has been obtained for the variety of the classifiers. From the results shown above, the precision of the random forest algorithm achieved the best perfumes (99%). However, the performance of the DT and SGD is also near to Random Forest with 98 % and 96% precision. While the precision of the other three classifiers, NB, LR, KNN achieved (90%), (94%), (82%).

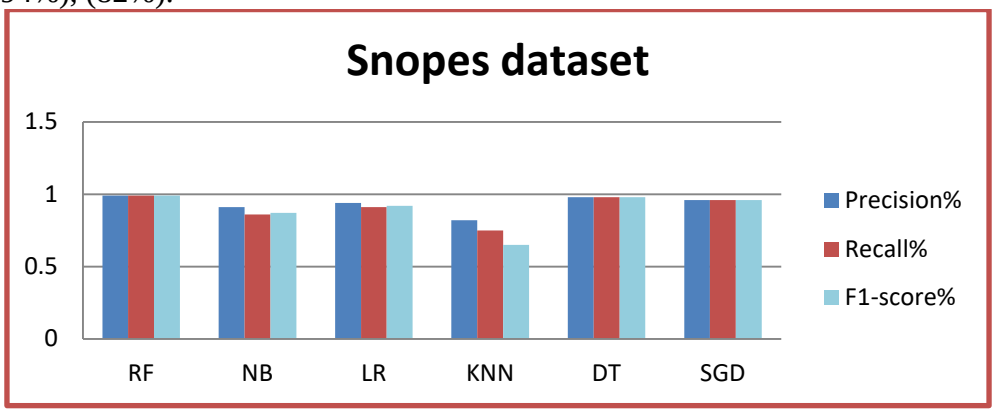

Figure 2-Results of evaluation metrics for each classifier about Snopes dataset

Through the analysis of various performance metric graphs in this section, it was found that RF performed the best (99%) in terms of accuracy.

5. Conclusions and Future Works

The goal of this research was to use supervised learning algorithms for detecting rumors in social media. We experimented with various ML classifiers, and based on the findings, the study came to the following conclusion:-

- RF performed the best (99%) with regard to the accuracy, while the performance of SGD, DT, was also very near to RF with 96% and 98% accuracy.
- The type of dataset collected (Snopes dataset) also has a considerable effect on the classification accuracy of this work.
- The preprocessing steps using our dataset give better results. These steps had an significant impact on increasing the accuracy of the classification.

For future work we suggest using deep learning algorithms with increment in the size of the dataset.

References

[1] S. R. Guruvayur and R. Suchithra, "A detailed study on machine learning techniques for data mining," *2017 International Conference on Trends in Electronics and Informatics (ICEI)*, 2017, pp.1187-1192,doi:10.1109/ICOEI.2017.8300900.
Available: https://ieeexplore.ieee.org/document/8300900.

[2] Sunita Beniwal, Jitender Arora, 2012, Classification and Feature Selection Techniques in Data Mining, International journal of engineering research & technology (IJERT) Volume 01, Issue 06 (August 2012). Available: https://www.ijert.org/classification-and-feature-selection-techniques-in-data-mining.

[3] S. Umadevi and K. S. J. Marseline, "A survey on data mining classification algorithms," 2017 International Conference on Signal Processing and Communication (ICSPC), 2017, pp. 264-268, doi: 10.1109/CSPC.2017.8305851. Available: https://ieeexplore.ieee.org/document/8305851.

[4] Jain, S., R. Raghuvanshi, and M. Ilyas, "*A survey paper on overview of basic data mining tasks",* International Journal of Innovations & Advancement in Computer Science (IJIACS), 2017. 6(9).

[5] Patil, L.H. and A. Mohammad, "*A multistage feature selection model for document classification using information gain and rough set,"* International Journal of Advanced Research in Artificial Intelligence (IJARAI), 2014. 3(11).

[6] Barbier, G., Liu, H. (2011). "Data Mining in Social Media," *In: Aggarwal, C. (eds) Social Network Data Analytics*. Springer, Boston, MA. https://doi.org/10.1007/978-1-4419-8462-3_12.

