iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.52778

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Identifying Fake News via Machine Learning and Web Scrapping

Devanshu Rathi[1], Vansh Rahangdale[2], Rajendrasingh Rajpurohit[3], Sandip Shinde[4], Rahul Pandita[5]
Computer Engineering, Vishhwakarma Institute of Technology, Pune , India

Abstract: ***After digitalization, the increase in use of social media has led the flow of information among the social network users unchecked.***
The news is immediately transported to social networks, where it is quickly read, marked with opinions (on Facebook), and shared (on Twitter and Facebook) without being verified as accurate or false numerous times.
In the modern world, fake news is an issue that has grown difficult to detect. Its impact on the judgement of commonfolk is noticeable, society has witnessed numerous events where the rampant flow of unverified news had affected the society as a whole. The sharing of fake news has become a major issue to the society. Our project aims to aid in reduction of e-crimes like use of social media for sharing of fake content. This will be achieved by training the machine to identify text based fake content. Using this the time for work of filtering out the fake content will be reduced by great lengths whilst helping the mass to get verified and credible news.
Keywords: Machine Learning, Natural Language Processing, Naïve Bayes, and Fake News.

I. INTRODUCTION

Fake news detection comes under the text classification [1] and in simpler term is the function of classifying the given news as true or false.
The several types of fake news comprise disinformation (with a goal to intentionally deceive the public), misinformation (wrong information without motive), and rumours, clickbait (misleading headlines), fake news, parodies [2].
Recent studies [2, 3] show that Fake news is spreading at an unprecedented rate, which has led to its widespread dissemination. The effects of such news can be observed in post-election instability or in anti-vaccine groups that hindered the efforts put against COVID-19. Therefore, it is crucial to halt the circulation of false information as soon as possible. The dissemination of false information against vaccinations and the myth that vaccines are harmful are both glaring examples of this.
The project follows to the previously put forward ideology [4] while increasing the number and quality of the dataset and reducing the number of inputs to lowest possible. Making the project a least user input to maximum accuracy experiment, which allowed us to come to a few conclusions that are further discussed in the report.

II. LITERATURE REVIEW

The elements that assist the experiment are Dataset and algorithms for machine. Dataset can contain various information, which can be classified into two types social context and news content. Social context simply being tweet features like followers, likes, shares and other factors that are attached to a tweet while news content being statements manually labelled into ‘true’ or ‘false’ [4].
Algorithm in the other hand can be of various types like state-of-the-art BART, BERT, GPT-2 model which can be used to train the machine, networks like Long Short-Term Memory (LSTM) network has also been used in frameworks like DEFEND [6].
Choosing the right algorithm for the job at hand is a necessary step in testing the model. Algorithms that were taken in consideration for the experiments were Random Forest, Support Vector Machines, and Naive Bayes will each be put to the test using various parameters to gauge their performance and accuracy [4].
The techniques used to identify fake news are often trained on the most recent data (at that moment) which might not generalize to upcoming activities. Most of the examples with labels from the confirmed fake news quickly becomes obsolete when current affairs are taken into view. For instance, a model trained in fake news data from before the period of COVID-19 may not appropriately categories false news during COVID-19.
The challenge of supplied data to target variable is termed as Concept drift [5], something which need to be overcome and the possible solutions are discussed further into the paper.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

III. METHODOLOGY

A. *Flowchart/ Theory*

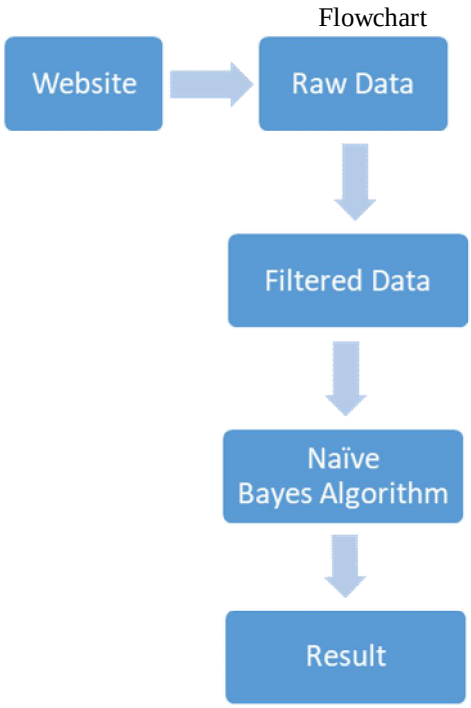


Fig. 1 – Process flow in the proposed system

B. *Theory*

As discussed before three algorithms were considered but since the experiment's goal is to take the least input while the random forest and SVM (Support vector machine) takes various parameter into consideration Naive bayes algorithm was selected for the experiment. Early-stage experiments were conducted via Kaggle dataset but was switched to dataset generated via web scrapping during final stages. So, the input data is pre-processed using natural language processing via NLTK.

