International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Unsupervised Misinformation Detection Model using Incremental K-Means Algorithm

Yashoda Barve[1], Jatinderkumar R. Saini*[2]

Submitted: 10/09/2022 **Accepted**: 20/12/2022

Abstract: The state-of-the-art misinformation detection techniques mainly focused on supervised learning approach, however, it requires a huge amount of labeled dataset resulting into manual efforts and delays in detecting misinformation. Thus, an unsupervised approach to misinformation detection is in demand. The researchers with unsupervised misinformation detection show average performance as they lack in generating important textual and user-specific features. Further, since the data in the real world is time- sensitive, a large amount of data is generated over a period of time and the models need to adapt to this newly arriving chunk of data. To tackle the above problems, the authors have proposed a first-of-its-kind unsupervised misinformation detection model using an incremental learning approach that can handle newly arriving data without needing to label the data. To evaluate the model's performance, the authors have used various metrics like silhouette score, purity, and importance of various features in cluster formation. The model showed a purity score of 0.92 % and average silhouette score of 0.57%.

Keywords: *Clustering, Incremental Learning, K-Mean, Misinformation Detection, Unsupervised Learning*

1. Introduction

Internet has evolved into individuals' primary source of information. It has gained popularity throughout the community as a result of its ease, viability, unlimited access, and affordable price [1]. Especially, online resources for health and medical information are abundant in the healthcare industry. It was found that while individuals browse the web for information about illnesses, infections, and their symptoms, doctors chose the web as a helpful information resource for medical practice, training as well as decision assistance. However, the accuracy and quality of the content that is made available online are not guaranteed. The credibility and accuracy of information are significant because they could cause misinformation to spread like wildfire. This pervasive misinformation has negative effects on every aspect of society, including individuals, businesses, the government, and the health system. The consequences of the false information are dire and could result in extinction [2], [3].

The researchers have developed models to detect misinformation using supervised learning techniques and have achieved great success. However, these models require a large amount of labeled data which involves manual efforts, and expert opinions and is a time- consuming process [4]. Thus, researchers have put in efforts to devise unsupervised approach-based models in the literature. For example, the authors of [5] proposed a method based on text content for unsupervised false news identification. This system ignored user information for news publications and used tensors to classify news according to its hidden content. The authors in [6] used transfer learning and semantic similarity technologies to determine whether the news was true, but they ignored the social context data. The authors of [7] proposed a method based on graphics that exploits user behavior to identify unsupervised bogus news, although this method ignores the textual substance of the news. Thus, it can be noted that considering textual as well as user-specific features plays a crucial role in detecting misinformation with an unsupervised machine learning approach to improve the performance of the model. Another important concern is to deal with incremental data appearing in chunks at different intervals of time. Because the information in the real world is time-sensitive and newly emerging data must be adapted accordingly by the model, efforts are required to furnish the newly arriving data [3]. In misinformation detection using the supervised approach, this newly arriving data needs to be labeled before building the model. This requires a significant amount of human effort, expert opinions, and time resulting into delay in detecting misinformation. Therefore, an unsupervised approach with incremental learning can reduce the data annotation efforts and also adapt to newly arriving or changing data.

To tackle above discussed problems, the authors in this research have proposed an unsupervised machine learning approach of misinformation detection with incremental learning. In this research, the authors have devised thirty textual and user-specific features and developed a module to compute feature importance using a k-means clustering algorithm and random forest classifier. Further, an unsupervised learning model is built using k-means forming two clusters. These clusters are updated in an incremental fashion using an incremental learning approach. This is achieved by dividing data into five different iterations and using Euclidean distance to map the newly generated cluster in the iteration with the original existing cluster. Following are the research contributions:

1. To generate novel features that would improve the performance of the misinformation detection model with an unsupervised

[1] *Suryadatta College of Management Information Research & Technology, Pune, India*
ORCID ID: 0000-0003-3422-2464
[2] *Symbiosis Institute of Computer Studies and Research, Symbiosis International (Deemed University), Pune, India*
ORCID ID: 0000-0001-5205-5263
** Corresponding Author Email: saini_expert@yahoo.com*

approach.
2. To develop a methodology to detect misinformation with an unsupervised machine learning technique to handle unlabeled data and also deal with newly arriving data over a period of time.
3. To evaluate the performance of the proposed model.

2. Related Work

In this section, the authors have studied three related topics: misinformation detection techniques using unsupervised learning, different types of features, and incremental learning approach for misinformation detection.

2.1. Misinformation detection using unsupervised learning

The misinformation detection is a problem of classification, different machine learning classifiers like Logistic Regression, Random Forest, and Support Vector Machine can be applied to classify data into true or false. Recent research in misinformation detection related to COVID-19 focuses more on the supervised approach [8]–[11]. Although these supervised methods have produced encouraging results, a crucial restriction prevents them from being fully utilized: they need a previously labeled dataset to build the model [12]. In real life scenario, since the huge amount of data evolves over a period of time it becomes a tedious task to annotate this data. Therefore, efforts are required to propose unsupervised learning-based models that would work without labeled data [13].

According to the literature, researchers are keen on developing unsupervised learning-based models. In research [14], authors have used an unsupervised learning approach with K-means, Non-negative Matrix Factorization (NMF), and Latent Discriminative Analysis (LDsA) algorithms for clustering. The experimental results showed poor performance of the K- means algorithm, this may be due to failure in extracting relevant features from the dataset as the authors have considered only TF-IDF features. LDsA algorithm showed better performance with an F1-score of 0.70. In another research, a temporal ensemble learning-based architecture with a convolution neural network was devised with a small annotated train dataset. Further, it combines the results of all prior epochs into a collective prediction that is anticipated to be more accurate than the unannotated inputs of unidentified labels. As a result, the labels deduced in this manner serve as an unsupervised training target [13]. In another research [4], authors have combined a Bi-directional GRU layer and self-attention layer with an autoencoder to detect fake news with an unsupervised approach. In a research, the semi- supervised expectation-maximization algorithm was used to detect rumors in Arabic tweets with content and user-based features. The results showed better performance than the Gaussian Naïve Bayes classifier [15]. Another study to detect political rumors on Twitter was based on an unsupervised clustering approach. In this study, similar clusters were combined into one cluster based on cosine similarity after tweets with the same URL link were grouped. The number of extreme users present in the cluster, along with a few rules, were used to detect rumors. The newly developed clustering process showed a recall score of 0.85 [16]. Thus, there are hardly a few researchers with unsupervised machine learning approaches using clustering techniques to detect false information. These methods have shown average performance and this is due to a lack of relevant feature extraction. Therefore, there is a pressing need to develop a cluster-based unsupervised machine learning model with novel features to detect misinformation.

2.2. Feature Extraction

The misinformation detection using the machine learning approach requires the generation of features from the data to train the model. The dominant categories of such features are textual or content-based, user- specific features, propagation-based features, and temporal and structural features [17]. In research [18], authors have engineered features like Part-of-Speech (POS) tagging, psycholinguistic features, and readability features to detect COVID-19-related misinformation. The features like Bag-of-Words (BoW), Term- Frequency-Inverse Document Frequency (TF-IDF) [10], [11], [19]–[22], and Word2Vec are widely used to develop the model fake news detection [23]. Also, features such as linguistic, sentimental, TF-IDF, and medical words are extracted to detect healthcare-related misinformation. In another research, authors have used similarity features to map text title with the body of the text, or to perform fact-checking [24]–[26]. Extracting relevant features from the dataset improves the model performance. Therefore, in this research authors have generated 30 different novel features in two categories viz. textual or content-based and user-specific features based on the ReCOVery dataset. The list of all the useful features along with the description is displayed in Table.

2.3. Misinformation Detection using Incremental Learning

The traditional issue of detecting incorrect information or misinformation only takes into account the static nature of the data. Using methods like ensemble learning or incremental learning, researchers have identified these issues in the literature [27], [28]. For example, authors have used ensemble machine learning techniques to detect fake news using a convolution neural network [13]. In another research, authors have used classifiers like the random forest, extra tree, and decision tree along with the bagging approach to aggregate the outputs of the classifiers [29]. However, ensemble learning techniques are more suitable for sudden changes in data [25]. Whereas a model is not retrained on the entire dataset; instead, incremental learning (IL) algorithms iteratively learn knowledge from freshly arriving data without forgetting previously obtained knowledge. Thus, it is believed that the incremental learning strategy performs better in terms of efficiency and is better suited to handle gradual concept drifts [30], [31]. In the literature, authors have devised a novel incremental learning-based veracity scanning model to detect and fact-check healthcare misinformation [25]. However, neither of these models have developed clustering-based unsupervised models using an incremental learning approach. Thus, there is a need to build an adaptive model that would combine an unsupervised approach with incremental learning to deal with newly arriving data to keep the data up-to-date and handle unlabeled data.

2.4. Potential Research Gap

1. According to the literature, there are only a few models related to the unsupervised machine learning approach showing average performance with k- means and alike clustering techniques. This is due to the lack of relevant feature extraction from the dataset. Thus, there is a pressing need to develop models that would extract relevant novel features and detect misinformation using an unsupervised machine learning approach.
2. Although researchers have used incremental learning for misinformation detection using machine learning, it mainly

focuses on supervised machine learning models and does not deal with unsupervised models. Thus, there is a need to build an incremental model that is able to handle unsupervised data.

3. Methodology

3.1. Dataset Description

An Unsupervised Misinformation Detection (UMD) model detects misinformation in healthcare web URLs using an incremental k-means clustering algorithm. To build the model authors have used ReCOVery[32] dataset consisting of COVID-19-related URLs. The dataset was generated in August 2020 soon after the COVID-19 pandemic with an aim to assist researchers in finding credible and non-credible web URLs. ReCOVery dataset consists of multi-modal features like news ID, news URL, publisher, text, country, political bias, and news credibility. The dataset size consists of 2029 URLs of credible and non-credible categories.

3.2. Feature Extraction

In this research, authors have extracted 30 different textual or content-based features and user-specific features from the ReCOVery dataset to build an unsupervised misinformation detection model. Table I lists the features extracted along with the description. To extract sentiment-related features viz. number and percentage of positive, negative, and neutral words, the total number of sentimental words and to compute sentence polarity, authors have used incremental sentimental Bag-of-Words (BoW) generated in the author's previous work. To extract topical features authors have used Latent Dirichlet Allocation Method (LDA). LDA is commonly used to identify various topics from the document and the corresponding number of words belonging to a particular topic. The authors have identified 10 different topics from the articles of the ReCOVery dataset. Another important feature is the readability index. Readability can help to identify the complexity of the text. Therefore, the authors have used Automated Readability Index (ARI) formula to compute the complexity of the text. This feature contributes efficiently to understanding the writing style and assists in identifying the stance of the article. Authors have also computed 4 features of Part-Of-Speech (POS) tags. In user-specific features, authors have considered the country as a feature to find the country-wise percentage of misinformation spread. The reliability feature specifies the credibility of URLs. The reputation value is computed by checking the publisher and its reliability. This is achieved by computing the sum of the reliability value of articles per publisher. If the sum value of true or reliable articles is greater than false, then reliability is assigned as 1 otherwise it is assigned to 0. The dataset contains a field named political bias which is also considered a feature to identify misinformation.

3.3. Finding Important Features

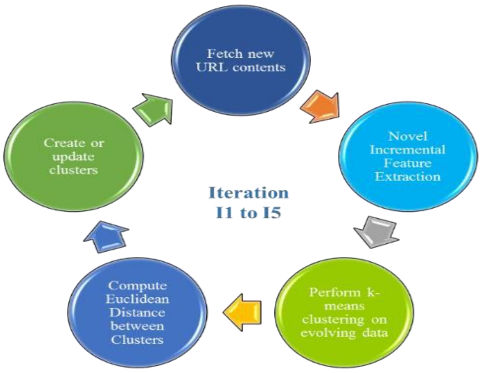
Upon extracting the required features authors have identified important features using a combination of k- means clustering and random forest classifier techniques. In this, the authors have considered the dataset of 2029 URLs and extracted 30 different features as discussed above, and performed k-means clustering with the value of k=2, as detecting misinformation results in the binary classification of either true or false values. Further, two clusters are formed with their respective labels as target variables. Then authors mapped the unsupervised approach with supervised learning by passing these clusters to Random Forest Classifier (RFC) which ultimately resulted in individual feature scores. These individual feature scores are then sorted to find the most important features.

3.4. Model Building

The authors applied Elbow method to identify the ideal number of clusters for ReCOVery dataset. Although it resulted into value of k=3, authors had to force fit the model to have k=2, since misinformation classification results into either true or false. Further, the model building phase starts with fetching of URLs from ReCOVery dataset. Initially, authors have extracted Term Frequency-Inverse Document Frequency (TF- IDF) features for clustering. However, it didn’t not show promising results and inter-cluster distance was too less. Next, for each URL 30 different features are extracted as discussed above and k-means clustering technique is executed. Further, Inter-quartile range is computed to remove the outliers. This resulted into clean data of 1181 URLs showing well-formed clusters. To evaluate the performance of model on incremental data, authors have splitted the data into train and test. Thus, 181 URLs are used for training and 1000 URLs are used for testing. The model was trained using 181 URLs. The test dataset is splitted up into five different iterations of 200 URLs each. In the first iteration I1 the features are extracted and model is built using k-means forming two clusters C1 and C2 respectively. In second iteration I2, the feature values are updated to get latest data, and the clusters C1 and C2 are updated by computing Euclidean distance between a URL from I2 and a random URLs picked from cluster C1 and C2. Based on the smallest difference, the URL from I2 is appended to the respective cluster. The same process is repeated for remaining iterations I3 to I5. Fig. 1 displays the model architecture.

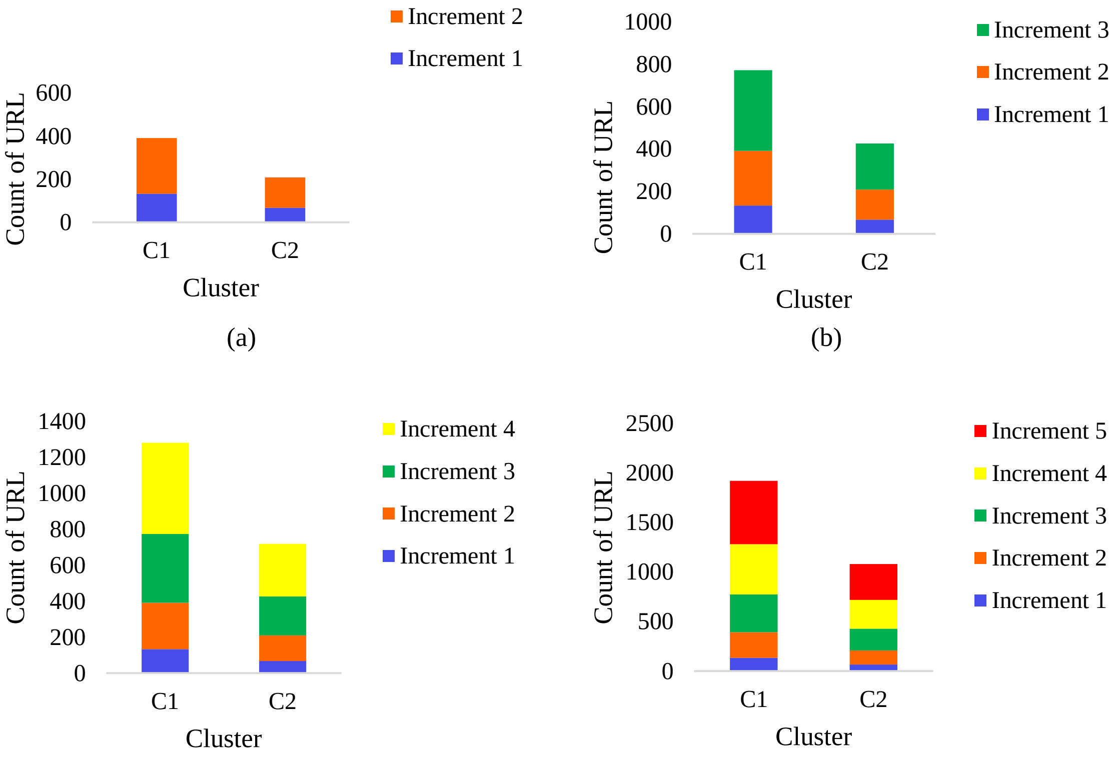
4. Experimental Results

This section elaborates on the experimental results and its analysis. Fig. 2 shows the clusters updated during each incremental phase. It can be seen that cluster C1 consist of more URL than cluster C2 during every increment. Thus total 638 URLs are part of cluster C1 while 362 URLs belong to cluster C2. Fig. 4 shows the publisher status. It can be seen that 51% of publishers share more false information thus labelled as bad, while 40% publishers spread more true information hence labelled as good, and 9% are average publishers sharing mix of true and false information. According to Fig.6 it can be found that cluster C2 has more bad and average publishers compared to C1. Thus it can be said that cluster C2 more comprises of false URLs. While cluster C1 has more good reputation status, depicting it contains more number of true URLs. Fig. 5. shows feature results based on ARI values. It can be observed that per iteration cluster C1 shows higher ARI value which means that cluster C1 contains more number of well written articles than cluster C2. Thus, it can be said that cluster C2 contains more false information spreading articles with lower ARI. From Fig. 3 it is clear that the percentage of misinformation in cluster C2 is high in every iteration than in cluster C1. Further, Fig.7 shows the country-wise percentage of misinformation with Iran having highest number of articles spread misinformation. Fig.13 shows the topical distribution among the clusters. It is observed that cluster 2 has high contribution of all the topics. However, Topic 3 has higher value in cluster C2 than in cluster C1. The Topic 3 contains words like ‘death’, ‘infection’, ‘COVID’, ‘outbreak’, etc. This also states that cluster C2 is more towards containing false URLs. Also, from Fig 8 and Fig.9. It is understood that the average positive count is more in cluster C1 and average negative count is higher in cluster C2. The silhouette score represent correctness of the cluster formation. The silhouette score near 1 means that the clusters are accurately formed. Fig. 10 shows the silhouette score based on various distance measures. It is observed that Euclidean, Squared Euclidean, Chebychev and Chi square measures has silhouette score near to 0.57 depicting average cluster formation. Same is the case with incremental learning as shown in the Fig.11. Fig.12 displays the word cloud of topic based features. This includes all the words from ten different topics identified by LDA technique. Fig.14 & Fig.15 displays the top ten important features identified using proposed technique of feature importance. It can be seen that in cluster C1 character count is found to be crucial with the highest score followed by ARI, topic 3, reputation, and positive count. Similarly, cluster C2 character count has the highest value followed by the positive count, negative count, ARI, and topic 3. Thus the findings state that cluster C1 contains a large number of true URLs while cluster C2 has more number of false URLs. Table II displays the time taken for each iteration using incremental k-means and simple k-means algorithms. It can be seen that the incremental learning model outperformed the simple k-means algorithm showing results less time of 12ms to 23ms for five different iterations whereas the simple k-means algorithm showed a high execution time of between 22ms to 37ms for five different iterations. The value of purity score for the proposed model is 0.92. According to the literature, authors from [1], [2] have used clustering technique for unsupervised misinformation detection, however they didn't receive promising results whereas the proposed unsupervised detection method showed better results in terms of purity with 0.92% and silhouette score of average 0.57%.



Fetch new URL contents
Novel Incremental Feature Extraction
Perform k-means clustering on evolving data
Compute Euclidean Distance between Clusters
Create or update clusters
Iteration I1 to I5

Figure 1. demonstrates the architecture of proposed model



(a) (b)

(c) (d)

Figure 2. Incremental updating of clusters in five iterations

Table 1. List of features used to detect misinformation using unsupervised learning approach

Sr. No.	*Category*	*Feature Name*	*Description*
1		#characters	Total number of characters
2		#words	Total number of words

3	Text or content based features	#sentences	Total number of sentences
4		ARI value	Readability index value
5		#sentimental words	Total number of sentimental words
6		#positive words	Total number of positive words
7		#negative words	Total number of negative words
8		#neutral words	Total number of neutral words
9		%positive words	Percentage of positive words
10		%negative words	Percentage of negative words
11		%neutral words	Percentage neutral words
12		sentence polarity	Sentiment based sentence polarity score
13		#nouns	Number of nouns
14		#pronouns	Number of Pronouns
15		#verbs	Number of Verbs
16		#adjectives	Number of adjectives
17		#Topic 1 words	Number of words in Topic 1
18		#Topic 2 words	Number of words in Topic 2
19		#Topic 3 words	Number of words in Topic 3
20		#Topic 4 words	Number of words in Topic 4
21		#Topic 5 words	Number of words in Topic 5
22		#Topic 6 words	Number of words in Topic 6
23		#Topic 7 words	Number of words in Topic 7
24		#Topic 8 words	Number of words in Topic 8
25		#Topic 9 words	Number of words in Topic 9
26		#Topic 10 words	Number of words in Topic 10
27	User-based Features	Country	Numeric value for 6 different countries
28		Reliability value	Boolean values for reliability
29		Reputation value	Boolean value for reliability
30		Political bias	Numeric value for 5 different political bias

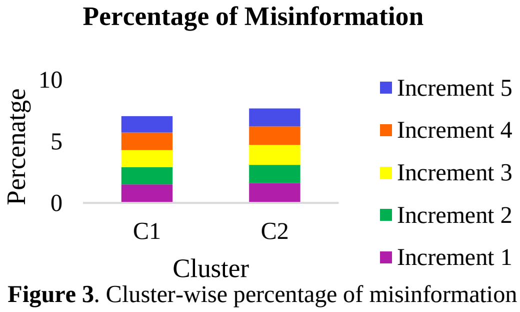


Figure 3. Cluster-wise percentage of misinformation

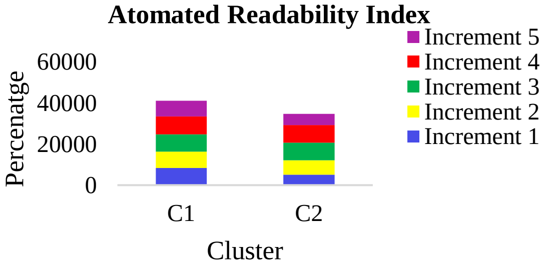


Figure 4. Publisher status analysis

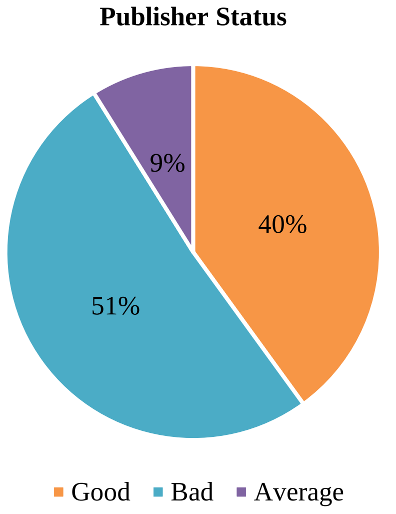


Figure 5. Cluster-wise ARI Analysis

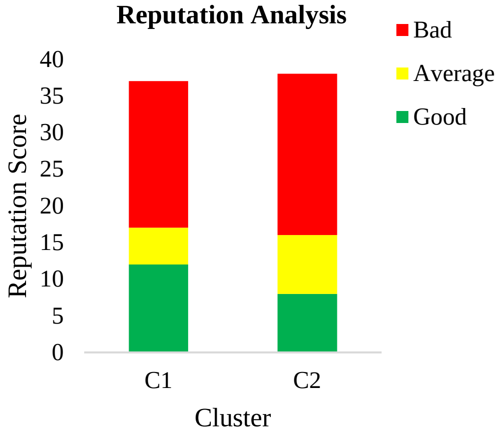


Figure 6. Cluster-wise reputation analysis

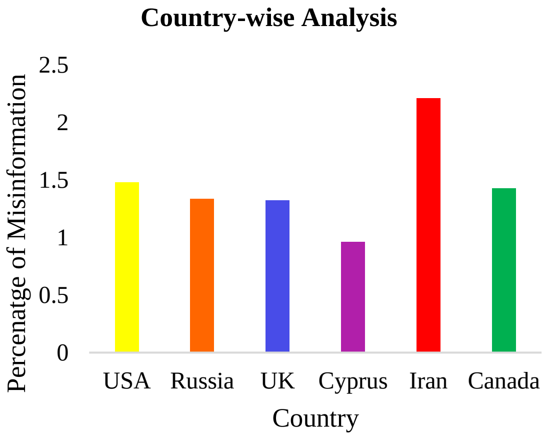


Figure 7. Country-wise misinformation detection

Figure 8. Cluster-wise negative sentiment analysis

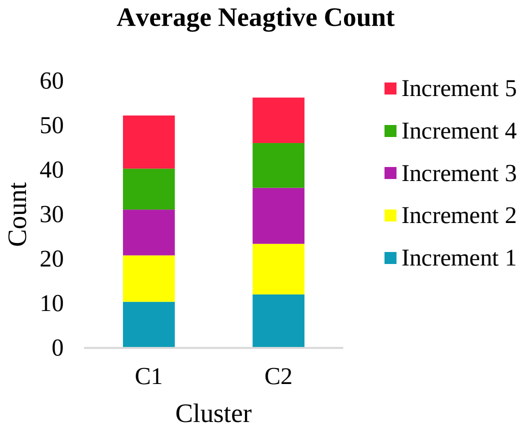


Figure 9. Cluster-wise positive sentiment analysis

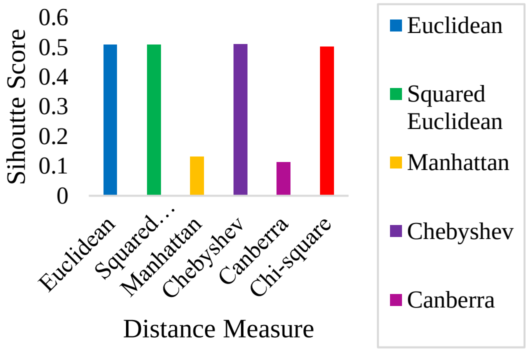


Figure 10. Silhouette score based on distance measures

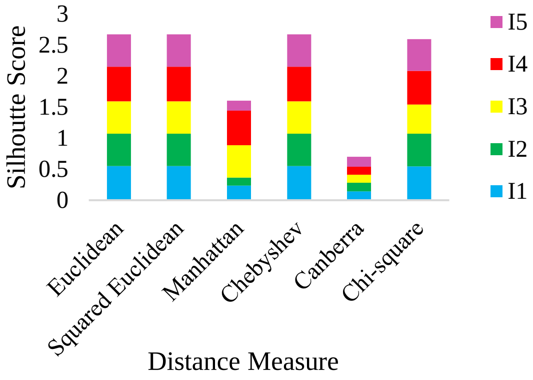


Figure 11. Silhouette score with incremental learning

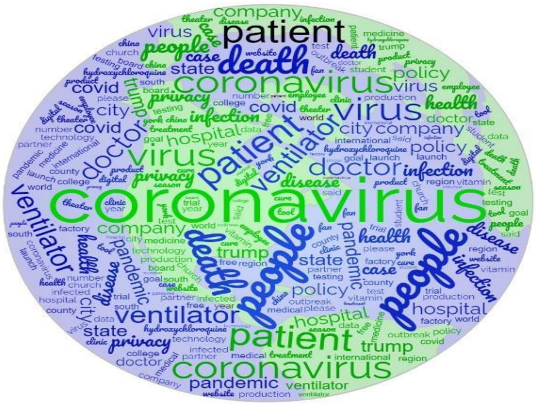


Figure 12. Word cloud of topical features

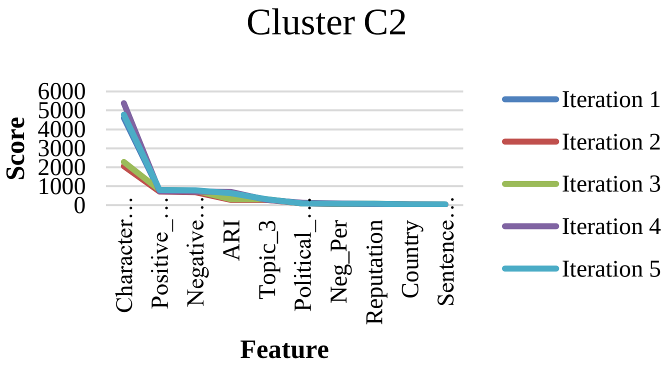


Figure 13. Cluster-wise topic analysis

Table 2. Execution time analysis using k-means and increment k-means approach

Iterations	*Incremental K-Means*	*K-means*
I1	12ms	22ms
I2	14ms	25ms
I3	17ms	29ms
I4	22ms	33ms
I5	23ms	37ms

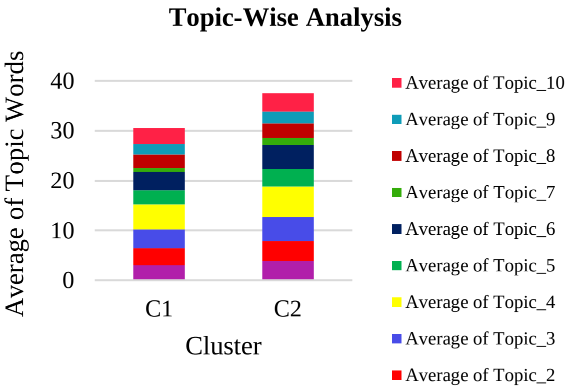


Figure 14. Cluster C1 top-ten important features

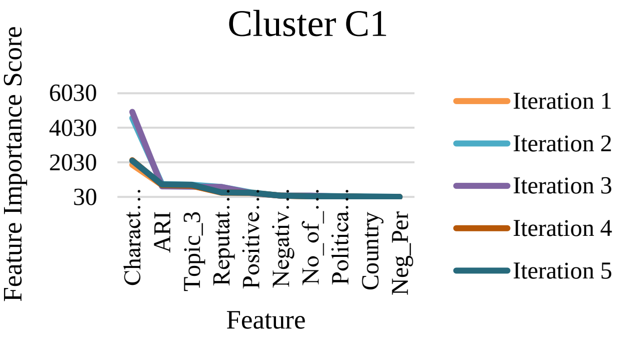


Figure 15. Cluster C2 top-ten important features

5. Conclusion

In this research authors have proposed a novel methodology of using the unsupervised and incremental learning-based approach for misinformation detection in the healthcare domain. Also, authors have identified novel features based on ReCOVery dataset and proposed a technique to compute feature importance using a combination of unsupervised and supervised machine learning approach. The experimental results showed average cluster formation with silhouette score of 0.57. It was observed that newly built user-specific features like reputation, political bias, country, and textual features like ARI and Topic 3 have contributed in the top ten features in cluster formation.

In the future author wants to develop a new clustering algorithm using an incremental learning approach to detect misinformation and compare the performance of the model with this research work.

References

[1] Z. Jahanbakhsh-Nagadeh, M.-R. Feizi-Derakhshi, and A. Sharifi, "A semi-supervised model for Persian rumor verification based on content information," *Multimed. Tools Appl.*, vol. 80, no. 28–29, pp. 35267–35295, 2021, doi: 10.1007/s11042-020-10077-3.

[2] Y. Barve and J. R. Saini, "Healthcare Misinformation Detection and Fact-Checking : A Novel Approach," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 10, pp. 295–303, 2021.

[3] Y. Barve, J. R. Saini, K. Pal, and K. Kotecha, "A Novel Evolving Sentimental Bag-of-Words Approach for Feature Extraction to Detect Misinformation," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 4, pp. 266–275, 2022, doi: 10.14569/IJACSA.2022.0130431.

[4] D. Li, H. Guo, Z. Wang, and Z. Zheng, "Unsupervised Fake News Detection Based on Autoencoder," *IEEE Access*, vol. 9, pp. 29356–29365, 2021, doi: 10.1109/ACCESS.2021.3058809.

[5] S. Hosseinimotlagh and E. E. Papalexakis, "Unsupervised content-based identification of fake news articles with tensor decomposition ensembles," *Proc. WSDM MIS2 Misinformation Misbehavior Min. Web Work.*, pp. 1–8, 2018, doi: 10.475/123.

[6] J. Gaglani, Y. Gandhi, S. Gogate, and A. Halbe, "Unsupervised WhatsApp Fake News Detection using Semantic Search," in *Proceedings of the International Conference on Intelligent Computing and Control Systems, ICICCS 2020*, 2020, pp. 285–289, doi: 10.1109/ICICCS48265.2020.9120902.

[7]

the detection of fake news articles," *Expert Syst. Appl.*, vol. 177, 2021, doi: 10.1016/j.eswa.2021.115002.

[14] S. G. Taskin, E. U. Kucuksille, and K. Topal, "Detection of Turkish Fake News in Twitter with Machine Learning Algorithms," *Arab. J. Sci. Eng.*, vol. 47, no. 2, pp. 2359–2379, 2022, doi: 10.1007/s13369-021-06223-0.

[15] S. M. Alzanin and A. M. Azmi, "Rumor detection in Arabic tweets using semi-supervised and unsupervised expectation–maximization," *Knowledge-Based Syst.*, vol. 185, 2019, doi: 10.1016/j.knosys.2019.104945.

[16] C. Chang, Y. Zhang, C. Szabo, and Q. Z. Sheng, "Extreme user and political rumor detection on twitter," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10086 LNAI, pp. 751–763, 2016, doi: 10.1007/978-3-319-49586-6_54.

[17] P. Meel and D. K. Vishwakarma, "Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities," *Expert Syst. Appl.*, vol. 153, 2020, doi: 10.1016/j.eswa.2019.112986.

[18] M. Chen, X. Chu, and K. P. Subbalakshmi, "MMCoVaR: Multimodal COVID-19 vaccine focused data repository for fake news detection and a baseline architecture for classification," in *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2021*, 2021, pp. 31–38, doi: 10.1145/3487351.3488346.

[19] K. Pogorelov, D. T. Schroeder, P. Filkuková, S. Brenner, and J. Langguth, *WICO Text: A Labeled Dataset of Conspiracy Theory and 5G-Corona Misinformation Tweets*, vol. 1, no. 1. Association for Computing Machinery, 2021.

[20] M. Mayank, S. Sharma, and R. Sharma, "DEAP-FAKED: Knowledge Graph based Approach for Fake News Detection," 2021, [Online]. Available: http://arxiv.org/abs/2107.10648.

[21] M. Isaakidou, E. Zoulias, and M. Diomidous, *Machine learning to identify fake news for COVID-19*. IOS Press, 2021.

[22] J. Ayoub, X. J. Yang, and F. Zhou, "Combat COVID-19 infodemic using explainable natural language processing models," *Inf. Process. Manag.*, vol. 58, no. 4, 2021, doi: 10.1016/j.ipm.2021.102569.

[23] K. Nath, P. Soni, Anjum, A. Ahuja, and R. Katarya, "Study of Fake News Detection using Machine Learning and Deep Learning Classification Methods," in *2021 6th International Conference on Recent Trends on Electronics, Information, Communication and Technology, RTEICT 2021*, 2021, pp. 434–438, doi: 10.1109/RTEICT52294.2021.9573583.

[24] Y. Zhao, J. Da, and J. Yan, "Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches," *Inf. Process. Manag.*, vol. 58, no. 1, 2021, doi: 10.1016/j.ipm.2020.102390.

[25] Y. Barve, J. R. Saini, K. Kotecha, and H. Gaikwad, "Detecting and Fact-checking Misinformation using 'Veracity Scanning Model,'" *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 2, pp. 201–209, 2022, doi: 10.14569/IJACSA.2022.0130225.

[26] I. Baris and Z. Boukhers, "ECOL: Early Detection of COVID Lies Using Content, Prior Knowledge and Source Information," *Commun. Comput. Inf. Sci.*, vol. 1402 CCIS, pp. 141–152, 2021, doi: 10.1007/978-3-030-73696-5_14.

[27] W. Zang, P. Zhang, C. Zhou, and L. Guo, "Comparative study between incremental and ensemble learning on data streams: Case study," *J. Big Data*, vol. 1, no. 1, pp. 1–16, 2014, doi: 10.1186/2196-1115-1-5.

[28] P. Ksieniewicz, P. Zyblewski, M. Choraś, R. Kozik, A. Giełczyk, and M. Woźniak, "Fake News Detection from Data Streams," *Proc. Int. Jt. Conf. Neural Networks*, 2020, doi: 10.1109/IJCNN48605.2020.9207498.

[29] S. Hakak, M. Alazab, S. Khan, T. R. Gadekallu, P. K. R. Maddikunta, and W. Z. Khan, "An ensemble machine learning approach through effective feature extraction to classify fake news," *Futur. Gener. Comput. Syst.*, vol. 117, pp. 47–58, 2021, doi: 10.1016/j.future.2020.11.022.

[30] A. Habib, M. Z. Asghar, A. Khan, A. Habib, and A. Khan, "False information detection in online content and its role in decision making: a systematic literature review," *Soc. Netw. Anal. Min.*, vol. 9, no. 1, 2019, doi: 10.1007/s13278-019-0595-5.

[31] A. Chefrour, "Incremental supervised learning: algorithms and applications in pattern recognition," *Evol. Intell.*, vol. 12, no. 2, pp. 97–112, 2019, doi: 10.1007/s12065-019-00203-y.

[32] X. Zhou, A. Mulay, E. Ferrara, and R. Zafarani, "ReCOVery: A Multimodal Repository for COVID-19 News Credibility Research," in *International Conference on Information and Knowledge Management, Proceedings*, 2020, pp. 3205–3212, doi: 10.1145/3340531.3412880.