Winter 12-10-2023

Cloud Optimization for Disinformation Detection and News Veracity

Lumbardha Hasimi
Lodz University of Technology

Aneta Poniszewska-Marańda
Lodz University of Technology, aneta.poniszewska-maranda@p.lodz.pl

Cloud Optimization for Disinformation Detection and News Veracity

Lumbardha Hasimi
Institute of Information Technology,
Lodz University of Technology, Poland

Aneta Poniszewska-Marańda[1]
Institute of Information Technology,
Lodz University of Technology, Poland

ABSTRACT

The recent advancements in technology and the widespread availability of information, has made it easier to reach massive audiences. However, the issue of fake news has reached a breaking point. It not only harms online social networks and news sites but also negatively impacts offline communities. Over the past few years, researchers have been challenged by the dangerous influence of fake news on politics, culture, and lifestyle, and now with the COVID-19 pandemic, the danger has extended to health and social well-being. Immediate action is necessary to counteract this problem. Therefore, the goal of this paper is to explore possible solutions to the problem of fake news and develop a suitable, effective, and user-friendly application that can identify disinformation and fake news by optimizing cloud-based tools. To achieve this, various papers and databases were analyzed, and it was concluded that a cloud-hosted web application and machine learning classifier would be a practical solution. The proposed model was implemented, and the results showed an accuracy rate of 93%.

Keywords: cloud computing, cloud-based detector, disinformation, fake news detection, optimization.

INTRODUCTION

With the advancement of technology, we started relaying more and more on decentralized sources of information than ever before. That means that access to information is

[1] Corresponding author. aneta.poniszewska-maranda@p.lodz.pl +48 426312796

easy, but assessing its credibility is both time-consuming and challenging. This results in such assessment being frequently overlooked, which in result makes such fake information even more widespread. To combat that a system supporting such verification process is needed. While fake news may not appear immediately dangerous, assessing their potential harm depends on various factors. The high intention to deceive the public, in particular, can lead to significant dangers (Nieminen and Rapeli, 2018). Not checked information about medical illnesses can lead to incorrect treatment and unverified claims can lead to serious consequences (Phippen et al. 2021, Rohman et al. 2021, Jain and Kasbe 2018, Vinhas and Bastos 2022, Paor and Heravi 2020, Gerbina 2021, Alharbi 2021).

The pressing need for technology that helps assess the authenticity of information is evident. In today's information-rich environment and easy access, there is a clear demand for advanced tools that can effectively assess the veracity of information. Moreover, news websites are a vast market. USA Today Network – one of the largest news media in US had 145 022 000 unique visitors in November 2021 alone. That shows that there is not only need but actual huge market for such solution. The project has two main challenges. Classifying given information as real or fake, and doing it both quickly enough and in the easy-to-use way, for both individual users, as well as news platforms (Pew Research 2023). System needs to have high reliability – it needs to classify information correctly – because otherwise there is no point in using it. It also needs to generate those results quickly – even though news media may wait for couple seconds, or even a minute before posting some news to check its authenticity, normal users will not have such patience.

Lastly, the ease-of-use. No individual user will bother to consider our system if learning how to use it, or using is more difficult than just checking authenticity manually. Similarly, no

business partner will bother with integration of our services if the difficulty of their use is disproportionate to the results. The problem is to be solved with simple, layered web application hosted in cloud (Pew Research 2023, Asr and Taboada 2023). Individual users will be able to use simple website, connected to main system through Web API, where after pasting piece of information they will see appropriate message, whereas businesses will be able to use Public API and integrate it with their existing systems. Both Web API and Public API will respond based on the results of classification from machine learning model. Cloud sources optimized for a real-time solution against fake news dissemination, bring advantages as compared to other environments (Göksu et al., 2020). By leveraging cloud computing, it is possible to scale up the detection infrastructure as the volume of data grows. This ensures that the detector can handle large volumes of data while allowing the fake news detector to quickly analyze data and provide results in real-time. Overall, cloud computing can provide a scalable, efficient, and effective way to detect fake news and disinformation with the use of powerful computing resources, machine learning algorithms, NLP, and data integration (Alarfaj and Khan, 2023).