1) *Web Scrapping:* The practise of deploying bots to gather information and material from a website is known as web scraping. In web scraping, a request is made to the server when open web scraper code is utilised and executed to the URL that has been copied. In response to the request, server transmits the data and permits view via a HTML or XML page. Following that, the code locates the data, extracts it, and parses the HTML or XML page. With this method, the code is utilised to efficiently collect data from many websites and social media platforms.
2) *Pre-processing:* In data processing, we begin with identification of Null or missing values and eliminated all such cases. All unnecessary components within a tweet like tagged username or links were also omitted to reduce noise. All stop words were eliminated using NLTK, and word tokenization was also carried out for greater accuracy.
3) *Naïve bayes Theorem:* Bayes' Theorem is stated as: P(class|data) = (P(data|class) * P(class)) / P(data) and is used to calculated probability for classifier models.

IV. DESIGN

The input as mentioned is aimed to be the least which led to adopting the naïve bayes algorithm, thus the only input needed is the keywords to be searched. Using HTML, CSS and Flask the webpage is developed through which the user interacts with the backend. The backend will require importing libraries like Tweepy, Pandas, etc. Most recent tweet containing the query passed by user through the front end will be extracted using tweepy which is an open-source Python module that provides access to the Twitter API. The scrapped tweet is then stored in database using MySQL and by natural language processing (NLP) the data is filtered. Further using Naïve Bayes approach probability for the event is determined by using the dataset which is created using web scrapping consisting of true statements and false statements. The final result obtained at last is show to the user at the front end.

V. EXPERIMENTS

A. *Experiment 1*

In first experiment, the model was trained via dataset of total 44000 labelled statements. The model achieved accuracy of 95.57%. It took 144 seconds to test the model.

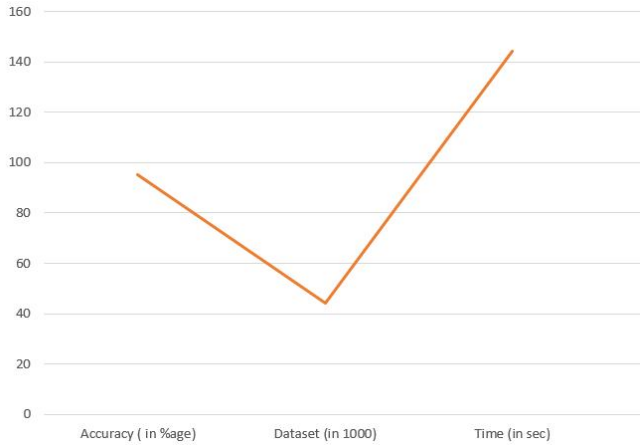


Fig. 2 – Expt. 1 Results

B. *Experiment*

Since the accuracy was good enough, dataset was kept the same, now the major setback was time. Thus, to reduce the time consumption parallel computing was implemented. This helped in reducing the time to 89 seconds.