[7] E. V. Altay and B. Alatas, "Detection of Cyberbullying in Social Networks Using Machine Learning Methods," *2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT)*, 2018, pp. 87-91, doi: 10.1109/IBIGDELFT.2018.8625321.

[8] Bingöl, Harun & Alatas, Bilal. (2019). "Rumor Detection in Social Media Using Machine Learning Methods". 1-4. 10.1109/UBMYK48245.2019.8965480.

[9] Ajeet Ram Pathak, Aditee Mahajan, Keshav Singh, Aishwarya Patil, Anusha Nair, Analysis of Techniques for Rumor Detection in Social Media, Procedia Computer Science, Volume 167,2020,Pages 2286-2296,ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.03.281..

[10] Zhang, Q., Zhang, S., Dong, J., Xiong, J., Cheng, X. (2015). Automatic Detection of Rumor on Social Network. In: Li, J., Ji, H., Zhao, D., Feng, Y. (eds) Natural Language Processing and Chinese Computing. NLPCC 2015 2015. Lecture Notes in Computer Science(), vol 9362. Springer, Cham. https://doi.org/10.1007/978-3-319-25207-0_10.

[11] G. Liang, W. He, C. Xu, L. Chen and J. Zeng, "Rumor Identification in Microblogging Systems Based on Users' Behavior," in IEEE Transactions on Computational Social Systems, vol. 2, no. 3, pp. 99-108, Sept. 2015, doi: 10.1109/TCSS.2016.2517458.

[12] Ma, B., D. Lin, and D. Cao, "*Content representation for microblog rumor detection"*, in *Advances in Computational Intelligence Systems*. 2017, Springer. p. 245-251.

[13] Vijeev, A., "A Hybrid Approach to Rumour Detection in Microblogging Platforms," Abhishek Vijeev and Anushreya Mahapatra and Arundhati Shyamkrishna and Srinivasa Murthy, 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI),{ 2018}, P.337-342.

[14] Pratiwi, A.R.D. and E.B. Setiawan, "*Implementation of Rumor Detection on Twitter Using the SVM Classification Method,"*. Journal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 2020. 4(5): p. 782-789.

[15] Dubey, Anil Kr, Apoorv Singhal, and Sarthak Gupta. "Rumor detection system using machine learning." *Int Res J Eng Technol (IRJET)* 7.05 (2020): 2395-0056.

[16] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In Proceedings of the 20th international conference on World wide web (WWW '11). Association for Computing Machinery, New York, NY, USA, 675–684. https://doi.org/10.1145/1963405.1963500

[17] Kannan, S., et al., *Preprocessing techniques for text mining.* International Journal of Computer Science & Communication Networks, 2014. 5(1): p. 7-16.

[18] Alhaj, Y.A., et al., "*A study of the effects of stemming strategies on arabic document classification,"*. IEEE Access, 2019. 7: p. 32664-32671.

[19] Abuhaiba, I.S. and H.M. Dawoud, "*Combining different approaches to improve arabic text documents classification,"* International Journal of Intelligent Systems and Applications, 2017. 9(4): p. 39.

[20] Casamayor, A., D. Godoy, and M. Campo, *Identification of non-functional requirements in textual specifications: A semi-supervised learning approach.* Information and Software Technology, 2010. 52(4): p. 436-445.

[21] Jain, C. and S. Vignesh, *Era of Sociology News Rumors News Detection using Machine Learning*. 2019.

[22] Srividhya, V. and R. Anitha, *Evaluating preprocessing techniques in text categorization.* International journal of computer science and application, 2010. 47(11): p. 49-51.

[23] Shaker, S.H. and N.M. Jaafar, *An E-Exam Management System under E-Network Management Course*.

[24] Wang, M.-J. and Y.-Z. Li. *Hash function with variable output length*. in *2015 International Conference on Network and Information Systems for Computers*. 2015. IEEE.

[25] Hotho, A., Nürnberger, A. & Paaß, G. (2005). "A Brief Survey of Text Mining," *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, 20, 19-62..

[26] Singh, M. P. (2004). *The practical handbook of internet computing*. Chapman and Hall/CRC.