METHODOLOGY AND THE PROPOSED ARCHITECTURE

While discussing methods of fact-checking, it is important to mention at the outset the ways in which news can be independently verified by users. There exists a set of criteria to which one should pay attention, for example, when choosing a news website. There are, among others, authority (like contact and owner information, author credentials), purpose, coverage (for example links to verifiable sources), timeliness (when the content was published or edited), objectivity, accuracy. These types of criteria can also be used as factors in automatic news verification (Comscore 2023). Moving on to machine learning one must mention word frequency. Many of methods and algorithms used in projects based on machine learning models

examine the frequency with which particular words occur and compare them with the frequency of occurrence in other articles on the same topic, thus checking the credibility of the news item on this basis.

While analyzing the types of algorithms to conduct the classification, we focused on two specific approaches: TF-IDF and the passive-aggressive classifier (Keita, 2021; Palriwala, 2020). This mainly because TF-IDF, is considered as a well-established algorithm, that excels at capturing the importance of terms, providing therefore valuable insights into the significance of words in the classification process (Ahuja et al., 2019). Similarly, passive-aggressive classifier is known for the ability to make quick updates to the model based on incoming data which is particularly advantageous in dynamic environments, making it well-suited for real-time or evolving datasets.

While dealing with textual data, that data first needs to be converted to a vector of numerical data by a process known as vectorization. This is a measure that measures the meaning or relevance of words or phrases in a document quantitatively within a set of documents. TF-ID vectorization calculates TF-IDS score for each word in a document and puts it into a vector. Thanks to this, it is relatively easy to compare checked article with a set of similar articles. TF-IDF is simple and handy algorithm popularly used as it is easy to use and computationally cheap. Unfortunately, it also has cons, TF-IDF cannot deal with semantic meaning, it ignores word order and can suffer from memory inefficiency (Keita, 2021, Dentith 2016). When it comes to passive-aggressive algorithms, they are family of algorithms used for large-scaled learning. They make updates to correct the loss that was found in a model – we should keep the model if the classification is correct (passive), but if the classification is wrong, we adjust the model to recognize misclassified examples better.

The proposed system focuses on classifying truthfulness of short fragments of text (around 160 characters) related to covid-19. The system provides two ways of interaction with the machine learning model (Fig. 1):

- Website where user can verify the text by pasting it in the text field.
- Public API which can be utilized in other systems by our business partners.

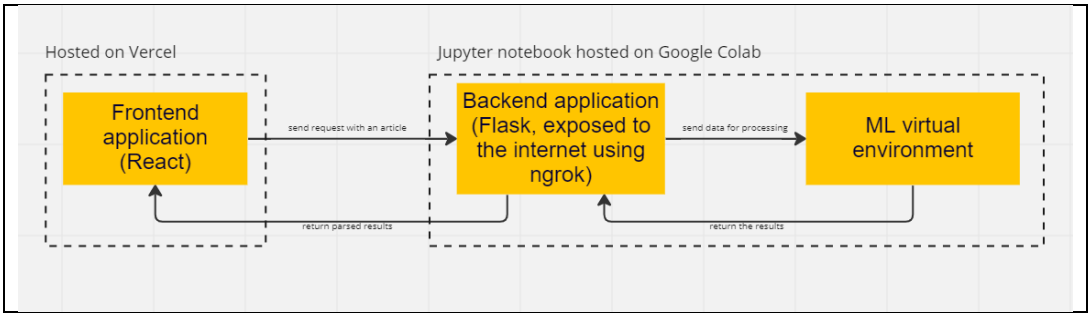

Figure 1. Deployment architecture of proposed solution of cloud-hosted web application and machine learning classifier

TOOLS AND TECHNOLOGIES

The tools and technologies used for the development of proposed solution for cloud optimization for disinformation detection and news veracity are characterized by the following features (Fig. 2):

Containerization: Docker is the most popular solution for containerization. Bundling components of the system into Docker containers provide consistent environment for each member of the development team and simplify deployment of the application in the cloud (Antoun et al. 2020, Çaglayan et al. 2022).

Machine Learning: Python is one of the most popular languages used in ML (Antoun et al. 2020, Çaglayan et al. 2022, Donthu et al. 2021, Radhakrishnan et al. 2017, Saini et al. 2022, Trujillo and Long 2018, van Eck and Waltman 2010). It offers great choice of ready to use, well-documented libraries, especially Scikit-learn for handling basic ML algorithms and NLTK for

working with natural language processing. For creating network API for machine learning model used we used Django framework. Using the same language as in ML model simplify integration and deployment of these two components.

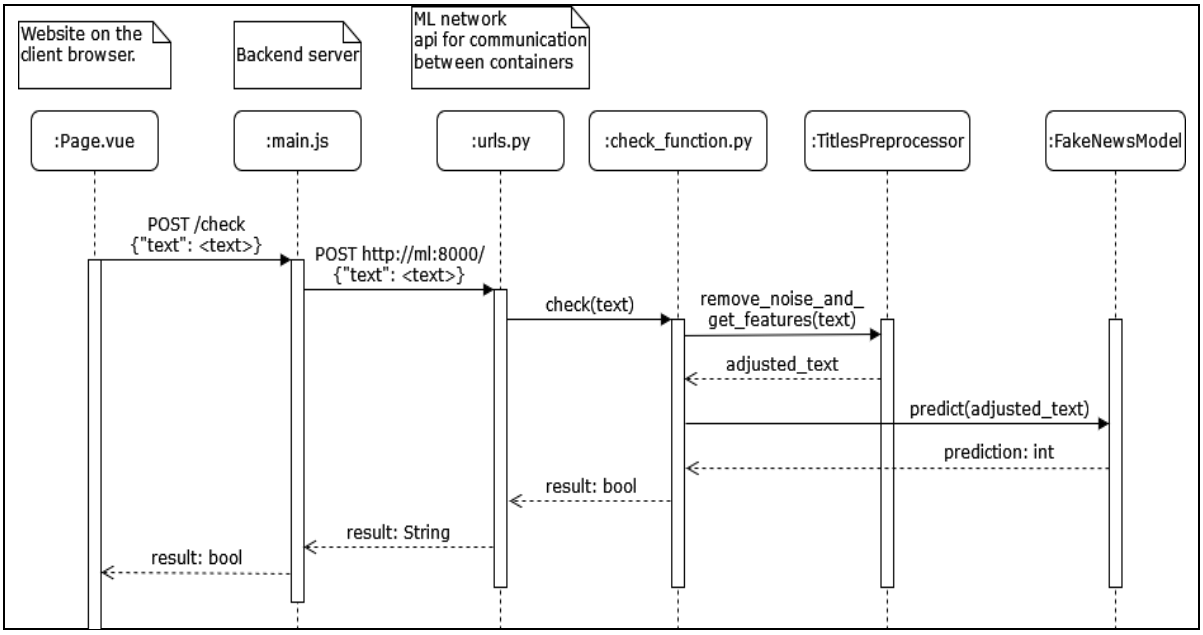

Figure 2 Main workflow of cloud-hosted web application and machine learning classifier

Backend: Node.js with Express framework is used on backend side of our application. Using this technology allows to serve static files which eliminates need for external HTTP server and reduce complexity of the system. *Frontend*: Vue.js is a framework well-suited for small projects. It is smaller in size compared to React or Angular (Reality Defender 2023). *Hosting*: Proposed application is lightweight and do not require much computing power, thus the main factors in choosing cloud provider are pricing and server location.

RESULTS AND DISCUSSION

In order to facilitate the process of fake news classification, as well as provide room for possible expansion in the future, our system is divided into three main, layered components (Fig. 3):

- *Machine learning model* – responsible for classifying given piece of information as real or fake data.
- *Node.js backend server* – responsible for maintaining both Public and Web API.
- *Vue frontend application* – responsible for providing user interface for our users.

It was decided to deploy each of the component as separate Docker container to improve modularity and scalability, as well as make whole project more lucid. Such approach also has other benefits – it makes the deployment process much easier (Wei and Zhang 2020, Zare-Farashbandi et al. 2014, Abbasi et al. 2022, Khivasara et al. 2020, Paschalides et al. 2019).

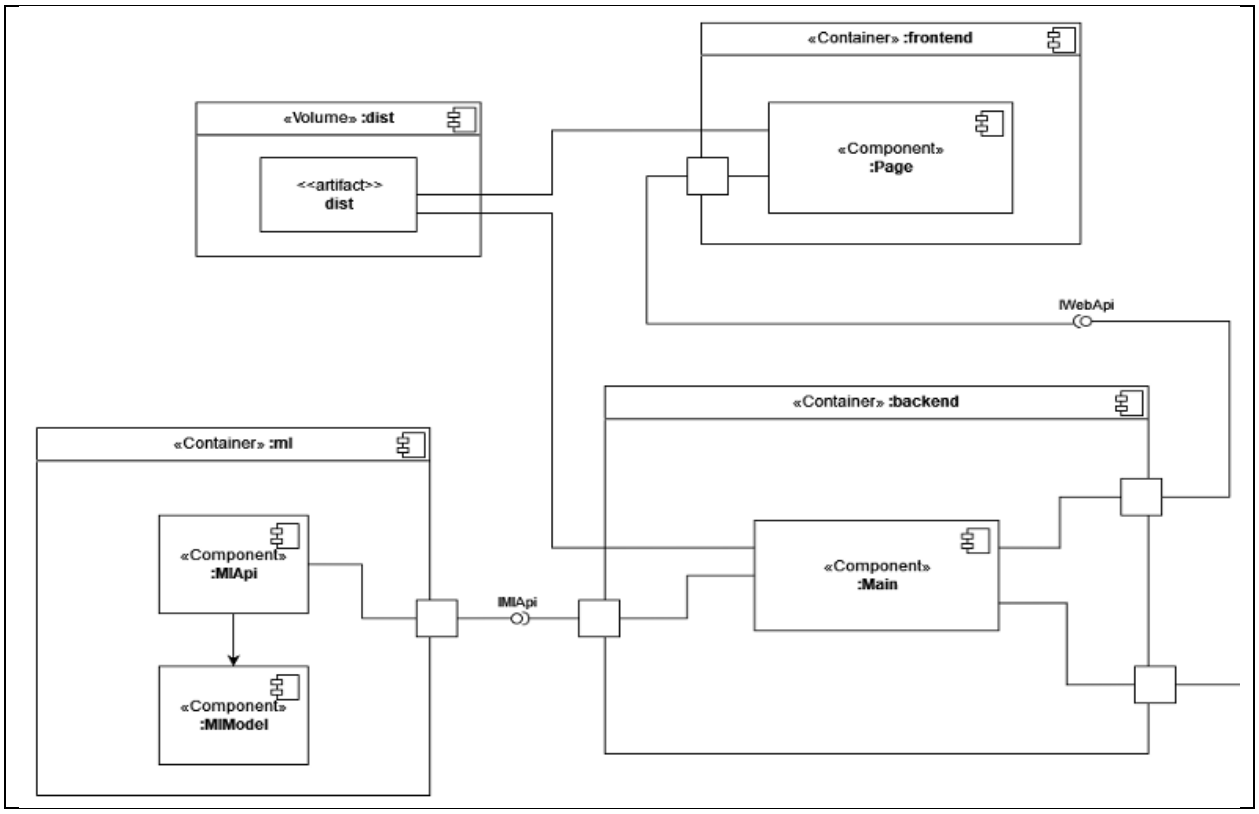

Figure 3. Architecture of proposed solution model

The application consists of three main modules: (1) Frontend application, (2) Backend server, (3) Machine learning model. Each module run in separate Docker container and communicate through HTTP protocol. Frontend module is implemented in form of single page

application. It provides graphical interface to interact with the system. When a user clicks on "Check!" button (Fig. 4) content of the text field is sent in POST request to check endpoint of the backend server. Relevant popup is displayed based on respond from the backend server. The main purpose of the backend module is to host the website and serve as middleware between frontend and ML modules. It is the only module which is exposed outside the VM. As a result, whole network communication with the system is performed through this module.

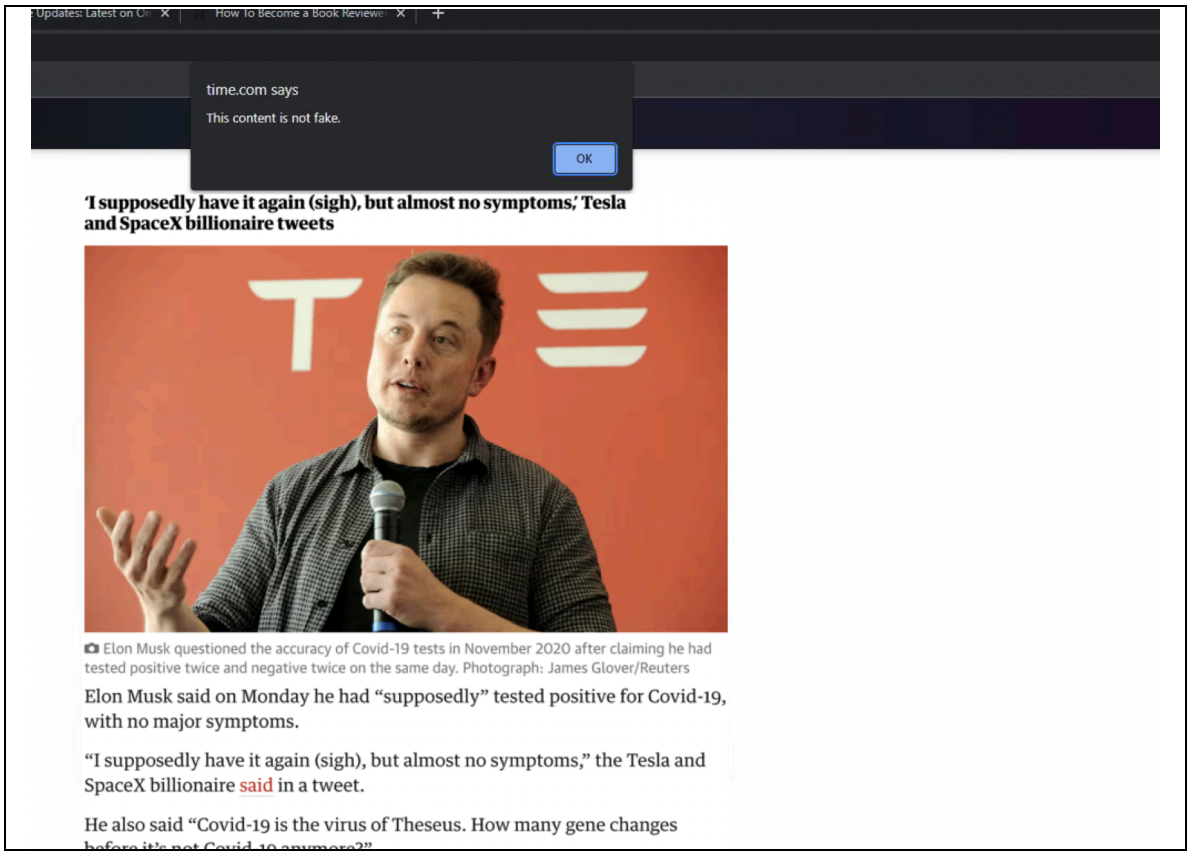

Figure 4. Frontend of proposed solution for cloud optimization for disinformation detection and news veracity

The machine learning module is the main part of the application. Evaluation is performed in following steps (Tab. 1):

- Unnecessary words and special characters are removed from the text.

- Pre-processed text is transformed into numerical representation using precomputed vectorizer.
- Vectorised text is used by classifier to predict if the news is real or fake.

This module includes API component which provides interface for convenient communication with backend module despite difference in programming languages (Song et al. 2021, Securing Cloud Functions 2023, The Factual 2023).