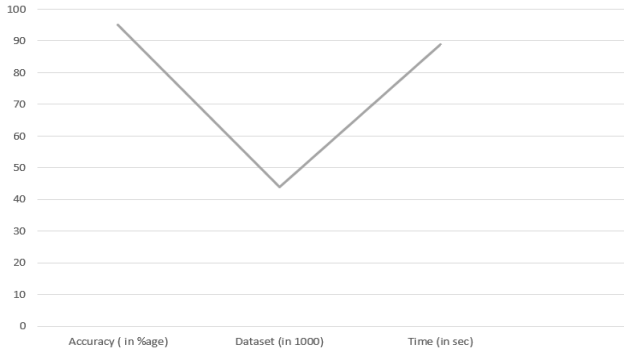


Fig. 3 – Expt. 2 Results

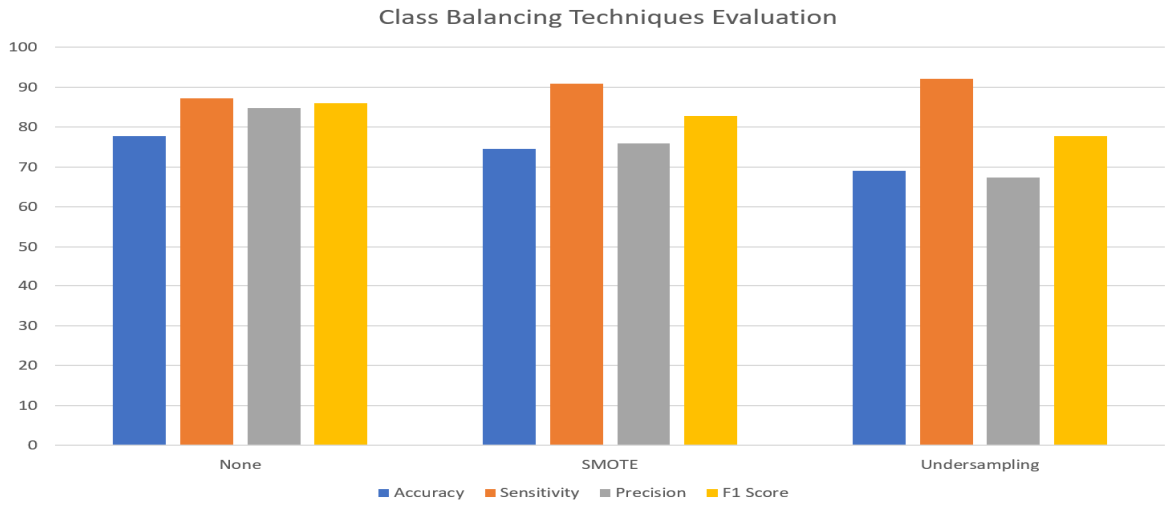


Fig.4 – Histogram of values

C. Experiment 3

It has been identified that now we must create a system that can keep up with the rate of updating with real time information, thus web scrapping has been implemented. In order to build a dataset that could update with real-time information, the prior dataset was discarded. It was achieved via web scrapping sites like PolitiFact. PolitiFact is a website which previous partnered with Facebook (now meta) to create a similar engine to ours which rated a post's credibility. The website PolitiFact assigns values as false, half-true, mostly-false, pants-fire, true, mostly-true, barely-true to any news as its credibility score. Given that the proposed model is a classifier type, such target values create noise in model. So false, half-true, mostly-false and pants-fire were taken as false values while true, mostly-true, barely-true as true values.

The website also offers a flip-o-meter that displays half-flip, full-flop, and no-flip values, none of which are useful because they are neither true nor false, therefore we eliminated them from the training model.

Beautiful soup and requests were libraries imported for web scrapping and web scraped data was organized into csv file using pandas. Current dataset has 3521 false values and 976 true values, so to resolve the class imbalance smote (oversampling), under sampling, smoteenn (oversamling+undersampling) were applied as an experiment.

Balancing techniques had better sensitivity but in all other parameter figures were better without performing balancing. The gain in sensitivity was not significant enough compared to the decrease in precision, accuracy and F1 score. As the dataset increased the accuracy jumped to 79.11%, concluding that with time the accuracy of the model also shows parallel growth. We also switched from system-based source-code editor to google collab resulting in increasing processing speed. The present execution time is only 15 seconds given the size of the dataset is much smaller as compared to previous experiments.

Confusion matrix was used to define the performance of our classifier algorithm, the result is shown below:

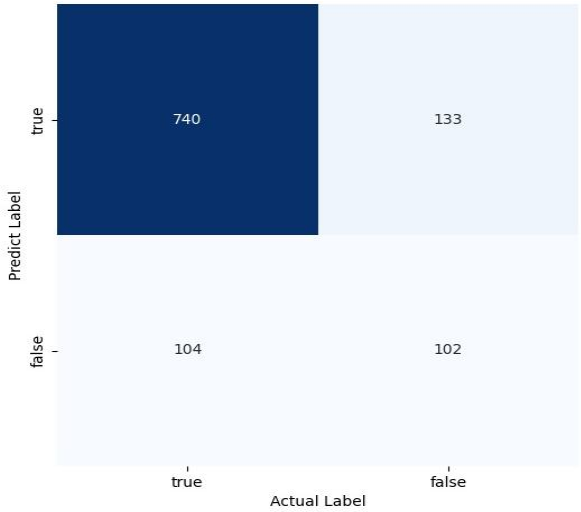


Fig.5 – Confusion Matrix

Calculated by confusion matrix the values of sensitivity, specificity and precision were 85.42%, 52.60% and 88.34% respectively.