[27] Kotsiantis, S.B., I. Zaharakis, and P. Pintelas, *Supervised machine learning: A review of classification techniques.* Emerging artificial intelligence applications in computer engineering, 2007. 160(1): p. 3-24.

[28] Dias Canedo E, Cordeiro Mendes B. Software Requirements Classification Using Machine Learning Algorithms. *Entropy*. 2020; 22(9):1057. Available: https://doi.org/10.3390/e22091057.

[29] Soofi, Aized & Awan, Arshad. (2017). "Classification Techniques in Machine Learning: Applications and Issues," Journal of Basic & Applied Sciences. 13. 459-465. 10.6000/1927-5129.2017.13.76.

[30] Kesavaraj, G. and S. Sukumaran. *A study on classification techniques in data mining*. in *2013 fourth international conference on computing, communications and networking technologies (ICCCNT)*. 2013.

[31] Singh, M., S. Sharma, and A. Kaur, "*Performance analysis of decision trees.* International Journal of Computer Applications," 2013. 71(19). Doi: 10.5120/12593-9232.

[32] Baradwaj, B.K. and S. Pal, "*Mining educational data to analyze students' performance,"* arXiv preprint arXiv:1201.3417, 2012.

[33] Dhande, L.L. and G. Patnaik, *Review of sentiment analysis using naive bayes and neural network classifier*. International Journal of Scientific Engineering and Technology Research (IJSETR), 2014. 3(7): p. 1110-1113.

[34] Kaur, S., P. Kumar, and P. Kumaraguru, *Automating fake news detection system using multi-level voting model.* Soft Computing, 2020. 24(12): p. 9049-9069.

[35] Bharadwaj, P. and Z. Shao, *Fake news detection with semantic features and text mining.* International Journal on Natural Language Computing (IJNLC) Vol, 2019.

[36] Iftikhar A., Muhammad Y., Suhail Y., Muhammad A., "Fake News Detection Using Machine Learning Ensemble Methods", *Complexity*, vol. 2020, Article ID 8885861, 11 pages, 2020. https://doi.org/10.1155/2020/8885861

[37] Hamood A., Sabrina T., Nazlia O., and Mohammed A., "Experiments on the Use of Feature Selection and Machine Learning Methods in Automatic Malay Text Categorization Procedia Technology," Volume 11, 2013, Pages 748-754, ISSN 2212-0173,https://doi.org/10.1016/j.protcy.2013.12.254.
Available: https://www.sciencedirect.com/science/article/pii/S2212017313004088.

[38] Tuerhong, G., M. Wushouer, and D. Zhang. *An Improved K Nearest Neighbor Classifier for High-Dimensional and Mixture Data*. in *Journal of Physics: Conference Series*. 2021. IOP Publishing.

[39] Aborisade, O. and M. Anwar. *Classification for authorship of tweets by comparing logistic regression and naive bayes classifiers*. in *2018 IEEE International Conference on Information Reuse and Integration (IRI)*. 2018.

[40] Indra, S., L. Wikarsa, and R. Turang. *Using logistic regression method to classify tweets into the selected topics*. in *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. 2016.

[41] Prasetijo, A.B., et al. *Hoax detection system on Indonesian news sites based on text classification using SVM and SGD*. in *2017 4th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*. 2017.

[**42**] Ozbay, F.A. and B. Alatas, *Fake news detection within online social media using supervised artificial intelligence algorithms.* Physica A: Statistical Mechanics and its Applications, 2020. 540: p. 123174.

[43] Shu, K., et al., *Fake news detection on social media: A data mining perspective.* ACM SIGKDD explorations newsletter, 2017. 19(1): p. 22-36.

[44] Tharwat, A. (2021), "Classification assessment methods", *Applied Computing and Informatics*, Vol. 17 No. 1, pp. 168-192. https://doi.org/10.1016/j.aci.2018.08.003

[45] Mokgonyane, T.B., et al. *Development of a text-independent speaker recognition system for biometric access control*. in *Southern Africa Telecommunication Networks and Applications Conference (SATNAC)*. 2018.

[46] Sefara, T.J. *The effects of normalisation methods on speech emotion recognition*. in *2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*. 2019.