CSE3024- Web Mining

Digital Assignment – III

Research Article Submission
(20 Marks)

Fake News Prediction

By

20BCE1731	Arnav Singh
20BCE1773	Archit Agarwal
20BCE1800	Deepansh Tripathi

B.Tech CSE

Submitted to

Dr.A.Bhuvaneswari,
Assistant Professor Senior,
SCOPE, VIT, Chennai

School of Computer Science and Engineering

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

April 2023

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Computing Science and Engineering

VIT Chennai

Vandalur - Kelambakkam Road, Chennai - 600 127

WINTER SEM 22-23

Worklet details

Programme	B.Tech	
Course Name / Code	Web Mining / CSE3024	
Slot	A2	
Faculty Name	Dr. A. Bhuvaneswari	
J Component Title	Fake News Prediction	
Team Members Name | Reg. No	Arnav Singh	20BCE1731
	Archit Agarwal	20BCE1773
	Deepansh Tripathi	20BCE1800

Team Members(s) Contributions – Tentatively planned for implementation:

Worklet Tasks	*Contributor's Names*
Data Collection and Preprocessing using Web Mining Techniques	Arnav Singh
Construction of Social Media Network Graphs using NetworkX	Deepansh Tripathi
Calculation of User-Related Metrics (Centrality Measures)	Archit Agarwal
Visualization of Results using Data Visualization	Deepansh Tripathi
Technical Report Writing	Arnav Singh
Presentation preparation	Archit Agarwal

Fake News Prediction

Arnav Singh
School of Computer Science and Engineering
Vellore Institute of Technology
Chennai, India
arnav.singh2020a@vitstudent.ac.in

Archit Agarwal
School of Computer Science and Engineering
Vellore Institute of Technology
Chennai, India
archit.agarwal2020a@vitstudent.ac.in

Deepansh Tripathi
School of Computer Science and Engineering
Vellore Institute of Technology
Chennai, India
deepansh.tripathi2020@vitstudent.ac.in

***Abstract* — Our project investigates the spread of fake news on social media and aims to identify user-related metrics that can differentiate between users who are likely to spread fake news and those who are not. We will use graph analysis techniques and the NetworkX package in Python to achieve this goal. The study of fake news spread on social media is complex and interdisciplinary, requiring knowledge from multiple fields, including computer science, sociology, and psychology.**

Our project will explore the role of user-related metrics, such as network centrality, in the spread of fake news on social media networks. By conducting a thorough analysis of social media networks, we hope to gain a deeper understanding of the factors that contribute to the spread of fake news. This information will be valuable in developing strategies for detecting and preventing the spread of fake news. Our findings will shape future efforts to mitigate the dangers posed by fake news and promote responsible information sharing practices.

Keywords— fake news, social media, graph analysis, NetworkX, user-related metrics, network centrality, detection, prevention

I. Introduction

Fake news is not a novel concept. It should be noted that publishers used incorrect and misleading material to advance their interests prior to the development of the Internet, so the concept existed before the Internet. Since the internet's inception, an increasing number of people have preferred online platforms for information distribution over traditional media outlets. The latter option is not only faster and more convenient, but it also allows consumers to access multiple periodicals at the same time. However, as content providers began to exploit what has become known as "clickbait," the concept of "false news" was redefined as a result of the evolution. When a person clicks on a link, they are taken to a web page with far less content than they expected. This is referred to as "clickbait." Clickbaits irritate many users, and as a result, they typically spend very little time on such websites. However, because the business side of using online advertisements is heavily reliant on web traffic, more clicks equate to more money for content creators. As a result, despite user concerns about these topics, content publishers have not made much of an effort to avoid using clickbait and publishing inaccurate information. At best, tech companies like Google, Facebook, and Twitter have made an effort to address this specific issue. Organizations have resorted to depriving people connected to such sites of the revenue they would have received from increased traffic, so these efforts have done little to help solve the problem. Users, on the other hand, must still deal with websites that post fake content and make it more difficult for readers to engage with legitimate news. Because the emergence and subsequent development of social media platforms have tended to compound the problem, companies such as Facebook are involved in the fake news problem. The majority of these websites, in particular, provide a sharing option that encourages users to spread the information on the page further. Because social networking sites make it simple and quick to distribute content, users can quickly spread false information. Following the data breach at Cambridge Analytica that affected millions of accounts, Facebook and other industry titans pledged to take additional steps to combat the spread of false information..

II. Literature Survey

A. Automated Fake News Detection: A Survey

This paper provides an overview of the challenges involved in detecting fake news on social media and surveys various approaches proposed to tackle the problem. The paper explores traditional machine learning techniques as well as deep learning models such as convolutional neural networks and recurrent neural networks. It also discusses the importance of feature engineering and dataset preparation in achieving accurate detection. The paper concludes with a discussion on the limitations and future research directions in automated fake news detection.

B. Detecting and Analyzing Rumors and Misinformation on Social Media

This paper presents a framework for detecting and analyzing rumors and misinformation on social media. The framework includes techniques such as content analysis, sentiment analysis, and network analysis to identify and track rumors and misinformation campaigns. The paper discusses the importance of real-time monitoring and verification and concludes with a discussion on the challenges and future directions in detecting and analyzing rumors on social media.

C. Deep Learning-Based Approaches for Fake News Detection: A Review

This paper reviews the different deep learning-based approaches proposed for detecting fake news on social media. It includes both supervised and unsupervised learning techniques such as convolutional neural networks, recurrent neural networks, and generative adversarial networks. The paper also discusses the importance of feature engineering,

dataset preparation, and evaluation metrics in achieving accurate detection. Finally, the paper concludes with a discussion on the limitations and future research directions in deep learning-based fake news detection.

D. Fake News Detection on Social Media using Geometric Deep Learning

This paper proposes a novel approach to fake news detection on social media using geometric deep learning techniques. The approach uses graph-based models to analyze the structural properties of social networks and identify misinformation campaigns. The paper also discusses the importance of feature engineering and dataset preparation in achieving accurate detection. Finally, the paper concludes with a discussion on the limitations and future research directions in fake news detection using geometric deep learning.

E. A Survey of Fake News Detection Methods: Algorithms, Datasets, and Future Directions

This paper surveys the different methods proposed for detecting fake news on social media. It includes traditional machine learning approaches as well as deep learning models such as convolutional neural networks, recurrent neural networks, and graph-based models. The paper also discusses the importance of dataset preparation and evaluation metrics in achieving accurate detection. Finally, the paper concludes with a discussion on the limitations and future research directions in fake news detection.

F. Multiple features-based approach for automatic fake news detection on social networks using deep learning

In recent years, the rise of Online Social Networks has led to proliferation of social news such as product advertisement, political news, celebrity's information, etc. Today, the majority of social network users get their news from online sources. The use of the Internet, however, has evolved into a prime platform for communication and the dissemination of false information as a result of OSNs' rising popularity. The dissemination of false information in the form of satires, false reviews, false rumours, ads, and misleading content. Right now, social media rather than traditional media is where bogus news spreads most quickly.To classify the contents as fake and genuine based on their features and behaviour, we have used various machine learning based classification approaches such as KNN,SVM and logistic regression.

G. Supervised Learning for Fake News Detection

Social media false news detection and interpretation have been studied recently by many. These works analyse source and social media features from news stories. We introduce new attributes and assess automatic false news detection methods and features. Fake news detecting feature applicability and relevancy are intriguing. Finally, we discuss false news identification challenges and prospects. Social media has changed how news is produced, disseminated, and consumed, creating unanticipated opportunities and challenges. Disinformation on social media threatens the news ecosystem's credibility. Social media news is unique in that anyone may publish without paying (for example, anyone can set up a Facebook page posing as a newspaper or news agency). Businesses are using social media for more than simply news. Because fake news may harm people and society, the lack of scalable fact checking is especially worrisome. Recent research has focused on automating fake news detection and comprehending it.

H. Optimal Resource Allocation Over Time and Degree Classes for Maximizing Information Dissemination in Social Networks

Since social media has connected billions of people, people in different roles and behaviors are prevalent. Users who bridge communities' profit from controlling the key information diffusion pathways, according to the structural hole (SH) theory. Understanding user behavior and social network architecture need SH spanners. We apply SHs theory to social network analysis, where SH spanners have informational and managerial advantages. Network centrality-based and information flow-based SH spanner detection techniques are examined. We show practitioners how SH theory may be used to business, social network information dispersion, software development, mobile apps, and machine learning (ML)-based social prediction. Our review covers SH origins, detection methods, and uses. The knowledge can help researchers and service providers use the theory better and produce valuable tools employing cutting-edge ML methodologies. To encourage research, we suggest network dynamics research themes.

I. Structural Hole Theory in Social Network Analysis: A Review

Since social media has connected billions of people, people in different roles and behaviors are prevalent. Users who bridge communities' profit from controlling the key information diffusion pathways, according to the structural hole (SH) theory. Understanding user behavior and social network architecture need SH spanners. We apply SHs theory to social network analysis, where SH spanners have informational and managerial advantages. Network centrality-based and information flow-based SH spanner detection techniques are examined. We show practitioners how SH theory may be used to business, social network information dispersion, software development, mobile apps, and machine learning (ML)-based social prediction. Our review covers SH origins, detection methods, and uses. The knowledge can help researchers and service providers use the theory better and produce valuable tools employing cutting-edge ML methodologies. To encourage research, we suggest network dynamics research themes.

J. Detecting Fake News with Tweets' Properties

Social media has become a major source of news dissemination, surpassing traditional media. The ease and accessibility of the internet has allowed news to spread faster and more easily on social media compared to traditional news sources. However, not all news on social media is trustworthy, as false information can be created and easily

spread, potentially misleading or misinforming readers. The widespread dissemination of fake news can have negative impacts not only on individuals, but also on society as a whole. This can affect how readers perceive online news on social media and indirectly impact their response to real news. While there are manual fact-checking websites to verify the authenticity of news, they cannot keep up with the volume of fast-spreading information online, particularly on social media. To address this issue, automated fact-checking applications have been developed to provide the necessary automation and scalability. However, existing applications lack an inclusive dataset with multi-dimensional information to accurately detect fake news characteristics. To overcome this limitation, they transformed social media Twitter data to identify significant attributes that would improve the accuracy of machine learning methods in classifying news as real or fake. In this paper, they present the mechanisms for identifying these significant Twitter attributes and the application architecture for systematically automating the classification of online news.

K. Understanding User Profiles on Social Media for Fake News Detection

Consuming news from social media has become a common practice, with users benefiting from fast dissemination, low cost, and easy access. However, the quality of news on social media is often lower than traditional news sources, leading to widespread fake news. Detecting fake news has become important due to its harmful effects on individuals and society. The performance of detecting fake news based solely on content is generally not effective, and it is suggested that incorporating user social engagement as additional information can improve the detection of fake news. This requires a deeper understanding of the correlation between user profiles on social media and fake news. In this paper, they have constructed real-world datasets measuring users' trust levels in fake news and selected representative groups of "experienced" users who can recognize fake news as false and "naive" users who are more likely to believe fake news. They have also performed a comparative analysis of explicit and implicit profile features between these user groups to reveal their potential in differentiating fake news. The findings of this research lay the foundation for future efforts in automatic fake news detection.

L. Behaviour forensics with side information for multimedia fingerprinting social networks

Multimedia social network users share and trade sophisticated media. Multimedia fingerprinting is used to study human behaviour in multimedia social networks. Side information is unrelated to the compromised multimedia material and can increase discovery odds. We study how side information impacts multimedia fingerprinting and show how the statistical means of the detection data can significantly boost the collusion resistance of the fingerprint detector. We model the fingerprint detector-colluder dynamics as a two-stage extended game with perfect information to probe side information. We can show that the min-max solution is a Nash equilibrium by modelling the colluder-detector behaviour dynamics as a two-stage game and using backward induction to find the game's equilibrium. This study shows how side information can improve system performance to nearly equal the ideal correlation-based detector. This expands fingerprinting system research by showing that side information may boost collusion resistance for any fingerprint code. We also help you choose the optimal collusion technique and detection to secure multimedia assets.