The overall growth with each experiment performed can be seen in the graph below,

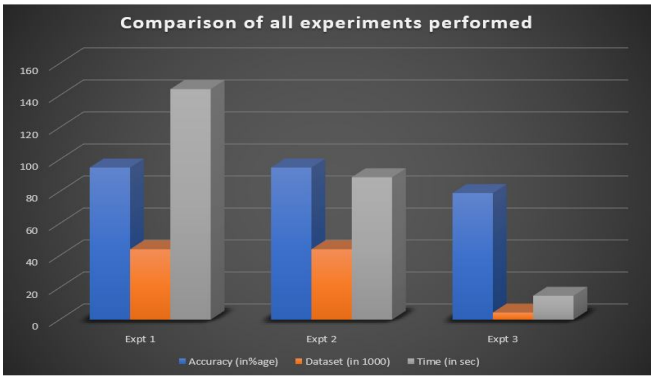


Fig.7 – Growth Chart during all Experiments

VI. RESULTS AND DISCUSSIONS

A. Result

As can be seen the result of the project is a frontend where the user puts in the query which is transferred to backend where the machine uses naïve bayes algorithm to calculate the possibility of the event. The final accuracy being 79.11% and constantly evolving with time and execution period being 15 seconds.

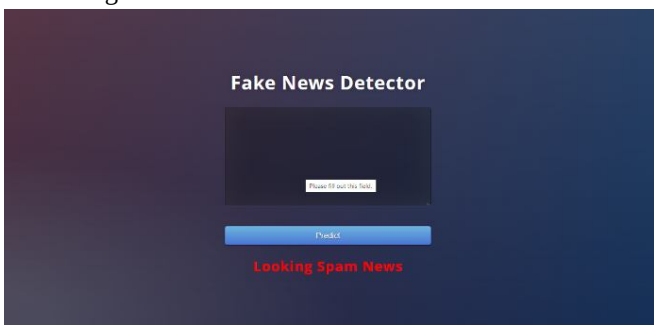


Fig. 6 – Front end

B. Discussion

The entire experiment has led us to the conclusion that working with real-time dataset is most prominent. A machine that can update the dataset continuously with fewer, more pertinent statements would therefore perform better. As demonstrated by the fourth experiment, accuracy may be subpar at first but will increase over time. The ideal system will be able to update statements every second, however if a query where past event is crucial segment is run, old news must also be preserved for precise output. So, creating a large database seems like an option but it turns into a stalemate as it will increase processing time exponentially as the data keeps getting scrapped. Concluding that keeping dataset massive or limited both options depend on the situation and neither choice is truly advantageous. Techniques like RNN (recurrent neural networks) can also be implemented which extends the memory so the results of the query executed can also be used as an input thus increasing accuracy with time. This project aimed at getting best result with minimum input, which was words of the statement, but new parameters can also be introduced which will enable the use of SVM and Radom Forest algorithms too.

VII. CONCLUSION

This paper presents a method of detecting fake news using naïve bayes, trying to reduce fake news circulation by detecting whether the news/tweet is true or false prior to spreading. The project mainly focused developing a machine which can detect fake news with least input which helps in preserving any user's privacy. Thus, various experiments were performed and conclusion were arrived on, which were discussed in the paper.

VIII. ACKNOWLEDGEMENT

The project has involved our group's efforts. Without our individual efforts and the kind assistance of our project guide, it wouldn't have been achievable. Therefore, we would want to express our gratitude to sir for guiding us through this project.

REFERENCES

[1] Liu, C., Wu, X., Yu, M., Li, G., Jiang, J., Huang, W., Lu, X.: A two-stage model based on BERT for short fake news detection. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 11776 LNAI, pp. 172–183 (2019). https://doi.org/10.1007/978-3-030-29563-9_17

