Scope
Volume 13 Number 3 September 2023

A Polyglot : To Detect Misinformation Spread and Auto Populate Real News Using Natural Language Processing

Dr B. Jyoshna[1*] Geethika Palavarapu[2] Nampally Rakshitha[3] Nadigopa Niharika[4]
Department of CSE[1,2,3,4], Keshav Memorial institute of Technology, Hyderabad, India

Abstract:

The prevalence of fake news and misinformation in today's digital age necessitates the development of effective tools for detection and prevention. Our work presents a comprehensive framework for fake news detection, incorporating information retrieval, natural language processing (NLP), a model for prediction and checking URL reputation. Since most contemporary works on false news detection are written in English, their applicability has been restricted. The efficiency of language-agnostic feature transfer across many languages, demonstrating positive results. The framework offers a comprehensive solution for fake news detection, leveraging the advancements in information retrieval, NLP, and ML to tackle the challenge of combating misinformation in the digital realm. It offers multilingual inputs to make it available globally.

Keywords: Fake news, Detection ,Information retrieval, Natural language processing (NLP),Model for prediction, Checking URL reputation, Multilingual

1. Introduction:

Fake news has proliferated in the digital age, posing a serious danger to the integrity of online communication and the veracity of information. In addition to distorting public perception, the willful dissemination of false or misleading content creates serious security, political, and privacy dangers for the country. As a result, researchers, decision-makers, and technological professionals are paying more attention to the detection and mitigation of fake news. A comprehensive strategy that harnesses the capabilities of Natural Language Processing (NLP) and cyber security measures is necessary to effectively address this developing danger.

Our work presents a novel and comprehensive framework for detecting false news by combining cutting-edge NLP methods with effective cyber security measures. We aim to create a more resilient defence against the spread of false information and the possible weaknesses it exploits by combining the best aspects of both disciplines. Our strategy's initial phase focuses on using NLP algorithms to automatically identify and verify the truth of textual information. advanced text processing techniques. We try to accurately categorise content as trustworthy or misleading by extracting relevant information from news articles and social media posts.

However, it is clear that those who create false news have been using more sophisticated ways to avoid detection in the fast changing field of cyber deceit. The second component of our integrated strategy, which addresses this difficulty, includes Multi-Antivirus Scanning, Machine Learning, and Behaviour Analysis. We can detect probable harmful URLs by combining these approaches. This proactive approach enables us to protect consumers from hazardous and misleading online content, resulting in a more secure online environment.

Identifying those responsible for disseminating false information is a priority in our campaign to raise awareness of it. Once identified, we take steps to interact with them and stop the propagation of rumours by giving them real, verifiable news. By addressing inaccurate information where it originates, we hope to advance media literacy and create a more knowledgeable and responsible online community. By reducing the possibility of false positives and false negatives that can have unintended repercussions, the suggested integrated strategy has the potential to improve the reliability and accuracy of fake news identification. A proactive approach against the spread of false information is also made possible by the combination of NLP with cyber security, enabling both individuals and organisations to protect their online information ecology.
Using the synergies between NLP and Cyber Security, this research article aims to pave the way for a cutting-edge and interdisciplinary method to tackle fake news. We seek to promote a more secure and dependable digital world where truthful information predominates and deceptive methods have a limited impact by combining the skills of language processing and threat analysis. Our model is designed to accommodate input in various languages, allowing users to interact with it using their preferred language. It is equipped to understand and respond appropriately in multiple languages, ensuring a seamless communication experience. The results generated by the model will be provided in the same language as the input, catering to the linguistic preferences and needs of our diverse user base. This flexibility in language usage enhances accessibility and usability, making the model a valuable tool for communication across linguistic boundaries.

2. Literature Survey:

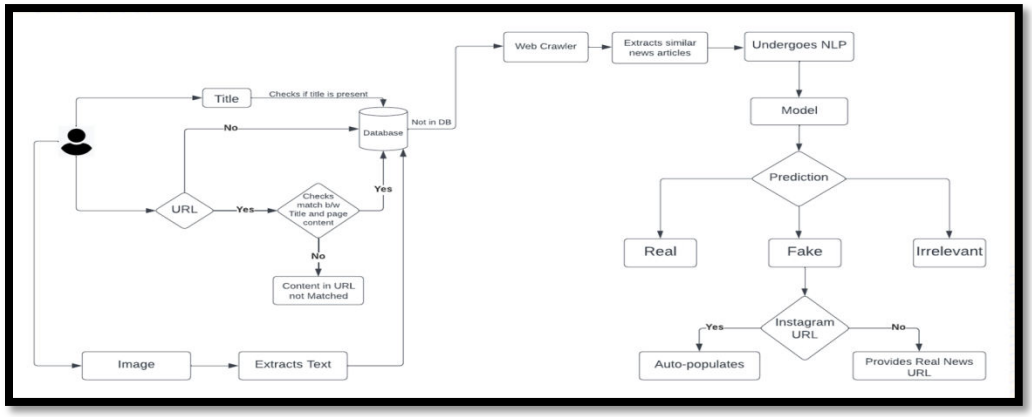
Natural-language processing (NLP) is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, and how to program computers to fruitfully process large amounts of natural language data
Word segmentation is a fundamental pre-processing step for Natural Language Processing, but current off-the-shelf solutions lack consistent benchmarking and the state-of-the-art deep learning system is slow and does not utilise sub-word structures. Research indicates that individuals with lower credibility, including malicious accounts or susceptible regular users, are more prone to disseminating fake news. To gauge the likelihood of spreading false information, user credibility scores can be utilized as a measure of trustworthiness.[12]
The proposed framework highlights the importance of considering the social context in detecting fake news on social media platforms. By utilizing the tri-relationship among publishers, news articles, and users, the framework aims to improve the accuracy of fake news classification.[3] Instances of fake news are often followed by fact-checks published on different media outlets.[10] Malicious users take advantage of online platforms by generating and spreading fake news to damage the reputation of individuals, businesses, and politics. [14] The motive behind spreading fake news is often to mislead readers, damage the reputation of entities, or gain from sensationalism. It is considered one of the greatest threats to democracy, free debate, and the Western order.[5] The earliest form of online social networking emerged through email, allowing individuals to share and exchange information using distinct email addresses. The rise of smartphones introduced popular social network applications like Facebook, Twitter, Snapchat, Tumblr, and Instagram, but also led to the proliferation of spammers within these platforms Irrelevant and redundant features adversely affect classifier accuracy and performance. To mitigate this, feature reduction techniques, such as excluding common words like "the," "and," "there," and setting a threshold for word frequency, are employed, including using a limited word set, lowercasing, and stop word removal. The proliferation of fake news has significant detrimental effects on individuals and society [16]. It disrupts the authenticity equilibrium within the news ecosystem, deliberately influencing consumers to adopt biased or false beliefs and altering their perception and response to genuine news, particularly in the realm of political information, with social media exacerbating the rapid spread of fake news.[9]

The main idea behind fake news is to manipulate the emotions and thinking of humans to make them believe something that isn't true. Mainstream social media platforms, including Facebook, WhatsApp, Twitter, Instagram, etc., are sources of such fake news. [11]
Misinformation refers to the dissemination of false information, disregarding the true intent. It arises from factors like mislabeling, inadequate fact-checking, and apathetic users who prioritize sharing over verifying the accuracy of what they read.[12].

3. Proposed Misinformation Detection Framework:

Our model is designed to detect the authenticity of news articles across different languages. It goes through several phases to determine whether the news is fake or real.Four unique phases make up the fake news detection implementation: news gathering, predicting, automatically supplying actual news to the false news spreader, and determining whether a given URL is malicious or not. Information retrieval (IR), natural language processing (NLP), and machine learning (ML) modules are included.
Web crawlers work concurrently in the first stage to gather information from numerous websites and social media platforms. The preprocessed data is subsequently fed into a machine learning model for the detection of bogus news. The IR module is in charge of web crawling during the data collection process and collecting news data from pertinent websites. The domain corpus for the following NLP module is made up of the obtained news data.
The user query serves as the entry point for getting data throughout the data collection stage. The system uses web crawlers to fetch and retrieve a list of related news articles for each news query. The featured data from this list of pertinent news is then analyzed and provided to the machine learning model. Every news query essentially serves the same purpose as a user inquiry, causing the web crawler to fetch and obtain a news list that is pertinent to the query.
The NLP module enters the picture when it receives the news material from the list that was collected and carries out several operations such as text segmentation, cleaning, and feature extraction. These procedures give the module the ability to successfully process the news information and extract valuable features that support the machine learning model.

Figure 1: The Flow of proposed work



3.2 Information Retrieval:

Information retrieval is essential to this paradigm because it makes it possible to access online news content, including both real and fraudulent news. Using a web crawler-based information retrieval system to gather news content from online sources is a successful strategy.

We have developed a proposed information retrieval module based on web crawling, which consists of two primary processes: feature extraction and news gathering. To gather relevant information from a multitude of news sources on the web, we employ web crawlers, automated agents, or robots. By carefully considering the vast amount of available data during this process.As a result of our web crawler-based information retrieval procedure, our model obtains a comprehensive list of news stories. Each news item in the list is selected based on its resemblance to the specified query. This ensures that the retrieved news items closely match the user's search criteria.

During the information retrieval phase, users can submit an image containing a headline. Our model extracts the text from the image using OCR and forwards it to web crawlers. These crawlers search for web pages that match the extracted text. By incorporating OCR, users can provide images as input, expanding the range of media for retrieving information.

3.3 Natural Language Processing (NLP):

By carrying out a number of crucial activities, Natural Language Processing (NLP) plays a crucial role in the data preparation stage. Word segmentation, data purification, stop-word removal, feature extraction, and word indexing are some of these activities. Clean text is produced from news content as part of the data preparation process, which is subsequently used in the feature extraction procedure.

Our information retrieval mechanism, which gathers user queries and delivers them to web crawlers to get pertinent news articles from the internet, provides data to the feature extraction module. The data preparation phase receives the news content that was retrieved during the query procedure.

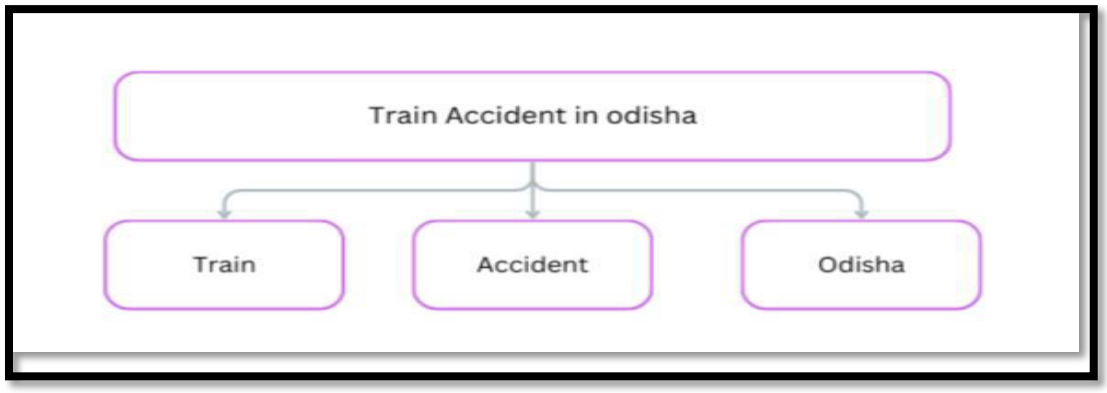


Figure 2: Process of tokenization

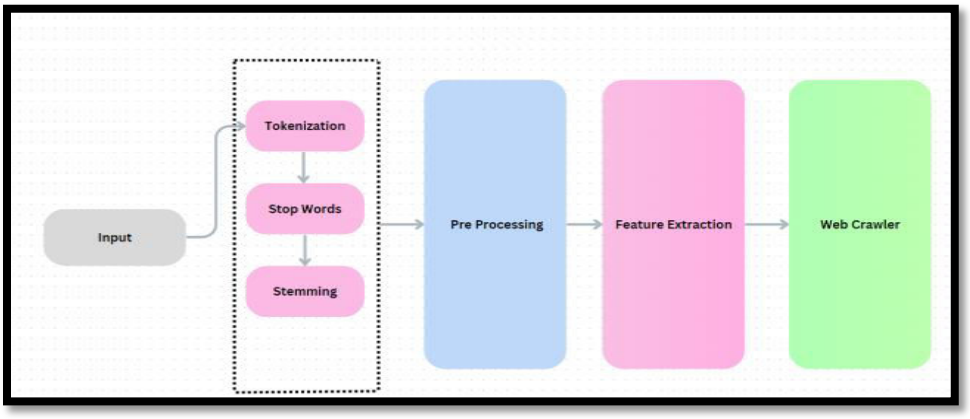


Figure 3: Removing Irrelevant data

According to Fig 2 and 3, we have implemented word tokenization and utilized a stopwords library to tokenize sentences and remove stopwords. This approach allows us to break down sentences into individual words or tokens efficiently. Additionally, by removing stopwords, which are commonly occurring words with little contextual meaning, we can focus on more significant terms within the text. This process has been successfully incorporated into our system to enhance text processing and analysis.

3.4 Developed Model:

News stories can be categorized as fake, legitimate, or irrelevant using our model. Our model is essential for automatically analyzing news content and establishing its veracity in the context of fake news identification. The NLP module processes textual data by carrying out operations such word segmentation, data purification, stop-word removal, and feature extraction. These models accept feature data extracted from the NLP module.

Our model compares the user query with the queries acquired from the web crawler during the information retrieval phase after obtaining the feature data. The goal of this matching procedure is to locate news items that are pertinent to the user's inquiry. Our model takes into account numerous aspects, including the existence of keywords and contextual information, when evaluating the similarity or relevance between the user query and the retrieved news items.

Our model then categorizes each news stories into one of the three predetermined categories after the matching process. The system can automatically analyze news information, compare it to user searches, and categorize it. This makes it possible to recognize and identify bogus news, assisting people in making defensible decisions and preventing the spread of false information.

3.5 Auto-populating:

Auto-populating the real news in response to a classified fake news post on Instagram involves utilizing web crawling and comment generation techniques. Once a news article is classified as fake using machine learning

models, the system can employ web crawling to search for and retrieve the corresponding real news article from credible sources.In this case, the web crawler is directed to search for the real news article related to the fake news being addressed. The crawler starts from a designated entry point, such as a search engine or a specific news website, and finds the relevant information.Upon retrieving the real news article, our model can then generate a comment or response to be posted under the fake news post on Instagram. The comment is composed using the content extracted from the real news article, providing accurate and reliable information to counteract the false claims made in the fake news.The generated comment can include a summary of the real news article, highlighting the facts and evidence that contradict the false claims in the fake news. It can also provide links or references to the original source for users to verify the authenticity of the information. By posting this comment under the fake news post on Instagram, it aims to raise awareness among users and provide them with

3.6 Evaluating the Safety and Legitimacy of URLs:

Using an API created expressly for threat intelligence or URL analysis, it is possible to determine whether a given URL is harmful or not. Such an API offers a thorough evaluation of the URL's legitimacy and safety based on a number of security factors. Our system can send an API request using the URL as input to carry out this action. After that, the API examines the URL using sophisticated algorithms and databases that have data on sites that have been banned, known malicious patterns, phishing attempts, malware dissemination, and other indicators of bad intent.

The API may go through several rounds of analysis, including looking at the URL's structure, comparing it to known dangerous patterns, searching reputation databases, using heuristic analysis, and evaluating other pertinent characteristics. These evaluations assist in determining whether the URL poses a possible security issue.

An assessment or risk score reflecting the chance that the URL is dangerous is often included in the API response. Additionally, it might include information about the type of danger, if malware is present, or any associated suspicious activity. Making conclusions concerning the URL's security is made easier with the help of this information.

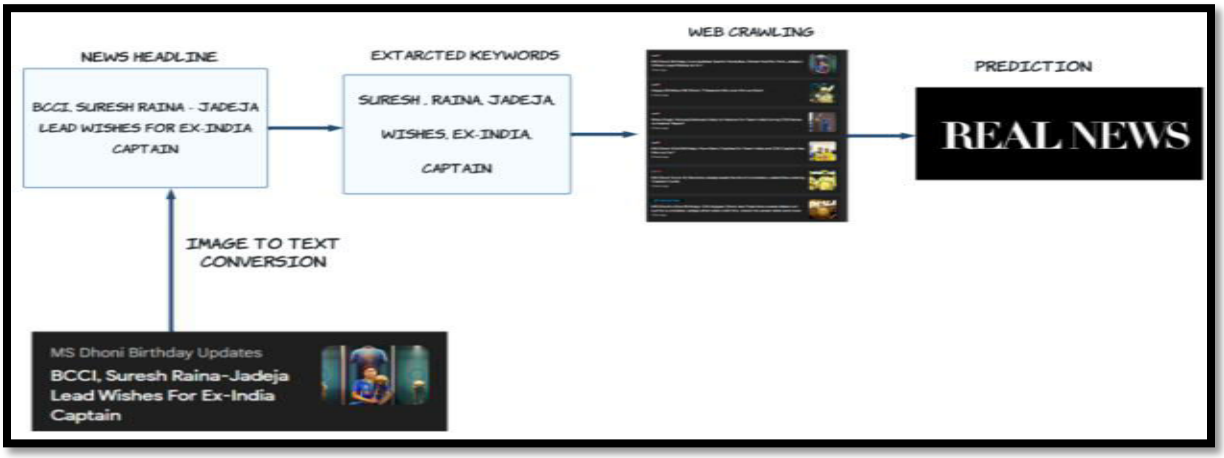


Figure 5: Visualization of the process

Results and Discussions:

The model is now able to recognize fake news as well as real news. Our model uses a thorough methodology that is divided into four phases: news gathering, prediction, automatic synthesis of real news for false news spreaders, and identification of harmful URLs. To provide reliable results, each phase includes elements of information retrieval (IR), natural language processing (NLP), and machine learning (ML).Web crawlers gather information from many websites and social media platforms concurrently in the initial stage. The preprocessed data is then sent into the ML model, which is in charge of identifying bogus news, after being collected. User queries are used as input during the data collecting phase to start the retrieval of data.

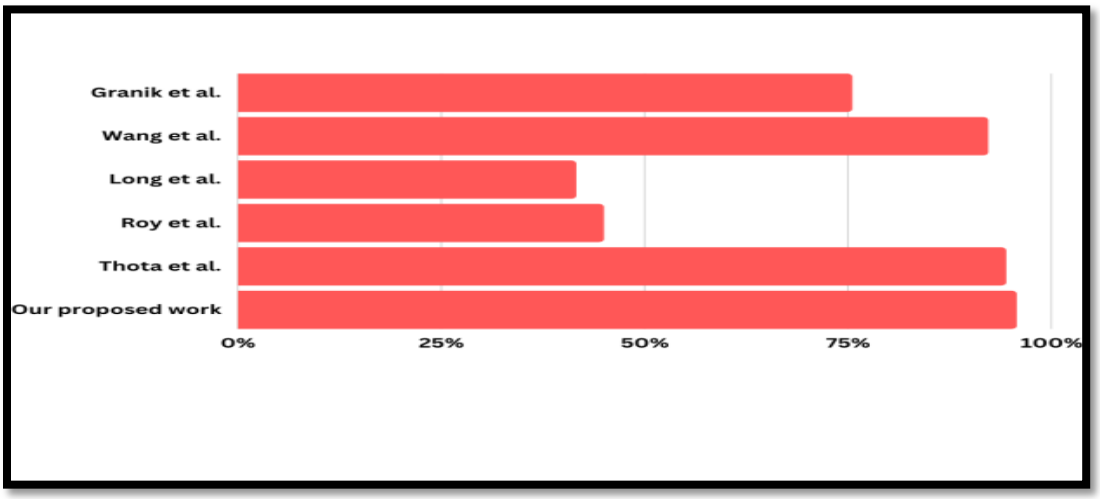


Figure 5: Comparative analysis between the proposed work and related work

Figure 6: An illustration of the truth and reliability of the news

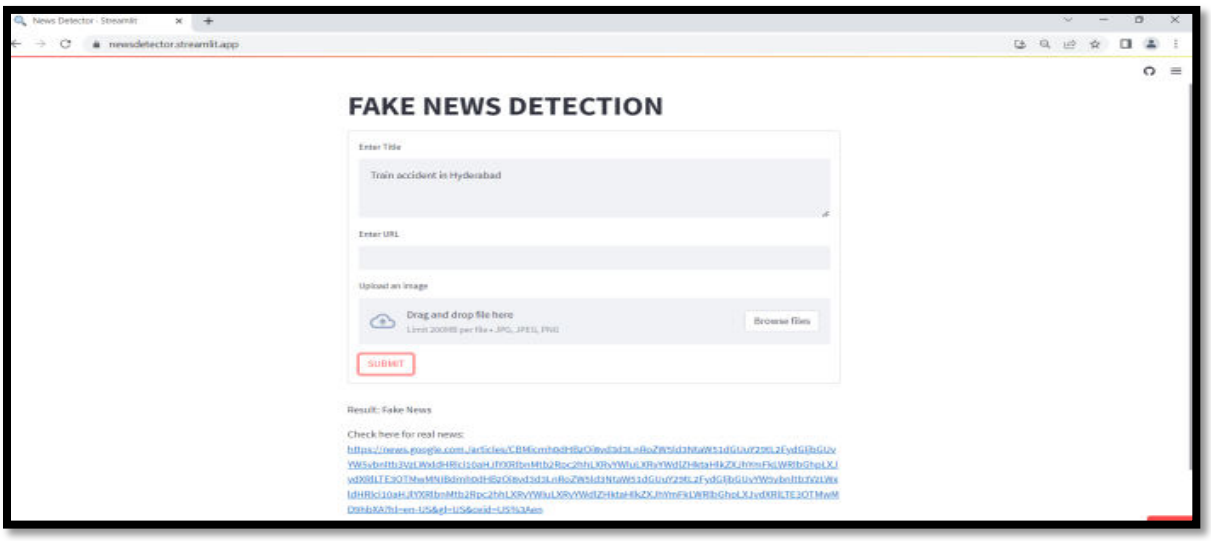


Figure 7: An illustration of fake news and a link to the real news

The IR module is essential to web crawling since it collects news data from pertinent websites and creates the domain corpus for the next NLP module.

learning (ML).Web crawlers gather information from many websites and social media platforms concurrently in the initial stage. The preprocessed data is then sent into the ML model, which is in charge of identifying bogus news, after being collected. User queries are used as input during the data collecting phase to start the retrieval of data. Web crawlers collect a list of news articles that are relevant to each inquiry. Relevant news stories are selected from this list, analyzed, and delivered to the ML model. In essence, every user enquiry serves as an inquiry that causes the web crawler to retrieve a list of pertinent news.

Table 1: Comparison Analysis of Proposed Work

Authors	**Data Source**	**Features Used**	**Labels**	**Multi-Language**
Granik et al.	Facebook posts	Posts texts	Mostly true,Mostly false	No
Wang et al.	News website and Kaggle	Text, images and title	Real and Fake	No
Long et al.	LIAR Dataset	Speaker information and topic	pants-fire,barely-true,half-true , mostly-true and true	No
Roy et al.	LIAR Dataset	Texts and speaker information	pants-fire,barely-true,half-true , mostly-true and true	No
Our proposed work	**News Websites**	**Texts , images , Title and URL**	**Real,Fake and Irrelevant**	**Yes**

When the collected news articles are delivered, the NLP module gets to work, carrying out tasks like text segmentation, cleaning, and feature extraction. These procedures give the module the ability to properly process the news data and extract useful features that support the predictions of the ML model.

Our model seeks to deliver precise and dependable detection of fake news across many languages by utilizing the power of web crawling, NLP methods, and ML algorithms. But it's vital to keep in mind that assessing the reliability of news pieces still requires critical thinking and verification from several sources.

We are able to receive precise results for the user inquiry because this procedure is being used in our project. We are also able to automatically auto-populate the genuine news to the false news spreader and determine the percentage of true news in it. Our work can be used anywhere in the world because it can accept any language.It even states whether or not the specified URL is malicious.

Authors	Main Idea	Model	Metrics	Accuracy
Granik et al.	Classify fake news based on text words	Naive Bayes	Accuracy	75.4%
Wang et al.	Analysis text and images, by cnn.	TI-CNN	F1-score	92.1%
Long et al.	Speaker profiles information	LSTM + Attention	Accuracy	41.5%
Roy et al.	Statement information and Speaker Profile information	CNN +Bi-LSTM	Accuracy	44.9%
Thota et al.	Find the relation between article and its headline	DNN + TF-IDF	Accuracy	94.31%
Our proposed work	Analyses text in multi-language images . Feature extraction and auto-populating.	NLP	Accuracy	95.6%

Table 2: Accuracy Analysis Matrix

Conclusion:

Internet usage has increased as a result of new computing trends (such as mobile cloud computing, wi-Fi, and applications for smart devices). Sharing information (text, audio, and video) is becoming simpler as a result. Recently, it has come to light that using various online platforms to disseminate misleading information and fake news plays a vital role in achieving a number of objectives, manipulating the stock market, or just sowing unwelcome feelings among online platform users, such as wrath and hatred. As a result, there is a rising need for automated systems that can accurately and effectively detect bogus news.Online communication and information integrity are seriously threatened by the spread of fake news in the digital age. A comprehensive approach that makes use of both cyber security measures and Natural Language Processing (NLP) is needed to combat this issue. We efficiently detected and stopped the propagation of fraudulent information by combining cutting-edge NLP algorithms for determining and confirming the veracity of textual content with proactive cyber security techniques like Multi-Antivirus Scanning, Machine Learning, and Behavior Analysis. A more informed and responsible online community has been created by conversing with people in charge of spreading fake news and encouraging media literacy. The combination of NLP with cyber security offers a

potent multidisciplinary strategy for combating false information, building a more trustworthy and safe digital environment where truthful information prevails.

References:

1. Chormai, Pattarawat, Ponrawee Prasertsom, and Attapol Rutherford. "Attacut: A fast and accurate neural thai word segmenter." arXiv preprint arXiv:1911.07056 (2019).
2. Shu, Kai, et al. "Mining disinformation and fake news: Concepts, methods, and recent advancements." Disinformation, misinformation, and fake news in social media: Emerging research challenges and opportunities (2020): 1-19.
3. Shu, Kai, Suhang Wang, and Huan Liu. "Beyond news contents: The role of social context for fake news detection." *Proceedings of the twelfth ACM international conference on web search and data mining*. 2019.
4. Meesad, P. "Thai Fake News Detection Based on Information Retrieval, Natural Language Processing and Machine Learning. SN COMPUT. SCI 2, 425 (2021)." (2021).
5. Shahbazi, Zeinab, and Yung-Cheol Byun. "Fake media detection based on natural language processing and blockchain approaches." *IEEE Access* 9 (2021): 128442-128453.
6. Ahmed, Sajjad, Knut Hinkelmann, and Flavio Corradini. "Development of fake news model using machine learning through natural language processing." arXiv preprint arXiv:2201.07489 (2022).
7. [9] Ireton, Cherilyn, and Julie Posetti. Journalism, fake news & disinformation: handbook for journalism education and training. Unesco Publishing, 2018.
8. Sharma, Sunidhi, and Dilip Kumar Sharma. "Fake News Detection: A long way to go." 2019 4th International Conference on Information Systems and Computer Networks (ISCON). IEEE, 2019.
9. Nasir, Jamal Abdul, Osama Subhani Khan, and Iraklis Varlamis. "Fake news detection: A hybrid CNN-RNN based deep learning approach." International Journal of Information Management Data Insights 1.1 (2021): 100007.
10. Chormai, Pattarawat, Ponrawee Prasertsom, and Attapol Rutherford. "Attacut: A fast and accurate neural thai word segmenter." arXiv preprint arXiv:1911.07056 (2019)..
11. Mookdarsanit, Pakpoom, and Lawankorn Mookdarsanit. "The covid-19 fake news detection in thai social texts." Bulletin of Electrical Engineering and Informatics 10.2 (2021): 988-998.
12. Shu, Kai, et al. "Mining disinformation and fake news: Concepts, methods, and recent advancements." Disinformation, misinformation, and fake news in social media: Emerging research challenges and opportunities (2020): 1-19.
13. Collobert, Ronan, et al. "Natural language processing (almost) from scratch." Journal of machine learning research 12.ARTICLE (2011): 2493-2537.

14. Granik, Mykhailo, and Volodymyr Mesyura. "Fake news detection using naive Bayes classifier." 2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON). IEEE, 2017.
15. Wang, William Yang. "" liar, liar pants on fire": A new benchmark dataset for fake news detection." *arXiv preprint arXiv:1705.00648* (2017).
16. Long, Yunfei, et al. "Fake news detection through multi-perspective speaker profiles." *Proceedings of the eighth international joint conference on natural language processing (volume 2: Short papers)*. 2017.
17. Roy, Arjun, et al. "A deep ensemble framework for fake news detection and classification." *arXiv preprint arXiv:1811.04670* (2018).

    Thota, Aswini, et al. "Fake news detection: a deep learning approach." *SMU Data Science Review* 1.3 (2018): 10.