M. Visual Analytics for Multimodal Social Network Analysis: A Design Study with Social Scientists

A growing area of interest for social network analysis (SNA) is differentiating between other kinds of such entities, in addition to actors and their relationships. Social scientists could, for instance, wish to look at unequal relationships in hierarchically rigid organizations or include non-actors like conferences and projects when examining co-authorship patterns. Multimodal social network analysis (mSNA) is the appropriate SNA for such networks. Multimodal social networks are ones where actors and relations belong to several categories, or modes. In this research, we describe a design study on how to complement mSNA with visual analytics tools that we carried out with several social scientist colleagues. We developed a visual representation known as parallel node-link bands (PNLBs) based on an open-ended, formative design approach that separates modes into distinct bands and displays connections between neighboring ones, much like the list view in Jigsaw. The tool was subsequently subjected to a qualitative review with the help of five social scientists, whose comments helped shape a later design phase that included more network indicators. Finally, we did a second qualitative assessment with our collaborators from the social sciences, which gave us further perspectives on the usefulness of the PNLBs representation and the promise of visual analytics for Msna.

N. Fake News Detection Using Machine Learning Approaches

The spread of fake news on social media and other platforms is a major concern because it has the potential to harm society and the country. There has already been a lot of research into finding it. This paper analyses the research on fake news detection and explores the best traditional machine learning models in order to develop a model of a product with supervised machine learning algorithm that can classify fake news as true or false by using tools like Python Scikit-Learn, NLP for textual analysis. To perform tokenization and feature extraction on text data, we recommend using the Python scikit-learn module, which includes useful tools such as the Count Vectorizer and Tiff Vectorizer. This method will yield feature extraction and vectorization. Then, based on the results of the confusion matrix, we will experiment and select the best-fit features to achieve the highest precision.

O. Fake News Detection System using Article Abstraction

Our civilization has been plagued by fake news. Thus, numerous researchers have investigated bogus news. Most

fake news detection algorithms involve language. They struggle to recognise very ambiguous fake news, which can only be detected after determining meaning and the latest linked information. We will show a new Korean fake news detection system based on a human-judged fact database. Our algorithm compares a proposition to semantically related entries in Fact DB to determine its truth. Bidirectional Multi-Perspective Matching for Natural Language Sentence (BiMPM), a deep learning model that performs well on sentence matching, is used for this. BiMPM works badly when the input sentence is too long and has trouble judging unlearned words or relationships between words. We will present a new matching method that combines article abstraction, an entity matching set, and BiMPM to circumvent these limits. Our technique increases bogus news detection in our trial.

III. Proposed Model

The proposed approach for this project involves several steps. Firstly, we will use web scraping and web mining techniques to collect data from various online sources, including social media platforms, online news websites, online forums, and other web pages. The collected data will then undergo pre-processing to remove irrelevant information and to format it into a suitable structure for analysis. This step will also involve data cleaning, normalization, and integration. Using the pre-processed data, we will construct a network of users and web pages to represent the relationships between them. We will use graph theory algorithms and the NetworkX package in Python for network construction.

Graph Theory Algorithms: Graph theory algorithms are used to construct the network of users and web pages based on the collected data. These algorithms use mathematical concepts to represent and analyze the network structure, including nodes (users or web pages) and edges (connections between nodes). For example, degree centrality measures the number of connections a node has, while betweenness centrality measures how often a node lies on the shortest path between two other nodes.

The constructed network will be analyzed to identify user-related metrics that can differentiate between users who are more likely to spread fake news and those who are not. The analysis will include centrality measures such as degree centrality and betweenness centrality, as well as community detection algorithms to identify groups of users with similar interests and behaviors.

Community Detection Algorithms: Community detection algorithms are used to identify groups of users with similar interests and behaviors within the constructed network. These algorithms partition the network into groups, or communities, based on the strength of connections between nodes. For example, the Louvain method and Fluid communities algorithm are popular community detection algorithms.

Percolation Method: The percolation method is another approach to fake news detection that involves randomly removing nodes from the network and measuring the impact on the network's connectivity. The idea behind this method is that the removal of certain nodes will have a disproportionate impact on the network's overall connectivity, which can be used to identify nodes that are critical to the spread of fake news.

Based on the results of the network analysis, we will develop algorithms for detecting fake news, taking into account both content-based features and user-related metrics Finally, the performance of the fake news detection algorithms will be evaluated. This approach will provide valuable insights into the ways in which fake news spreads on the internet, and will be useful in developing new strategies and algorithms for detecting and preventing the spread of fake news.

IV. Algorithms Involved / Psuedocode

Step 1: Collect data from various sources including social media platforms, online news websites, online forums, and other web pages using web scraping and web mining techniques.

Step 2: Pre-process the collected data to remove any irrelevant information and to format it into a suitable structure for analysis. This step will also involve data cleaning, data normalization, and data integration.

Step 3: Construct a network of users and web pages using the pre-processed data. The network will represent the relationships between users and the web pages they engage with. Use graph theory algorithms and the NetworkX package in Python to construct the network.

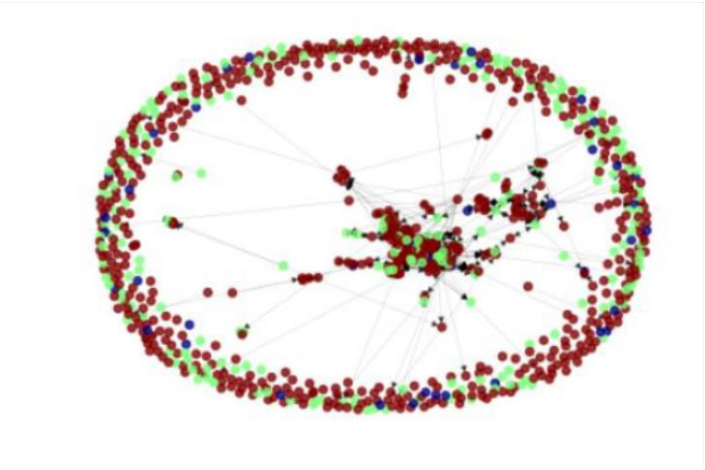
Step 4: Analyze the constructed network to identify user-related metrics that can differentiate between users who are more likely to spread fake news and those who are not. The network analysis will include centrality measures, such as degree centrality and betweenness centrality, as well as community detection algorithms to identify groups of users with similar interests and behaviors.

Step 5: Evaluate the performance of the fake news detection algorithms using metrics such as precision, recall, and F1-score. Perform the evaluation on a separate set of data that was not used in the network analysis and algorithm development steps.

Step 6: Visualize the results of the network analysis and fake news detection algorithms for better interpretation and understanding.

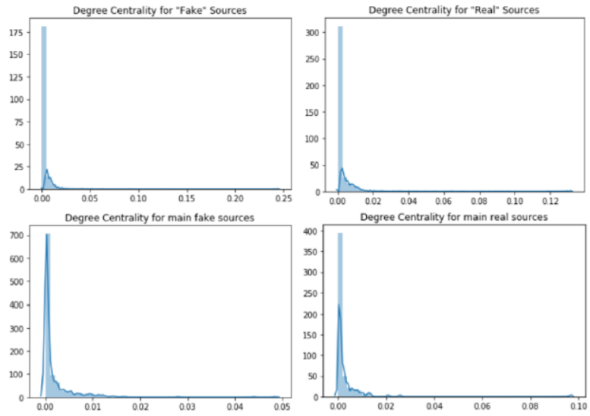
V. Results and Discussions

A. Network Graph

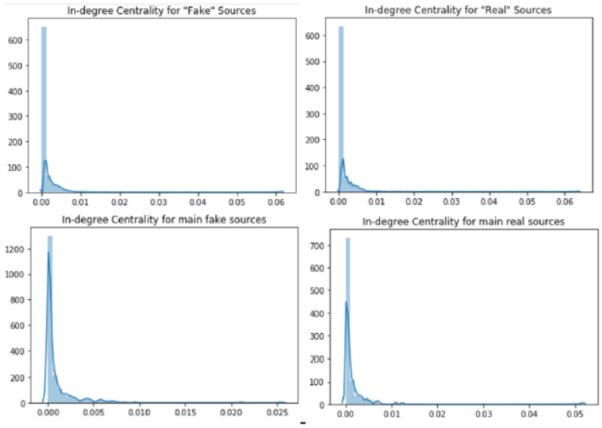


B. *Network Analysis*

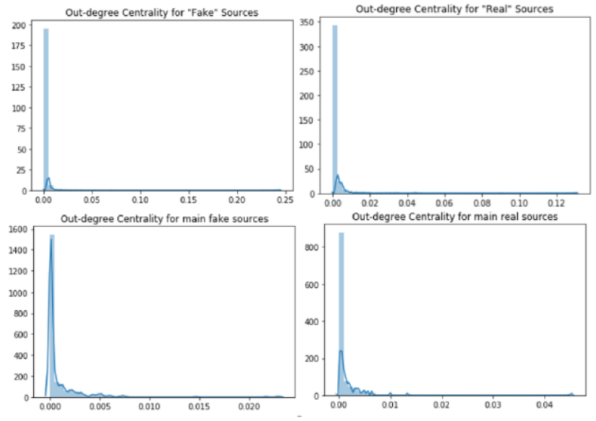
a) Degree Centrality



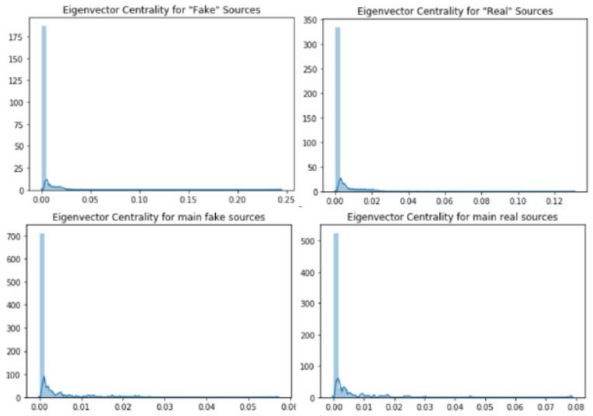
b) In-Degree Centrality



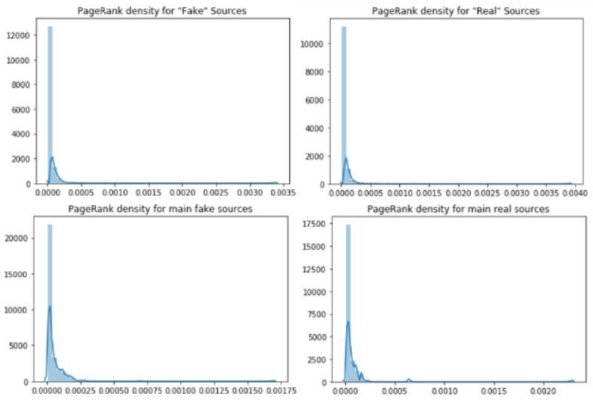
c) Out-Degree Centrality



d) Eigenvector Centrality



e) Page Rank



C. *Community Detection*

a) Louvian Method

	length	nb_fake	nb_real	nb_main_fake	nb_main_real	nb_small_users	unknown credibility
0	5004	3068	1551	129	52	4814	385
1	12365	7451	3949	283	116	11935	965
2	5435	3283	1720	134	45	5246	432
3	4	4	0	0	0	4	0
4	272	160	88	8	1	263	24
5	122	74	34	4	1	117	14
6	9	8	1	1	0	8	0
7	2	1	1	0	0	2	0
8	2	1	1	0	0	2	0
9	8	4	3	0	0	8	1
10	33	22	8	3	0	30	3
11	2	0	2	0	0	2	0

b) Fluid Communities Algorithm

	length	nb_fake	nb_real	nb_main_fake	nb_main_real	nb_small_users	unknown credibility
0	2428	1475	760	57	21	2343	193
1	3630	2170	1154	76	38	3508	306
2	5925	3591	1869	154	60	5701	465
3	2797	1700	878	79	20	2696	219
4	4402	2660	1403	107	41	4246	339

c) Percolation Method

```
[ ] Pcommunities = list(community.k_clique_communities(G_u, 5, cliques=None))

[ ] print(len(Pcommunities))
    l=[]
    for i in Pcommunities:
        l.append(len(i))

    4659
```

The Louvain algorithm is one of the most effective ways to find communities. It works by maximising the modularity score, which is a measure of how many edges are in a community compared to how many are outside of it. When the Louvain algorithm is run on the full, undirected version of the graph, the 214 communities listed in the data frame above are the results. For each community, we know how long it has been around, how many fake and real sources there are, how many fake and real sources the most important users have, how many small users there are, and how many users we don't know anything about.

Then, we tried the Fluid community algorithm, which is based on the same idea as Kmeans. We told the algorithm how many communities to find, and it should find the best ones. To find the best number of communities, we try it with different numbers and figure out the coverage score. However, it looks like the best fit is between two communities. We finally tried a Clique Percolation method, which finds all cliques of k nodes (k is set by the user), draws a graph of cliques with an edge if two cliques share k-1 nodes, and defines as a community all the nodes in a connected part of the new graph. Again, the results (4,695 communities) weren't very useful, so we decided not to include metrics about the communities.

VI. Conclusion

Fake news is something that can have an effect on an individual's belief regarding particular circumstances, and an individual's perspective and ideology towards a scenario can lead to other people developing similar beliefs. This is something that our project addresses by utilising NetworkX and a few additional programmes. We are aware of the dissemination of false information throughout social media networks, and we plan to acquire a more in-depth understanding of the factors that lead to the dissemination of false information by conducting an in-depth investigation of social media networks.

Our findings will assist to guide future initiatives that aim to mitigate the dangers posed by fake news and encourage information-sharing behaviours that adhere to ethical standards.

References

[1] Reeya Baria, Sheshang Degadwala, Rocky Upadhyay, Dhairya Vyas, "Theoretical Evaluation of Machine And Deep Learning For Detecting Fake News", 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), pp.325-329, 2022.

Pronaya Bhattacharya, Shivani Bharatbhai Patel, Rajesh Gupta, Sudeep Tanwar, Joel J. P. C. Rodrigues, "SaTYa: Trusted Bi-LSTM-Based Fake

[2] News Classification Scheme for Smart Community", IEEE Transactions on Computational Social Systems, vol.9, no.6, pp.1758-1767, 2022.

[3] Swatej Patil, Suyog Vairagade, Dipti Theng, "Machine Learning Techniques for the Classification of Fake News", 2021 International Conference on Computational Intelligence and Computing Applications (ICCICA), pp.1-5, 2021.

[4] Ashwaq Khalil, Moath Jarrah, Monther Aldwairi, Yaser Jararweh, "Detecting Arabic Fake News Using Machine Learning", 2021 Second International Conference on Intelligent Data Science Technologies and Applications (IDSTA), pp.171-177, 2021.

[5] Sashank Sridhar, Sowmya Sanagavarapu, "Fake News Detection and Analysis using Multitask Learning with BiLSTM CapsNet model", 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp.905-911, 2021.

[6] Aditya Rao, Ankush Shetty, Aditya Uphade, Puneet Thawani, Priya RL, "A Proposal for a novel approach to analyze and detect the fake news using AI techniques", 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), pp.582-589, 2020.

[7] Somya Ranjan Sahoo, B.B. Gupta, "Multiple features-based approach for automatic fake news detection on social networks using deep learning", Journal of Ambient Intelligence and Humanized Computing, vol.13, no.7, pp. 799-809, 2022.

[8] Julio C.S Reis, Andre Correia, Fabricio Muria, "Supervised Learning for Fake News Detection", 2019 International Joint Conference on Neural Networks (IJCNN), pp.1-7, 2019.