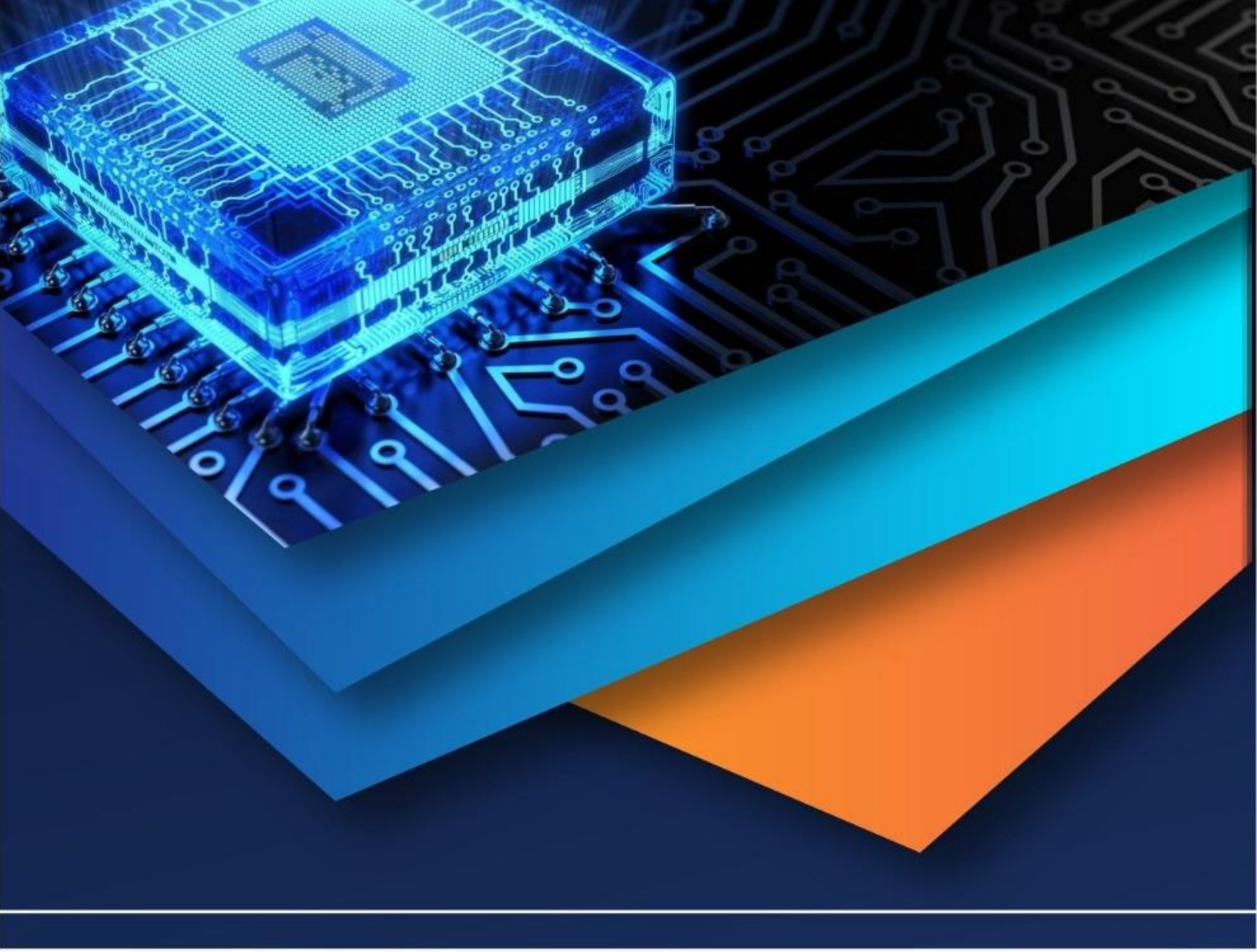
[2] Zhou, X., Zafarani, R.: A survey of fake news: fundamental theories, detection methods, and opportunities. ACM Comput. Surv. (2020). https://doi.org/10.1145/3395046

[3] Shu, K., Wang, S., Liu, H.: Beyond news contents: The role of social context for fake news detection. In: WSDM 2019—Proceedings of 12th ACM International Conference on Web Search Data Mining, vol. 9, pp. 312–320 (2019).

[4] Ciprian-Gabriel, Cusmuliuc & Cusmuliuc, Georgiana & Iftene, Adrian. (2018). IDENTIFYING FAKE NEWS ON TWITTER USING NAÏVE BAYES, SVM AND RANDOM FOREST DISTRIBUTED ALGORITHMS.

[5] Hoens, T.R., Polikar, R., Chawla, N.: V: Learning from streaming data with concept drift and imbalance: an overview. Prog. Artif. Intell. 1, 89–101 (2012)

[6] Raza, S., Ding, C. Fake news detection based on news content and social contexts: a transformer-based approach. Int J Data Sci Anal 13, 335–362 (2022).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY