iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VII **Month of publication:** July 2021

DOI: https://doi.org/10.22214/ijraset.2021.37123

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: 7.429
Volume 9 Issue VII July 2021- Available at www.ijraset.com

Fake News Classification Using Outliner Detection and Trend Analysis

Anurag Srivastava[1], Abhinav Kaushik[2], Archit Singh[3]
[1, 2, 3]*Dept. of Computer Science and Engineering, IMS Engineering College, Ghaziabad, India*

Abstract: In modern era fake news is one of the major causes for disrupted social harmony, impact of fake news can lead to various unforeseen situations and thus affect the society as a whole. This paper proposed the use of anomaly detection and trend analysis for detecting fake news.
Keywords-fake news, outliner detection, trend analysis, anomaly detection, machine learning

I. INTRODUCTION

Fake news has always been pervasive, it is not new. Journalists may be sued for libel anytime. Whether in the domain of celebrity tabloid magazine or even in the most formal and respectable broadsheet, non-credible or non-factual information have been used and abused to influence, mislead, or deceive. While there are laws penalizing libellous articles, but with the rapid increase of social media, it is tremendously difficult to control spreading of misinformation. Popular platforms such as Facebook and Twitter were intentionally designed to make it easy for people to share any article, credible or not, with even the motivation to make it viral. Especially when nowadays most people get their news from the web being it is free and convenient, thus making these platforms enabler of fake news. According to a report named "The spread of true and false news online" by MIT scholars, it is found that about 126K rumours were spread by around 3 million people on twitter from 2006-2017.

Thus, concluding that false news reached more people than the truth; the top 1% of false news reached between 1000 and 100,000 people, whereas the truth rarely reached to more than 1000 people. This study describes a methodology that can be adopted for detecting fake news.

II. RELATED STUDIES

Humans are unreliable detectors of fake news. This is because people are susceptible to bias. People tend to believe those that do not contradict their preconceived ideas; Share articles just because it was shared by a celebrity (or by a celebrity fake account). But then again, automated detection tools or even Artificial Intelligence alone does not cut it. AI's machine learning approach detects content whether it is false or not based on patterns from training data, however, it does not tell if a statement is a lie or a joke. There are already studies done in the domain but the possibility of fake news spread by new conventions always leaves some space for further research.

A. Web Scraping

Web scraping is an automated approach by the means of which subsequent data can be collected from the web. Web scraping plays an essential role in data collection and thus making the prediction model more efficient. In this project, web scraping is used to collect data from various news portals and process it to strip noise out of it and further using it to train the machine learning model. Web scraping is also used to analyse the current ongoing trends and thus helping the outliner detection model to predict the scraped article.

B. Natural Language Processing

Satirical news is deceptive as it causes false beliefs among its readers, thus they are fake news. Satire and irony are parts of the complexities of the human language. Natural Language Processing (NLP), a branch of Artificial Intelligence, is applied for detecting the crux of the statement. NLP techniques with machine learning algorithms processes news content through language pattern detection, sentiment analysis, lemmatization, etc. In the project the scraped articles are first processed using NLP, this converts the textual data into non-textual or numerical format that are technically known as count vectors. These vectors are used to further classify the article as real or fake.

C. Principal Component Analysis

Principal component analysis (PCA) is a dimensionality reduction technique that is used to reduce the number of features in the given dataset. This project utilizes PCA for reducing the number of features in the generated count vector. Initially count vector contains a total of 18K features which was a lot and it comprises of a lot of noise too. After applying PCA the number of features gets reduced from 18K features to 50 features which was the right amount that we needed for our model.

©IJRASET: All Rights are Reserved

D. Outliner Detection

Outliners are defined as extreme values that differ from previous observations or deviates from seen data. Outliner detection is also known as "Anomaly Detection". This project utilizes outliner detection for detecting fake news and analysing trend for the specific keywords from the article. For outliner detection isolation forest algorithm is used in this project. Outliner detection using isolation forest is a two-stage process.

The first (training) stage builds isolation trees using subsamples of the training set. The second (testing) stage passes the test instances through isolation trees to obtain an anomaly score for each instance.

The result after applying outliner detection:

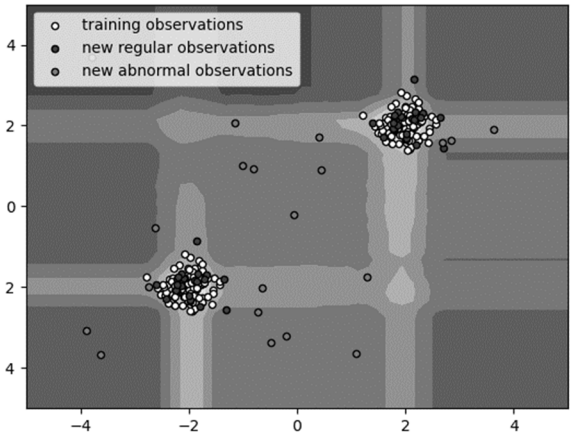


Figure 1: Isolation Forest on scraped data

It can be seen that data gets classified or outlined into two categories that contain two scores 2 and -2, the score of negative 2 gets classified as Fake and the similarly positive gets classified as Real.

E. Trend Analysis

Analysing trend and predicting whether the trend is an anomaly in the system is a tedious task. For doing trend analysis this project utilized the google trends API by Google. This API gives the insights about the ongoing trend. Through this API we collected data about various ongoing topics and trained a second anomaly model using this data. For prediction we utilized this data along with the previous model and make a decision to classify the article as fake or real.

These insights helped a lot in getting various information about the topic, and getting the ins and outs of the trend. Thus, trend analysis plays a major role in classifying fake news.

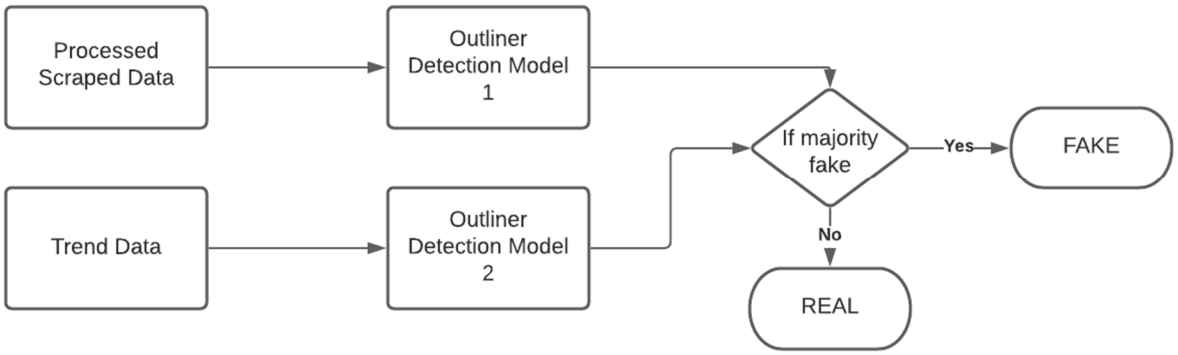


Figure 2: Proposed model for outliner detection

III. METHODOLOGY

To explicate machine learning techniques for automatic fake news detection, this paper uses Comparative Study. Presenting a Descriptive Comparison of each machine learning tool.

©IJRASET: All Rights are Reserved

IV. RESULTS

Each algorithm mentioned in this paper provides promising efficiency as indicated by the considerably high confidence rate, that is based on the number of articles collected, based on its own metrics, and primarily based on its own definition of fake news thus various cues.

V. CONCLUSIONS

Understanding fake news is a typical task. Consequently, fighting it requires multifaced strategies. Considering that the technology that counter fake news is the same technology that created them, neutralizing it may take more than just the expertise of top tech companies.

