SEUSL

Sri Lankan Journal of Technology

Research

SLJoT
2022 Sp Issue 32 - 38
ISSN 2773-6970

Detection of Fake News on Twitter Regarding COVID-19: An Analysis of Machine Learning Algorithms with n-gram Modeling

R. Yasotha[1], H. Ahamed Rikas[2],M.I. Fathima Nuska[2], M.A. Fathima Shanas[2] and A.F. Sharfana[3*]

[1]Dept. of Physical Science, University of Vavuniya., Sri Lanka
[2]Department of Information and Communication Technology, University of Vavuniya, Sri Lanka
[3]Department of Information and Communication Technology, South Eastern University of Sri Lanka

[*]Corresponding Author: sharfana.atham@seu.ac.lk ‖ ORCID: 0000-0002-9080-2431

Received: 04-10-2022 * Accepted: 01-11-2022 * Published Online: 30-11-2022

***Abstract*—**Fake news is fabricated information that notably impacts our social lives. The massive propagation of fake news by humans or robots severely impacts society and individuals. After the massive increase in the reach of social media platforms, the spread of fake news is unavoidable. Automatic detection of fake news will increasingly reduce the spread of misinformation on digital media platforms. As a contribution to solving this issue, this study recommends a better machine learning algorithm for detecting digital fake news by using a different set of extracted features, namely, regional features and text n-gram. This study uses various machine learning algorithms such as Support Vector Machine (SVM), logistic regression, decision tree, random forest, KNN classifier, MultinomialNB, Passive Aggressive, and Gradient Boost are analyzed with the efficient features for content-based text analysis. Among all the other algorithms, SVM produced outstanding outcomes with an average accuracy of 99.13% and the highest accuracy of 99.3% on the COVID-19 FNIR Dataset.

***Keywords*—**Fake News, Natural Language Processing, feature extraction, n-gram, COVID-19

I. Introduction

With the proliferation of digital mediums of communication, written newspapers were relegated. Readers now have faster access to the most recent updates to the news. Social media sites like Facebook, Twitter, Instagram, and others provide a free, independent forum for us to express our ideas and opinions while maintaining a high level of anonymity. The majority of internet information is not validated using the correct procedures, which is how fake news spreads online.

Fake news is defined as a made-up story and it is a phenomenon that is having a big influence on social life and is making headway quickly. World Health Organization has reported that acting on the wrong information can kill. In the first 3 months of 2020, nearly 6000 people around the globe were hospitalized because of corona misinformation and at least 800 people have died due to misinformation related to COVID-19 (Kupferschmidt, 2022). Patients with painful and persistent symptoms are more likely to use the Internet to learn about or manage their condition than those who are asymptomatic, a practice that has dramatically increased as a result of the SARS-Cov-2 pandemic (Arena *et al*., 2022). They might therefore be more susceptible to the danger of finding false information and spreading it. Social media provided the platform to spread misinformation as quickly as possible. Compared with the amount of information generated on social media, the detection of misinformation is strenuous.

Research on fake news identification is still in its early stages since, at least in terms of the public's attention, this subject is relatively new. To provide precise and dependable methods to identify bogus news, some research has been conducted. Most of them were very useful for scratch this study.

Automatic Deception Detection (Conroy *et al*., 2015) a hybrid approach which is combined with linguistic cue approaches with machine learning, and network analysis approaches which gave 91% accuracy, Fake News Detection: A deep learning approach (Masciari *et al*., 2020) using Artificial Intelligence and Natural language processing technologies with 94.21% accuracy, The data used for this were derived from the Emergent Dataset created by Craig Silverman. Emergent Research (EMERGENT RESEARCH BLOG Www.Emergentresearch.Com, n.d.) is a research and consulting firm been focused on the most dynamic sector of the global community and were able to achieve 94.21%, (Umer *et al*., 2020) Fake News Stance Detection Using Deep Learning Architecture using hybrid Neural Network architecture that is combined with the capabilities of CNN and LSTM, another two-dimensionality reduction approaches, such as Principle

