A Machine Learning Approach For Early Detection Of Social Media Rumors

Subash. A[1], Dr. P. Vijayakumar[2]
[1]Dept of Computer Science and Information Technology
[2]Associate Professor and Head, Department of Computer Applications
[1, 2] Sri Jayendra Saraswathy Maha Vidyalaya College of Arts & Science, Coimbatore-641005

Abstract- *Social media platforms have become an integral aspect of our lives in the modern day. Social media platforms such as Facebook, Instagram, Twitter, SnapChat, and YouTube are used to connect people and promote companies. Twitter is a massive communication and sharing network where users can express themselves and advertise their companies via 140-character tweets. Each month, around 42 million new Twitter accounts are established. Rumors are easily propagated and spread in crowds, particularly in Online Social Networks (OSNs), owing to their open nature and large user base. Introduced SBR approach for early detection of social media rumors and malicious social bots in this paper. To validate the datasets, to use machine learning algorithms with varied window sizes to pick essential components and local characteristics. Classification was performed using KNN, DT, SVM, NB, and RF classifiers and achieved an accuracy of 97.9%. Experiments using twitter datasets reveal that the proposed model outperforms previous content-based approaches in detecting and verifying early rumors.*

Keywords- OSN, ML, EDSMR-RD, Social Media, Behavior

I. INTRODUCTION

Twitter is a popular online social networking and microblogging service that was founded in 2006. Its defining feature is its extraordinary simplicity [1]. Its community interacts by text-based communications dubbed tweets. Online Social Networks (OSNs) such as Facebook, Instagram, and Twitter have become ingrained in the daily lives of millions of people worldwide. Indeed, they are the world's most popular online applications, with around 3 billion users [2]. By linking users, OSNs allow them to interact with, consume, produce, and trade material with one another. Today, social media has conventional modes of communication, resulting in a shift in the way information is provided to a broad audience, making it simple to share information in a short amount of time [3].

However, the accuracy of the information or news that is distributed over social media is rarely validated, and it has often resulted in confusion among people worldwide [4]. Social media platforms such as Facebook and Twitter lack a technology that can identify any information being circulated on the website as a rumor and pull it down immediately if required before it causes any problems [5]. At the moment, these massive social media sites rely entirely on their staff and available time to label any information as a rumor, which may have worked successfully 10 or 15 years ago when social media usage and data were smaller, but now that social media sites have billions of users and a massive amount of data growing exponentially, it becomes extremely difficult to manually verify the veracity of all the information flowing on social media [6] [7].

In June 2016, a fake "Facebook privacy notice" spread like wildfire on social media, urging users to copy and paste a specific piece of text onto their Facebook wall in order to retain their profile privacy, including the things they share, photos they upload, and personal information, which ultimately proved to be a rumor, and millions of Facebook users shared it, becoming victims of this false claim [8]. A rumor is an unconfirmed piece of information whose authenticity and source are unknown and which may cause damage in a variety of ways [9] [10].

Additionally, several feature selection techniques have been applied to reduce redundant and superfluous traits. Second, the data is pre-processed using KNN, SVM, Nave Bayes, Random Forest, and Decision Tree algorithms. The authors describe a unique approach for identifying tweets created by actual persons or bots that combines behavior detection with the EDSMR-RD algorithm.

II. REVIEW OF LITERATURE

Corcoglioniti et al. (2018) suggested a recommendation system that utilizes machine learning methods to predict factors ranging from fundamental SM qualities to domain-specific user profile attributes.

Cao et al. (2017) developed an app recommendation matrix factorization-based latent factor model using numerical ratings and textual input from a variety of mobile platforms.

The rating matrices from these multiple platforms were factored, and the textual content was included into the topic models.

Faulkner et al. (2014) pioneered a unique approach to document-level stance classification by using two feature sets to capture the linguistic properties of the stance-taking language. Additionally, they developed a corpus of annotated student essays for position in order to assess the use of elements derived from linguistic studies on argumentative language. This corpus provides a more representative sample of argumentation language than the often disorganized internet conversation material used in attitude classification studies.

Guidi, B., and Michienzi, A. (2020) advised doing a research to better understand the behavior of (human) users and (automated) bots in BOSMs in order to develop a mechanism for bot detection in future work. The authors picked Steem as a case study since it was the first BOSM proposal and, at the time of writing, the most successful. Around 29 million blocks from the Steem blockchain compose the dataset, which includes human activity and an initial bots collection. According to the authors' statistics, actual users were more active than bots. Nonetheless, this last group engages in a broad variety of activities and was, on average, far more involved than the general population. Additionally, the authors discovered that reputation was not a desirable characteristic for a bot identification method. Contrary to common assumption, bots have a positive reputation.

Hercig et al. (2015) developed a strategy for identifying an online debater's point of view. A new database of Czech news comments was constructed, and classifiers based on maximum entropy, support vector machine, and convolutional neural networks were evaluated.

The authors Heredia, B. et al. (2018) investigate the effect of social bots on the tone of tweets concerning both presidential candidates. Between September 22 and November 8, 2016, data from the sentiment140 dataset was utilized to train a convolutional neural network (ConvNet). The dataset previously had 705,381 distinct accounts; however, inactive accounts were deleted, leaving 570,532 distinct persons.
Jiménez-Bravo et al. (2019) developed a recommendation system that displays expert profiles to other users. Consumers must be able to take in exciting information from experts. The expert suggested to a user will be determined by the information they publish and if the material is of interest to the user.

Rodrguez et al. (2016) created a fuzzy logic approach for following a Twitter tip. This strategy treats ideas as if they were link prediction problems. It compares two users based on three criteria: similarity of tweets, followed ids, and followee tweets. These similarities are determined throughout the user profile extraction process. Tran et al. (2018) presented the hashtag recommendation technique, which significantly improves the performance of hashtag recommendation systems based on content analysis of tweets, user characteristics, and currently popular Twitter hashtags.

Sobhani et al. (2015) pioneered a novel approach to argument labelling. News comments were categorized according to the NMF-retrieved topics. Following that, the clusters were named using the top keywords for each cluster.

Zappavigna (2014) examined ambient affiliation in order to ascertain how a Twitter user enacts relational identities while executing discourse fellowships.

III. PROPOSED METHOD

People are more willing to share their skills on the internet as a result of SM (Social Media). They may also communicate with people from all around the world via it. It has developed into a wonderful area to experiment with new attacks with all its focus. The volume, velocity, and diversity of user data (e.g., user-produced data) in OSNs have increased dramatically. As a consequence, novel approaches for collecting and interpreting such large amounts of data have been studied. SBs have proved beneficial in automating analytical services and enhancing the quality of client care. On the other hand, RD has suffered real-world consequences as a result of false news (i.e., disinformation). As a consequence, identifying and removing RD (Rumor Detection) from OSNs is critical. The overwhelming majority of known approaches for RD identification are quantitative in nature.

Since a result, analytical precision is poor when utilizing SBs, as they are capable of readily imitating these traits. HybridRD, a novel approach for identifying RDs based on hybrid learning automata, was suggested. The algorithms KNN, SVM, Naive Bayes, Random Forest, and Decision Tree were used as pre-processors for the machine learning models. The authors describe a unique approach for identifying tweets created by actual persons or bots that combines behavior detection with the EDSMR-RD algorithm. The first simulation results are really encouraging. Experiments on the Kaggle dataset reveal that this strategy is capable of outperforming state-of-the-art bot detection systems.

3.1 Architecture Diagram

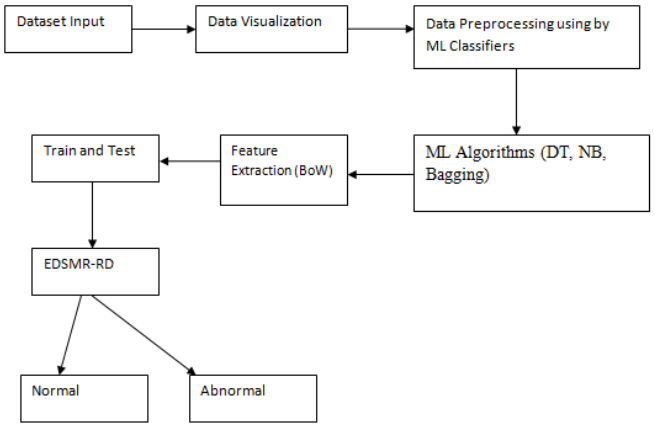


Figure 1: HybridRD System Model

3.2 Rumor Detection

This method has introduced the ML technique for efficient RD detection using the EDSMR-RD algorithm to overcome the restrictions of methods. This method generates Twitter data set from the kaggle.com website. The proposed architecture is shown in figure 1.

3.3 Datasets:

Datasets are collected from kaggle.com. All of the entries in this data collection are unique instances of data that are organized in a certain way. There are certain data structure kinds that are associated with the data set entries. Rumor and ham are the two types of Twitter rumor detection data sets. In ham radio communications, it might be difficult to tell whether a signal is genuine or not (i.e., hard to know). Since this might be mistaken for rumor, in some cases it is. The Kaggle.com website provided the data sets for free.

Dataset Details:

- Types of accounts:

i) Legitimate user
ii) Rumor user 1 (political)
iii) Rumor User 2 (product promotion)
iv) Rumor User 3 (eCommerce ads)

3.4 Exploration and Analyzing the Data:

This is an essential stage for a research data analyst in comprehending and investigating what is going on in the data collection. At this point, the primary features of any data may be evaluated. As a result, analyzing relevant patterns and features for big and scattered datasets is critical.

3.5 Data Visualization

For visualizing the data counts of user accounts concerning each group. This approach plotted the histogram of the account type of each group against the number of accounts. The plotting provides insights into the strength of accounts present in each subsequent twitter group. Users denote the legitimate users, and Rumor-1, Rumor-2, and Rumor-3 are the three automated rumor groups.

3.6 Data Pre-processing:

Stop words and stemming are the major components of this technique. To eliminate the stop words, stemming was used, and the verbs and adverbs of the views were extracted. There is a stop word removal procedure in which numbers, punctuation, and words that are not necessary in language processing are deleted. derived terms (smoothly, "smoother" to the base or root word is "smooth") are reduced to their base form in stemming.

ML pre-processing involves transforming raw data into a form that is easier to comprehend. For further implementation study, this technique processed the dataset after its presentation. Once the data has been pre-processed, it is cleaned and normalized using the various machine learning classifiers. Prior to analyzing, pre-processing removes noise from the data, inserts text, selects functions and normalizes cleaning procedures to ensure that the final product is consistent with the original data. ML models may benefit from pre-processing as well. Whitespace, punctuation, hyperlinks, and other data set noise may be removed using Text Cleaning. It's common practice to convert lowercase letters, erase dotted and white space as well as numbers. There's also word stream and lemmatization. Text documents are prepared for NLP events via the standardization process. Classifiers such as KNN, SVM, Nave Bayes, Random Forest, and Decision Tree were utilized to pre-process the data.

3.7 Extraction of Features:

The preparation of text data for machine learning (ML) is critical. Removal of words from text data necessitates a high degree of caution. It is called tokenization, and it is used to refer to this procedure. When using ML, it must be transformed into numbers since it cannot handle language directly The science-study tokenization and subsequent functional extraction are hence legitimate. BoW-ECM technology was developed based on how many word events occurred. It is common practice for extraction layers to convert words to "int" since most algorithms need numeric values (int or float).

Research has shown that bigram models give better perplexity with tolerable computing load than higher n-gram models, resulting in computational expense and over fitting. Bigrams are formed by concatenating neighboring words in the text; hence this model relies on them. Using predetermined lexicons, bigrams are tested against each other in order to reduce the amount of input required. Using bigrams instead of unigrams helps enhance text classification since a few words have various polarity meanings when evaluated as bigrams. As an example, "not good" and "not cool" will have distinct polarity meanings when they are created as unigrams. The Sentiment sensitive thesaurus identifies the word's sentiment.

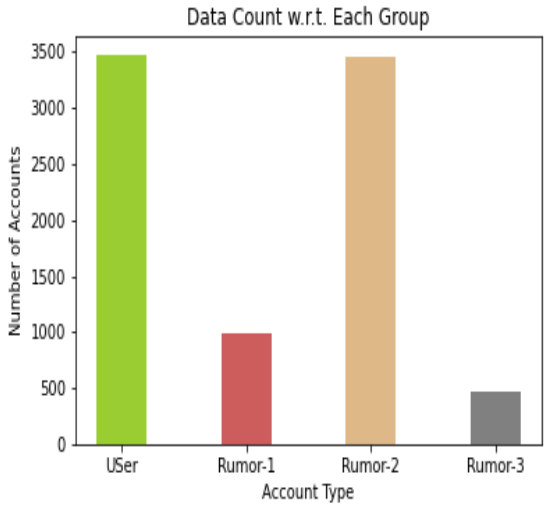


Figure 2: Comparison chart for Data Count for Each Group.

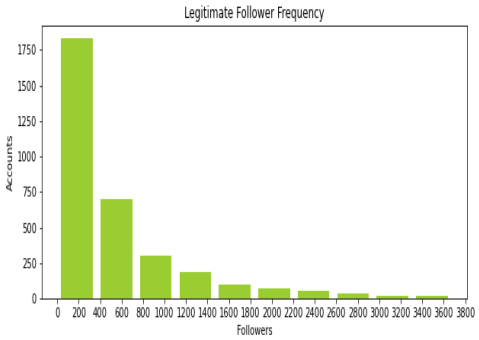
The Various types of accounts are represented in the graphical view, as shown in figure 3.2. The various accounts are user account and Rumor1, Rumor2, Rumor3. In X-axis denotes the Account type, and Y-Axis indicates the Number of Accounts.

3.8 Training Classifiers

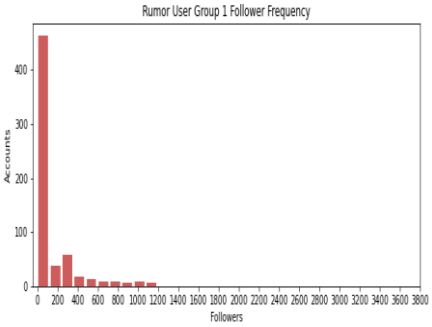
This approach offers a generic framework for detecting Twitter rumors and analyzing user behavior. A classification problem entails deciding on an input vector of data. To learn from examples of each class, the system uses a supervised learning method. The method should produce meaningful output for inputs that were not encountered during learning, known as a generalization.

3.9 Behavior Detection:

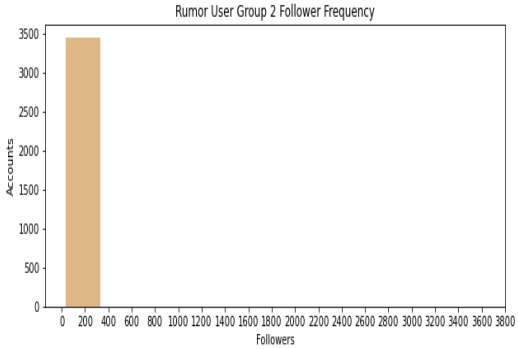
With the help of behavioral similarities analysis, This approach aims to find out how different groups of rumor accounts are similar or different from legitimate accounts groups.



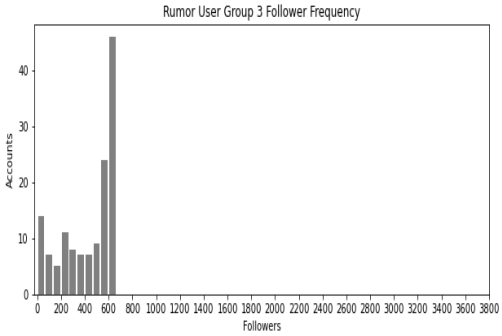
(a)



(b)



(c)



(d) Figure 3: Follower Frequency

The various user accounts are different levels for the follower frequency represented in Figure 3. In that, a, b, c, d

denotes the Legitimate, Rumor user group 1, 2, 3, respectively. In X-axis indicates the followers, and Y-axis represents the Accounts count.

3. 10 EDSMR-RD: DETECTION ALGORITHM

The generator T generates False rumor samples, and the discriminator TP cannot distinguish between the false samples and the real samples as input for the network classifier. ML classifiers are used to label real rumor samples and the hybirdRD analyses this label to identify rumor behavior.

Algorithm: 3.1: EDSMR-RD

Input: Set of users *P*= {p_1, .p_n} in twitter, T: Number of timeslots T_f; Threshold value, e; Reward parameter

Output: T: a set of trust values for all legitimate users with a list of legitimate users, S_b: a set of RD

Assumptions: Let LA={la_1,la_2,….la_n} be set of LA, where la_i represents hybrid learning automata for each participant.

Begin

1: Sb=$\emptyset, \beta = \emptyset, T = \emptyset$
2: EDSMR is activated for each users p_i
3: for each users $p_i \in$P do
4: for t=1,2,….T do
5: Compute trust value of p_i ($Tp_i(t)$)
6: Compute action probability value $p_r(t)=1-T_{pi}(t)$
7: if (TPi(t)<Tf) then
8: Concatenation of set β with a string 1 and β is updated with concatenated values
9: else
10: Concatenation of set β with a string 0 and β is updated with concatenated values
11: end if
12: end for
13: $\beta = \emptyset$
14: end for
15: return T with a list of legitimate users and S_b

Using the classifier result, trust value, and strength value, the weighted average (P) is calculated. During training, the weights v, w, and z are calculated. When a message is flagged as rumor, a rumor template is automatically generated. Incoming rumor messages are compared to an existing database of rumor message templates in order to determine whether they are a new sort of rumor. A rumor template match will cause the message to be filtered. In this case, the rumor message is tokenized, resulting in a rumor template being created. Only messages that meet the threshold (Th) are allowed to be posted; otherwise, they are rejected. The trust ratings between users are also updated on a regular basis. If the user who sent the message is closely linked to the recipient of the message, the trust value of the user will be lowered once the message is classed as rumor using the weighted average. If the sender and recipient aren't connected in any way and the message is rumor, the trustworthiness of the most trusted individual in the chain is lowered. Senders and messages are stored for a certain amount of time. A judgment is made on whether or not the trustworthiness of the communications can be enhanced depending on the class of messages after that time period.

3.11 Analyzing Behavior of Twitter Rumor Accounts

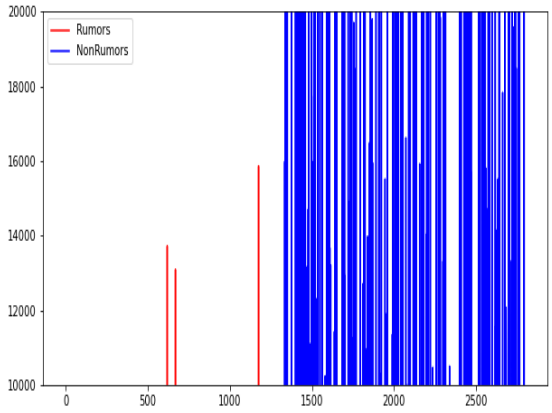


Figure 4: Twitter Rumor Account Count

The dataset has detected rumor account details as proposed, and the existing method is shown in figure 4. The dark blue is demonstrated as the proposed, and light blue is denoted as an existing method. The X-axis represents the rumor account detection, and the Y-axis indicates the user's account number.

IV. EXPERIMENTAL RESULTS

Datasets are collected from kaggle website as twitter dataset. The python programming language was used to create the HybridRD. It was necessary to gather data using several performance criteria such as accuracy and f-measurement. In this study, the results were compared to 20% to 40% of rumor messages. According to the findings, this model was well-executed intuitively and has a classification accuracy of better than 99 percent. Using this method, you'll get the most accurate score possible. With proper training, this model was able to reach an accuracy of more than 98%. HybridRD, on

the other hand, has a 99 percent accuracy and recall rate for Naive ayes. Additionally, the F-Measure of RD is around 78%, while the score of HybridRD is up to 97.9% with properly adjusted training.

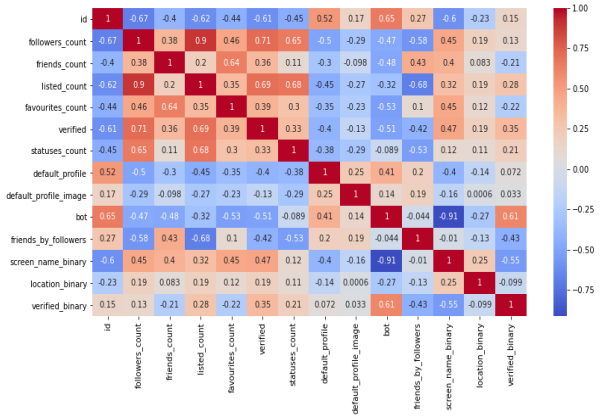


Figure 5: Confusion Matrix

The proposed solution has achieved 99.9% accuracy; the confusion matrix is calculated and displayed in figure 5.

4.1 Evaluation Metrics

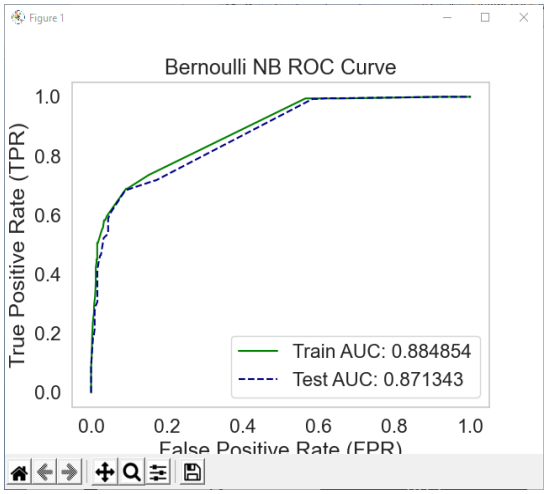
To evaluate the effectiveness of this proposed HybridRD approach

- **True Positive(TP):** the quantity of SBs correctly recognized;
- **True Negative (TN):** the number of malicious accounts correctly identified;
- **False Positive (FP):** the number of malicious accounts incorrectly recognized as SB;
- **False Negative (FN):** the quantity of SB erroneously recognized as human accounts.

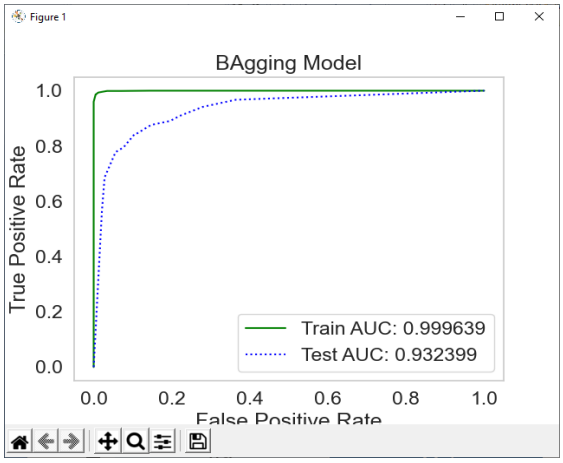
This approach evaluates the classifiers' performance for each test set using the following common evaluation metrics:

- Precision, the ratio of predicted positive points, i.e., Twitter rumors, that are indeed real positives: $\frac{TP}{TP+FP}$;
- Recall (also known as sensitivity) the ratio of real positive points that are indeed predicted as positives: $\frac{TP}{TP+FN}$;
- Specificity, the ratio of real negative points, i.e., malicious accounts, that are correctly identified as negative: $\frac{TN}{TN+FP}$;
- Accuracy, the ratio of correctly classified users (both positives and negatives) among all the users; $\frac{TP+TN}{TP+TN+FP+FN}$;
- F-measure, the harmonic mean of precision and recall: $2 * \frac{Precision.Recall}{Precision+Recall}$

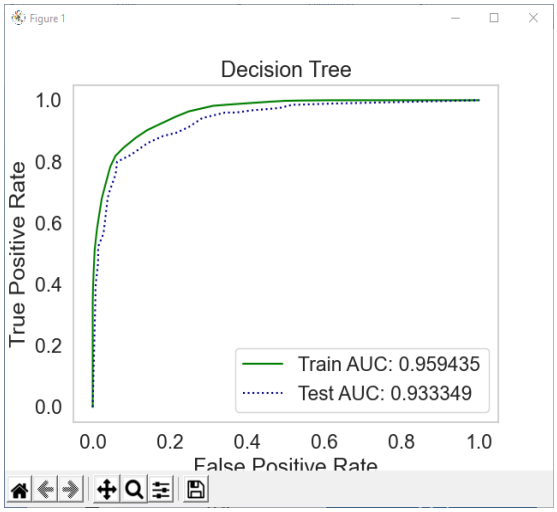
The ROC curve is a graphical illustration of a binary classifier system's analytical capability. The false-positive ratio is the likelihood of incorrectly rejecting the null hypothesis when an actual positive is tested. Because the goal is solely to detect fraudulent human accounts, the corpus is cleansed of any rumor accounts. Even after the screening, there may be a few false human and rumor accounts in the corpus. An additional 15,000 articles were manually generated to seem like they were created by people rather than machines and included in the corpus for the study reasons.



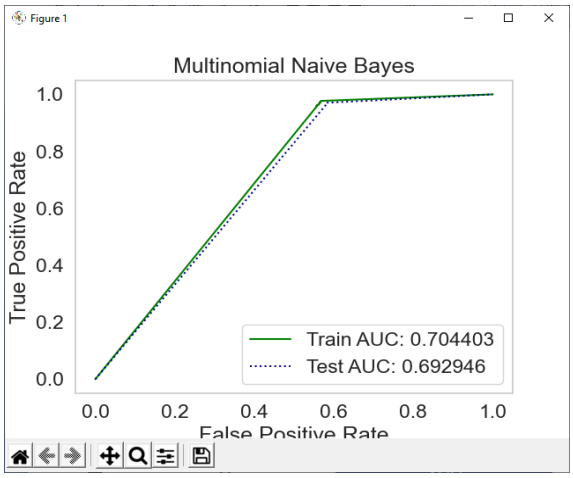
(i)



(j)



(k)



(l)

Figure 6: ROC Value in various Algorithms

The ROC value has been calculated with training, and the Testing Accuracy level is measured and shown as I, j, k, l.

Table 1: Accuracy Comparison table

Sno	Algorithm	Training	Testing
1	BNB	88	87
2	DT	95	93
3	Bagging Model	99	93
4	MNB	70	69
5	EDSMR-RD	98	97

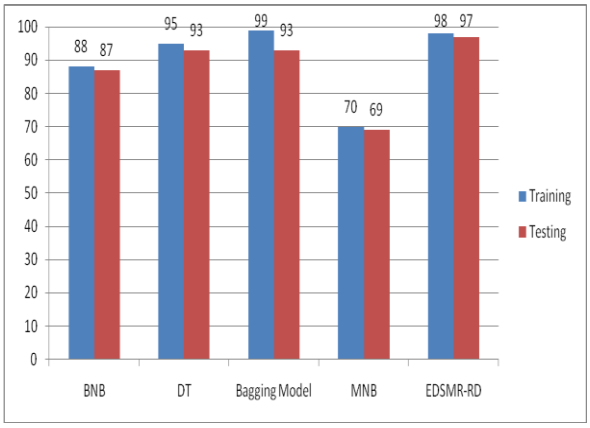


Figure 7: Accuracy Comparison chart

The proposed method has archived the highest accuracy is represented the figure 7.

The proposed EDSMR-RD method compared with various algorithms like DT, BNB, MNB, B, XGB algorithms, and EDSMR-RD has achieved 97.9% as a proposed method.

V. CONCLUSIONS

Twitter users and SBs' profiles, prior tweets, and social diagrams were utilised to identify spammers in the early stages of the process. spammers may swiftly penetrate OSN and disseminate damaging content, requiring the deployment of advanced methods such as machine learning and deep learning to counteract rumor. In light of the dangers posed by hazardous data, this study developed an ML-based rumor detection system that can successfully remove Twitter rumor. A hybrid RD model proposed in this paper is based on word embeddings to detect Twitter rumors from real accounts. There's no need to know anything about the target account's profile, friends list, or prior activity to use this method. According to this research, to use word embeddings in an ML model to detect tweet-only-dependent rumors without the need for significant feature engineering. The results of the initial simulations are very encouraging. These findings on the open-source training data set kaggle.com to show that this method is competitive with intensive human feature engineering. This advantage enables the rumor detection system to develop and deploy quickly and easily. The accuracy level has been reached at 97.9%.

REFERENCES

[1] Corcoglioniti, F, Nechaev, Y, Giuliano, C, Zanoli, R, "Twitter user recommendation for gaining followers", In: Ghidini C., Magnini B., Passerini A., Traverso P. (eds) AI*IA 2018 – Advances in Artificial Intelligence. AI*IA

2018. Lecture Notes in Computer Science, vol. 11298. Springer, Cham,(2018).

[2] Cao, D, He, X, Nie, L, Wei, X, Hu, X, Wu, S & Chua, TS, "Cross- platform app recommendation by jointly modeling ratings and texts", ACM Transactions on Information Systems, vol. 35, no. 4, pp. 1-27,(2017).

[3] Faulkner, A, "Automated classification of stance in student essays: An approach using stance target information and the Wikipedia link- based measure", In: Proceedings of the Twenty-Seventh International Flairs Conference, 174-179,(2014).

[4] Guidi, B., & Michienzi, A,"Users and Bots behaviour analysis in Blockchain Social Media". 2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS). doi:10.1109/snams52053.2020.93365.

[5] Hercig, T, Krejzl, P, Hourová, B, Steinberger, J & Lenc, L, "Detecting stance in czech news commentaries", In: Proceedings of the 17th ITAT: CEUR Workshop, pp. 176 – 180,(2015).

[6] Heredia, B., Prusa, J. D., & Khoshgoftaar, T. M,"The Impact of Malicious Accounts on Political Tweet Sentiment". 2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC). doi:10.1109/cic.2018.00035.

[7] Jiménez-Bravo, DM, De Paz, JF & Villarrubia, G, "Twitter's experts recommendation system based on user content. In: Rodríguez S. *et al*. (eds) Distributed Computing and Artificial Intelligence", Special Sessions, 15th International Conference DCAI 2018. Advances in Intelligent Systems and Computing, vol. 801. Springer, Cham,(2019).

[8] Rodríguez, FM, Torres, LM & Garza, SE, "Followee recommendation in Twitter using fuzzy link prediction", Expert Systems, vol. 33, no. 4, pp. 349-361,(2016).

[9] Sobhani, P, Inkpen, D & Matwin, S, "From argumentation mining to distance classification", In: Proceedings of the Workshop on Argumentation Mining, pp. 67–77,(2015).

[10] Zappavigna, M, "Enacting identity in microblogging through ambient affiliation", Discourse & Communication,vol. 8, pp. 209–228,(2014).