VI. FUTURE STUDIES

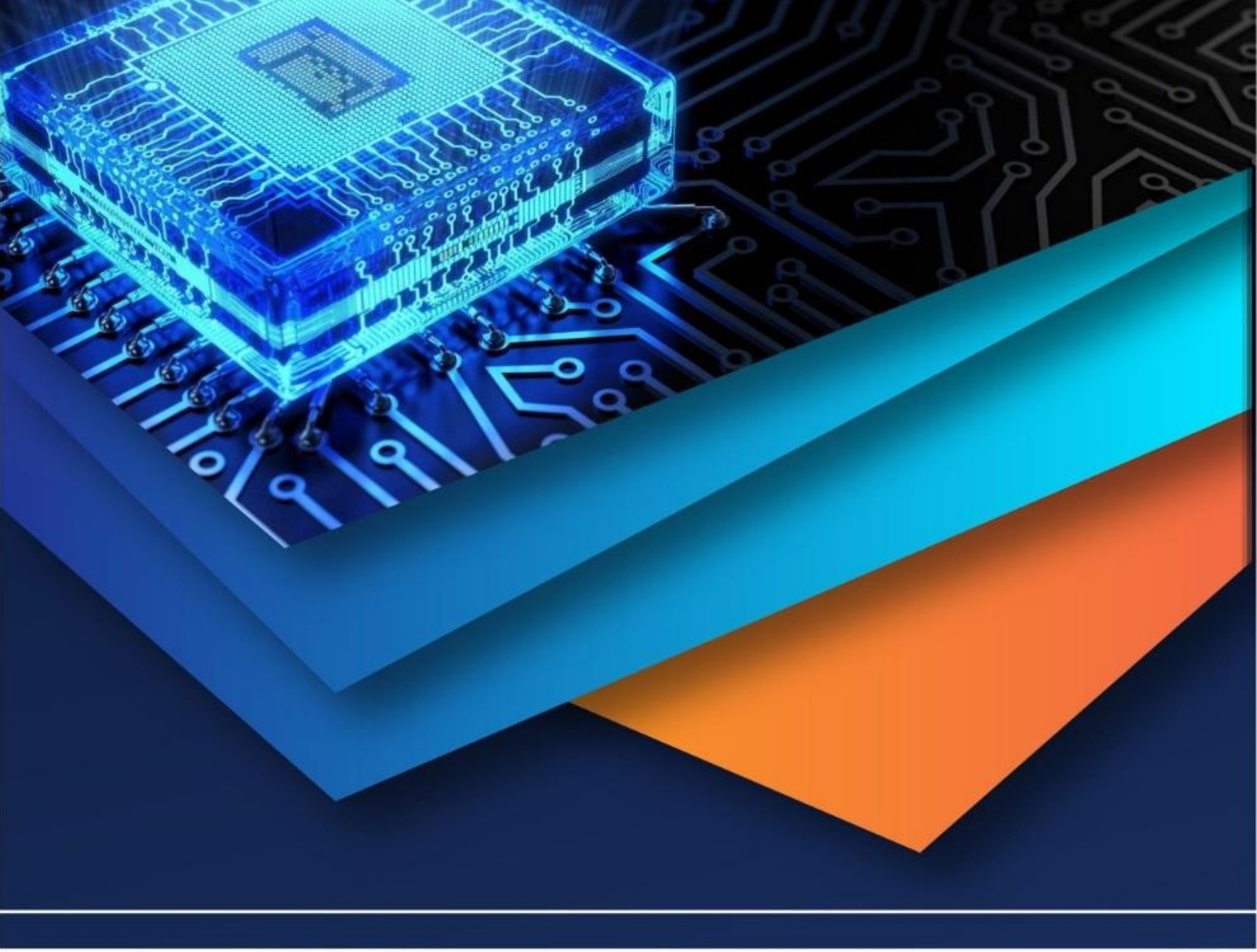
For further future studies this paper recommends the following:

A. Better feature extraction to extract related keywords from the articles.
B. A bigger valid dataset to classify the news more efficiently.
C. Making using of DNN to further enhance the prediction score of the model.

REFERENCES

[1] Doron Sadeh, “Fake News Classification via Anomaly Detection”
[2] Armida Panganiban Salazar, AI Tools on Fake News Detection: An Overview and Comparative Study
[3] Dun Li, Haimei Guo, Zhenfei Wang, and Zhiyun Zheng, Unsupervised Fake News Detection Based on Autoencoder
[4] Soroush Vosoughi1, Deb Roy1, View ORCID ProfileSinan Aral, The spread of true and false news online

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)