Component Analysis (PCA) and Chi-Square The datasets were acquired from the fake news challenge website. Which has four types of stances: agree, disagree, discuss, and unrelated done same as like (Masciari *et al*., 2020) and got an accuracy of 97.8%, Detecting Fake News in Social Media Networks, (Aldwairi & Alwahedi, 2018) the authors have noticed and they came with the analyzes through clickbait using Logistic classifiers, they got an accuracy of 99.4%,The study of (Lakshmanarao *et al*., 2019) was done in makes an approach to detection of fake news with the use of machine learning algorithms such as Count Vectorizer, Tf-idf Vectorizer, a Naive Bayes Model, and natural language processing and got the best accuracy of 90.7%,The authors of (Patel, 2019) describe the challenges involved in fake news detection and also describe the related task. They also methodically review and compare all tasks. They got datasets from three different websites namely: POLITIFACT.COM, CHANNEL4.COM, and SNOPES.

Bauskar *et al*.(2019) presents a Novel machine learning language model build on NLP Techniques for detecting fake news. They use both content features and social features of news to get a remarkable result. Datasets were taken from PolitiFact and Buzzfeed and achieved an accuracy of 90.33% on a standard dataset. (Ibrishimova & Li, 2020) also used novel framework. They use a hybrid framework based on automating incident classification about previous work. Our model is substantially different from the model proposed by Ibrishimova & Li (2020), instead of relying on a hybrid model (i.e. either content-based or social features). The study has used an efficient machine learning approach using content-based and regional features and these are the researches to be highlighted.

This study undertakes to identify higher predictive machine learning algorithms which automatically detect deceptive information. The research study attempts to analyze the best approach and model for detecting fake news, "whether it is true or fake?" using existing Machine Learning methods and by which classifier, can we detect fake news with a high degree of accuracy.

The whole methodology depends on changing the n-gram size and the range of features to various classifiers including Support Vector Machine (SVM), Logic Regression, Random forest, Gradient boost, and K-Nearest Neighbor. 7000 articles were used for training and testing purposes. Here, dependent and independent variables are content and regional features. The three regions included in this study are India, Europe, and the United States.

The proposed solution to the problem concerned with fake news gives the best approach and model for detecting fake news and implementing the feature extraction methods and capturing the feature extraction such as regional features and n-gram-based features, etc. to increase the higher prediction of the model.

Further, the remaining sections of this paper are structured as follows. Section 2 contains related works. Section 3 discusses the research methodology. Section 4 presents the experimental results achieved on various classifiers and the evaluation of each classifying algorithm. Section 5 makes concluding remarks and discusses future work.

II. Research Methodology

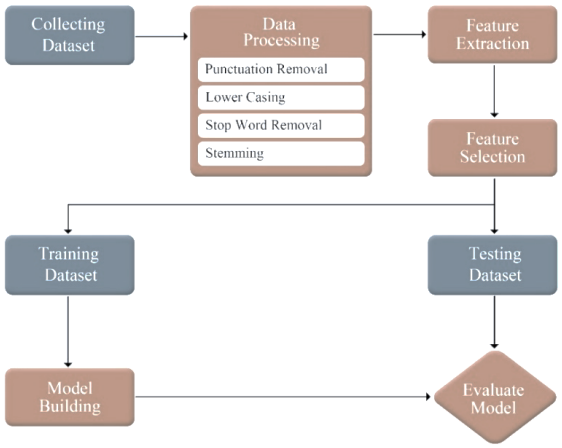

Figure 1: Flowchart of proposed methodology

A. Dataset Description: COVID-19 FNIR DATASET

The particular dataset used for the fake news detection model was collected from the IEEE Data Port Websites (Saenz *et al*., 2021) A CoVID-19-specific dataset called CoVID19-FNIR is made up of verified Twitter handles from news organizations and fact-checked fake news that was scraped from Poynter. The online posts from social media platforms were collected from the regions of India, the United States of America, and Europe between February 2020 and June 2020. The dataset underwent prepossessing procedures, which included removing special characters and irrelevant data.

	Text	Region	Label	content
2549	Nobel Prize laureate claims the new coronavir...	Europe	0	Europe Nobel Prize laureate claims the new co...
784	Coronavirus live news: UK to open temporary Co...	Europe	1	Europe Coronavirus live news: UK to open tempo...
1542	Says Nancy Pelosi's "daughter is on the board...	United States	0	United States Says Nancy Pelosi's "daughter i...
3407	A Rohingya Muslim believes that the thermal s...	India	0	India A Rohingya Muslim believes that the the...
2293	Just in: 16 more #COVID-19 positive cases (15 ...	India	1	India Just in: 16 more #COVID-19 positive case...
1241	Pakistani doctor Osama Riaz's last message fo...	India	0	India Pakistani doctor Osama Riaz's last mess...
2951	According to a tally of Covid-19 cases reporte...	India	1	India According to a tally of Covid-19 cases r...
3463	As Covid-19 cases spike, some states came out ...	India	1	India As Covid-19 cases spike, some states cam...
1850	News about COVID-19 can only be shared after ...	India	0	India News about COVID-19 can only be shared ...
2688	Full disclosure? Hedge funds navigate COVID he...	Europe	1	Europe Full disclosure? Hedge funds navigate C...

Figure 2: Covid-19 FNIR Final Dataset

Figure 2 shows some samples of the Covid-19 FNIR Final Dataset. Initially, the dataset was optimized as suitable to the proposed model. In order to enhance the performance of the dataset, the following processes were applied to the dataset.

1) Dropped NULL Values from both datasets.
2) Dropped unnecessary features from both datasets (except Text, Region & Labels).
3) Merge both datasets.

4) Merged text and region feature and added as a new feature named Content.
5) Shuffled the dataset at frequency level 1.

The final dataset after the optimization process consists of 7,588 unique labelled instances which is the combination of 3795 Fake News and 3793 True News. The details of the labels are shown in Figure 3. The news is related to 3 regions namely Europe, India, and United States. Regions played a major role in the deceptiveness of the news. Figure 4 shows the quick overview of data belonging to each region.

B. Data Preprocessing

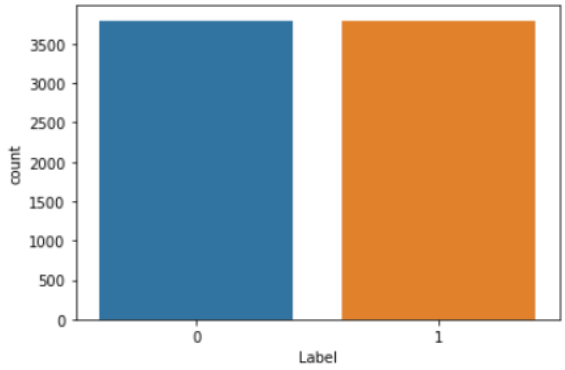

Figure 3: Fake News vs True News

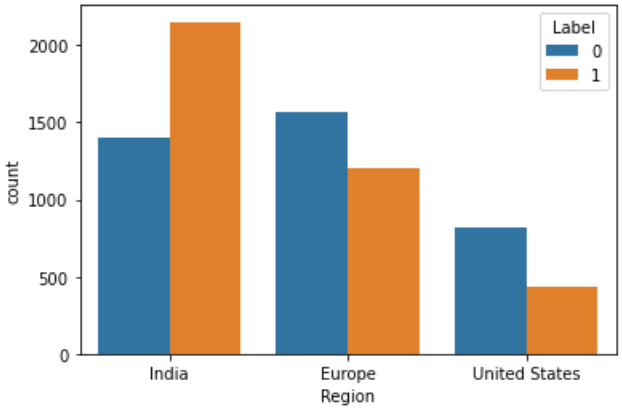

Figure 4: Labels over regions

Data preprocessing is the foremost step to prepare the data to configure, how a machine learning model can understand and the effectiveness of machine learning algorithms is increased by preprocessing the data since some of the algorithms require data to be in a specific format.

Preprocessing is also used for dimensionality reduction. In the vector space model, each term represents an axis. The text or document is made up of vectors in multi-dimensional space. The number of dimensions we utilized in our study is indicated by the number of distinct terms we used.

1) Special Characters Removal:: he special characters in the text do not add any value to the content but offer a grammatical context for each phrase in the natural language. When special characters are added to any word, it becomes difficult to distinguish it from other word

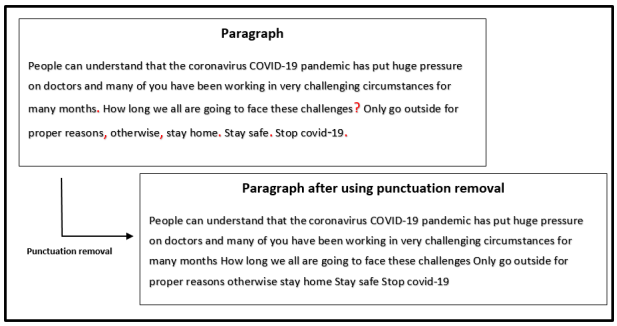

Figure 5: Special Character Removal

2) Lower Casing:: Lowercasing helps to maintain the consistent flow during the preprocessing technique. Words having the same meaning like covid-19, Covid-19 and COVID-19 if they are not converted into lowercase then these both will constitute non-identical words in the vector space model. It will treat all these words as different tokens. After the lower casing, all three words are treated as a single word. This procedure is required because all of the text in the news corpus requires a consistent representation format.

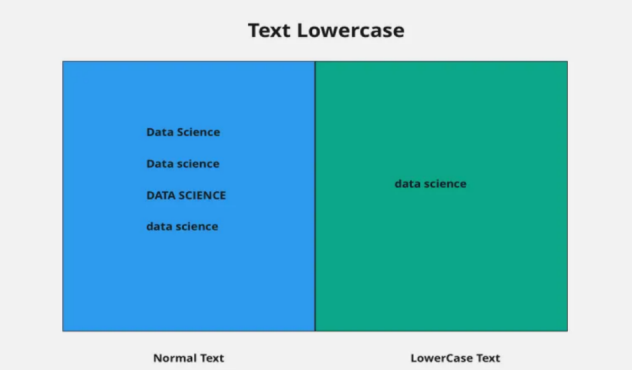

Figure 6: Lower Casing

3) Stop Word Removal:: Stop words are insignificant words in a language that contains less relevant information and they cause noise when used as features in text classification. Stop words includes, conjunctions, prepositions, and Common words and "so", “on" operate more like a connecting portion of the sentences, and now will be removed. Stop words are less important, can waste processing time. To do so, we used the Natural Language Toolkit (NLTK) package.

4) Stemming:: mming is a strategy for minimizing word inflection. (e.g. connected, connection) to their root form (e.g. connect). Stemming is the process of removing prefixes and suffixes from a word until just the stem remains. To make categorization faster and more efficient, we use stemming. We also employ Porter stemmer, which is one of the most widely utilized stemming algorithms due to its precision.

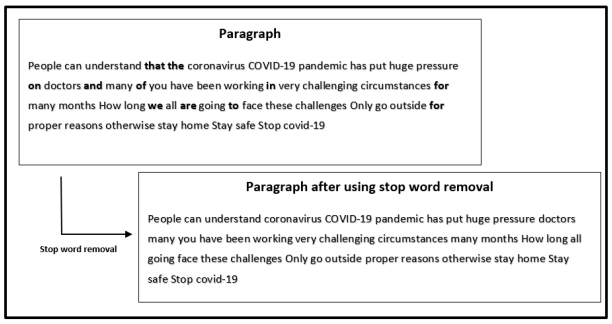

Figure 7: Stop Word Removal

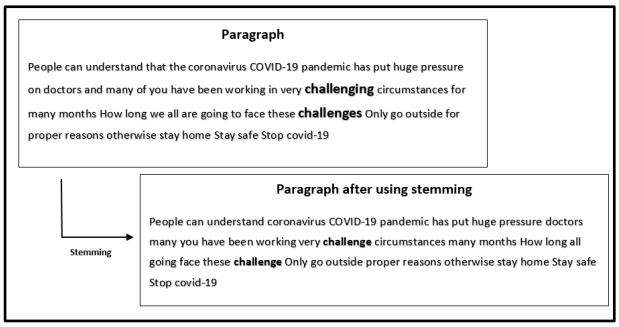

Figure 8: Stemming

C. Feature Extraction

Irrelevant or redundant features can affect the efficiency, accuracy and performance of the classifiers. Thus, this is the superior feature to reduce the text feature size, dimensionality reduction of the text word space and the establishment of its mathematical model. Here we have used CountVectorizer for feature extraction from the sci-kit-learn library in Python.

1) CountVectorizer: ountVectorizer is a utility offered by the Python sci-kit library. The CountVectorizer of Scikit-learn is used to transform a collection of text materials into a token count vector. It also allows the preprocessing of text data before the vector representation is generated. This capability makes it a very versatile text display module. Figure 5 shows how CountVectorizer vectorizes a sentence.

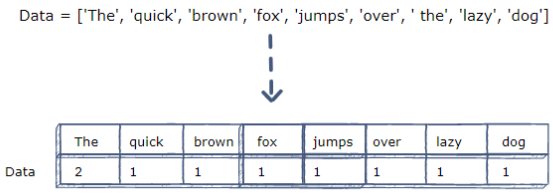

Figure 9: CountVectorizer

2) N-gram Model: -gram modelling is a popular approach used in language modelling and the processing of natural languages. N-gram is an adjacent sequence of n-length elements. The word, byte, syllable or character sequence can be used in categorizing text, the most common n-gram models are word-based and n-grams based on characters. We utilize word-based n-gram for this study to describe the document context and create document classification features. To distinguish between false and factual news, we construct a single N-gram classifier. Figure 6 shows uni-gram, bi-gram and tri-gram breaks the sentence to the size of n.

D. Feature Selection

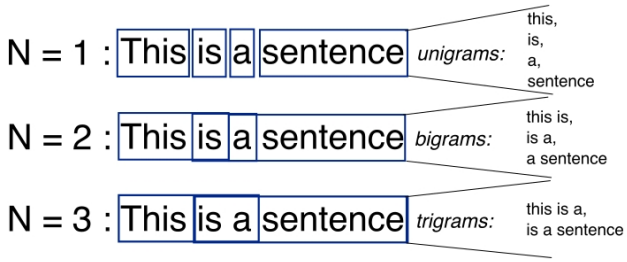

Figure 10: N-gram Model

The process of decreasing the number of inputs when developing a predictive model. It is essential to reduce the count of inputs to the model to reduce the computational cost of the model and as well as to increase its performance. Here, the study used the supervised method since it is a classification problem and has labelled data in the dataset. Therefore, the maximum amount of classification algorithms was used in this section such as Multinomial NB, Logistic Regression, Passive Aggressive Classifier, Decision Tree Classifier, Gradient Boosting Classifier, Random Forest Classifier, K-Nearest Neighbor, and Support Vector Machine.

E. Train and Test Data

In machine learning models, the Train and Test data split is used to evaluate the performance of the model. In this study, the Python programming language provides SKLEARN Package which contains the train-test-split tool which is used to split the dataset into train and test datasets. The current COVID-19 FNIR Dataset is divided by 80% and 20% respectively for the training and test datasets.

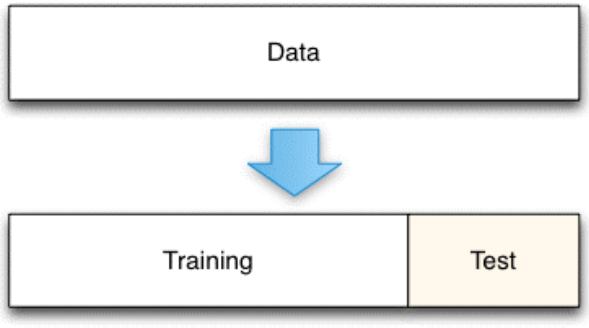

Figure 11: Train Test Split

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.25, random_state=0)
```

Figure 12: Train Test Split Equation

F. Model Evaluation

Accuracy is the ratio of the number of right predictions to the total number of predictions in the data, whether True or False. For balanced datasets, as indicated by the equation below, accuracy is extremely dependable.

$$Accuracy = \frac{correct\ predictions}{total\ predictions} \times 100 \qquad (1)$$

III. Results and Discussion

A. Experimental setup

The machine learning model started to run on the current COVID-19 FNIR Dataset to classify the news whether it is "fake" or "true". The experiment started by studying the impact of the size of the n-gram on the machine learning model's performance. The process started with unigram to n=4, n was raised one at a time.

Additionally, in each increment of n, the number of maximum features was at 4 different states. Such as 1000, 5000, 10000 and 50000 respectively. The experiments use 5-fold cross-validation, with the dataset split 80% for training and 20% for testing in each validation cycle.

Eight Learning models were created using various classification methods, and the trained models were utilized to predict the labels assigned to the testing data. The findings of each step's experiments were then presented and reviewed.

B. Experimental Results

The study was done on the Count Vectorization feature extraction method and then, the size of n-gram increased to 4 from 1. Also, the number of features ranging was 1,000 to 50,000.

The results achieved through the experiments shown in Figure 13, the highest accuracy was achieved is 99.3% throughout the whole experiment on Support Vector Machine (SVM) and Random Forest Classifiers. However, the rest of the classifiers achieved good results too. However, among the other classifiers, SVM had the highest average accuracy. Figure 8 shows the confusion matrix produced by the SVM classifier.

IV. Conclusion

This research work presents the task of automatic detection of fake news. It presented the method of detecting fake news using n-grams of text content as the best-chosen feature extraction technique. Moreover, this study has implemented two extraction features and compared them with eight different machine learning techniques. We found the most prominent framework for detecting fake news using a combination of content-based and regional features.

e research work concludes that all the chosen machine learning algorithms achieved good accuracy, among them, the Support Vector Machine (SVM) achieved the highest accuracy of 99.3% and the highest average accuracy of 99.13% throughout the whole experiment. The study recommends that SVM provides the best of accuracy for this particular case study and dataset among all classifiers, because of its higher accuracy achieved.

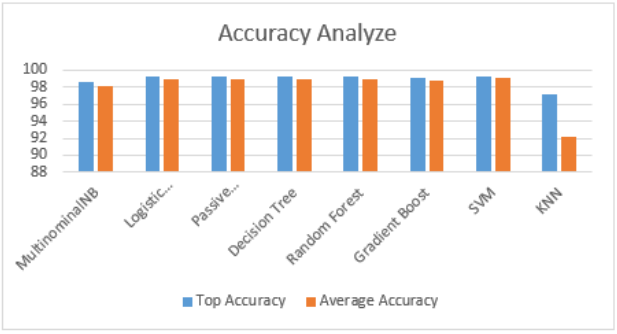

Figure 13: Accuracy analysis of each classifier

The results achieved through the experiments shown in Figure 16, the highest accuracy was achieved is 99.3% throughout the whole experiment on Support Vector Machine (SVM) and Random Forest Classifers. However, the rest of the classifiers achieved good results too. However, among the other classifiers, SVM had the highest average accuracy. Figure 14 shows the confusion matrix produced by SVM classifier

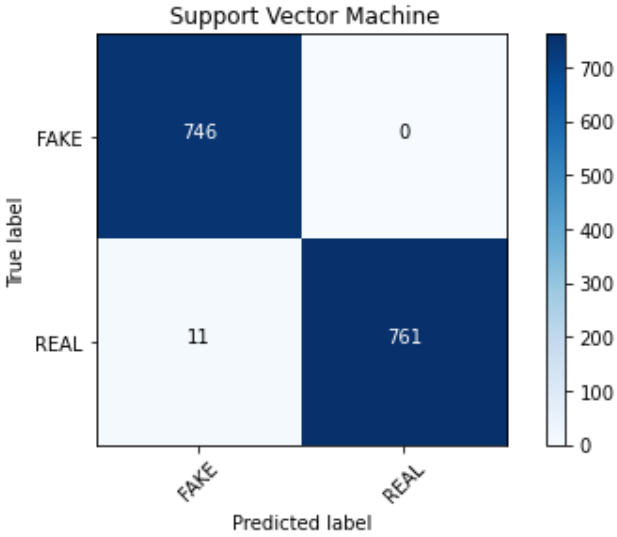

Figure 14: Confusion Matrix of SVM

References

Aldwairi, M., & Alwahedi, A. (2018). Detecting fake news in social media networks. *Procedia Computer Science*, 141, 215–222. https://doi.org/10.1016/j.procs.2018.10.171

Table I: Accuracy of classifiers

Classifier		Number of features				Average Accuracy
MultinomialNB Algorithm	**N-Gram**	**1,000**	**5,000**	**10,000**	**50,000**	98.06%
	Uni - gram	98.3	98.1	97.7	97.6	
	Bi - gram	98.4	98.2	97.7	97.2	
	Tri - gram	98.4	98.6	98.2	97.7	
	Four - gram	98.3	98.5	98.2	97.8	
Logistic Regression	Uni - gram	99.0	99.1	98.9	98.9	98.98%
	Bi - gram	99.2	99.0	98.9	98.9	
	Tri - gram	99.1	99.0	98.9	98.9	
	Four - gram	99.1	99.0	98.9	98.9	
Passive Aggressive Classifier	Uni - gram	99.0	99.1	98.9	98.9	98.98%
	Bi - gram	99.2	99.0	98.9	98.9	
	Tri - gram	99.1	99.0	98.9	98.9	
	Four - gram	99.1	99.0	98.9	98.9	
Decision Tree Classifier	Uni - gram	98.9	99.1	99.2	98.9	98.91%
	Bi - gram	98.9	99.1	98.9	98.9	
	Tri - gram	98.7	98.7	98.9	99.0	
	Four - gram	98.7	98.8	99.0	98.9	
Random Forest Classifier	Uni - gram	99.2	99.1	99.2	99.3	98.89%
	Bi - gram	98.9	99.0	98.9	98.9	
	Tri - gram	98.8	98.8	98.9	98.4	
	Four - gram	98.8	98.8	98.6	98.6	
Gradient Boost Classifier	Uni - gram	98.9	98.7	98.9	98.6	98.78%
	Bi - gram	99.0	98.6	98.9	98.6	
	Tri - gram	99.1	98.6	98.9	98.6	
	Four - gram	98.9	98.8	98.8	98.6	
Support Vector Machine	Uni - gram	99.3	99.2	99.3	99.3	99.13%
	Bi - gram	99.2	98.9	99.3	99.0	
	Tri - gram	99.1	98.9	99.2	99.0	
	Four - gram	99.1	98.9	99.2	99.2	
KNN Classifier	Uni - gram	96.1	91.7	89.7	86.9	92.11%
	Bi - gram	96.9	94.0	90.9	80.7	
	Tri - gram	97.1	95.8	93.8	86.4	
	Four - gram	97.2	96.0	94.0	86.5	

Arena, A., Degli Esposti, E., Orsini, B., Verrelli, L., Rodondi, G., Lenzi, J., Casadio, P., & Seracchioli, R. (2022). The social media effect: the impact of fake news on women affected by endometriosis. A prospective observational study. *European Journal of Obstetrics & Gynecology and Reproductive Biology, 274*, 101–105. https://doi.org/10.1016/j.ejogrb.2022.05.020

Bauskar, S., Badole, V., Jain, P., & Chawla, M. (2019). Natural Language Processing based Hybrid Model for Detecting Fake News Using Content-Based Features and Social Features. *International Journal of Information Engineering and Electronic Business, 11*(4), 1–10. https://doi.org/10.5815/ijieeb.2019.04.01

Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology, 52*(1), 1–4. https://doi.org/10.1002/pra2.2015.145052010082

EMERGENT RESEARCH BLOG www.emergentresearch.com. (n.d.). Ibrishimova, M. D., & Li, K. F. (2020). A machine learning approach to fake news detection using knowledge verification and natural language processing. In *Advances in Intelligent Systems and Computing* (Vol. 1035). Springer International Publishing. https://doi.org/10.1007/978-3-030-29035-$1_2$2

Kupferschmidt, K. (2022). Science.org. https://doi.org/10.1126/science.abq1754

Lakshmanarao, A., Swathi, Y., & Srinivasa Ravi Kiran, T. (2019). An effecient fake news detection system using machine learning. *International Journal of Innovative Technology and Exploring Engineering, 8*(10), 3125–3129. https://doi.org/10.35940/ijitee.J9453.0881019

Masciari, E., Moscato, V., Picariello, A., & Sperli, G. (2020). A Deep Learning Approach to Fake News Detection. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12117* LNAI(3), 113–122. https://doi.org/10.1007/978-3-030-59491-$6_1$1

Saenz, J. A., Gopal, S. R. K., & Shukla, D. (2021). Covid-19 Fake News Infodemic Research Dataset (CoVID19-FNIR Dataset). *IEEE Dataport*.

Umer, M., Imtiaz, Z., Ullah, S., Mehmood, A., Choi, G. S., & On, B. W. (2020). Fake news stance detection using deep learning architecture (CNN-LSTM). *IEEE Access*, 8, 156695–156706. https://doi.org/10.1109/ACCESS.2020.3019735