Table 1. Results and accuracy achieved in different datasets

Dataset	Accuracy Train_label	Accuracy Test_label	Size
WELFake (training dataset)	93.76%	91.46%	72134 articles
Fake news from Kaggle Competition	93.2%	89.3%	20387 articles
Fake News by Hassan Asim	88.81%	72.63%	6060 articles

CONCLUSIONS

Based on the research and examination of existing solutions, the most crucial feature of the information detection is the time taken to review the authenticity of a content. Many applications offer the ability to read articles already checked, but it is important for the user to be able to check independently the content read. The most intuitive way for many users is to paste a link with found article into the search box and wait for reliability result. The TF-IDF and passive-aggressive classifiers allow text extraction and word frequency checking, which are verified and widely used fact-checking methods. Such methods are in particular effective considering the simple and transparent enough process of use. Aiming to tackle the issue of fake news through different technologies with the use of cloud systems and resources we proposed a detector that incorporates three layered components. The main objective was to explore possible solutions to the problem of fake news and develop a suitable, effective, and user-friendly application that can identify disinformation and fake news by optimizing cloud-based tools. To

achieve this, various papers and databases were analyzed, and it was concluded that a cloud-hosted web application using machine learning classifiers is the most practical and efficient solution. The proposed model was implemented, and the results showed an accuracy rate of 93%, while being tested in three different datasets.

Considering various aspects of the fake news and disinformation detection issue, cloud computing offers promising avenue for scalable and efficient solutions. Particularly when considering the computational resources, accessibility and the opportunities for real-time analysis of massive datasets. Cloud-based systems are scalable and flexible, allowing for the integration of various machine learning algorithms and data analysis techniques to identify patterns and anomalies in social media data and news articles. The collaborative options of the platforms further support and simplify the use of natural language processing and sentiment analysis techniques to identify misleading and biased content. Seemingly, cloud computing offers a powerful platform for developing and deploying effective solutions to combat the growing threat of fake news and disinformation.

REFERENCES

Abbasi, A., Javed, A.R., Iqbal, F., Kryvinska, N., and Jalil, Z. 2022. "Deep learning for religious and continent-based toxic content detection and classification," *Scientific Reports*, (12:1), pp. 1-12.

Ahuja, R., Chug, A., Kohli, S., Gupta, S., and Ahuja, P. 2019. "The Impact of Features Extraction on the Sentiment Analysis," *Procedia Computer* Science, (152), pp. 341–348.

Alarfaj, F.K., and Khan, J.A. 2023. "Deep Dive into Fake News Detection: Feature-Centric Classification with Ensemble and Deep Learning Methods," *Algorithms*, (16), 507.

Alharbi, R., Vu, M.N., and Thai, M.T. 2021. "Evaluating Fake News Detection Models from Explainable Machine Learning Perspectives," *Proceedings of the IEEE International Conference On Communications* (ICC 2021).

Antoun, W., Baly, F., Achour, R., Hussein, A., and Hajj, H. 2020. "State of the Art Models for Fake News Detection Tasks," *Proceedings of 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies* (ICIoT), Doha, Qatar, pp. 519-524, doi: 10.1109/ICIoT48696.2020.9089487.

Asr, F.T., and Taboada, M. 2023. "Discourse Processing Lab," [Online]. Available: http://fakenews.research.sfu.ca/. Last accessed: Jan 11, 2023.

Çaglayan Akay, E., Yılmaz Soydan, N.T., and Kocarık Gacar, B. 2022. "Bibliometric analysis of the published literature on machine learning in economics and econometrics," *Social Network Analysis and Mining*, (12:109) https://doi.org/10.1007/s13278-022-00916-6.

Comscore, Inc, "Rankings - Comscore, Inc." [Online]. Available: www.comscore.com/Insights/Rankings. Last accessed: Sept 20, 2023.

Dentith, M.R.X. 2016. "The Problem of Fake News," *Public Reason*, (8:1-2), pp. 65-79.

De Paor, S., and Heravi, B. 2020. "Information literacy and fake news: How the field of librarianship can help combat the epidemic of fake news," *Journal of Academic Librarianship*, (46:5), pp. 102218.

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., and Lim, W.M. 2021. "How to conduct a bibliometric analysis: An overview and guidelines," *Journal of Business Research*, (133), pp. 285-296, doi: 10.1016/j.jbusres.2021.04.070.

Gerbina, T.V. 2021. "Science Disinformation: On the P roblem of Fake News," *Scientific and Technical Information Processing*, (48:4), pp. 290-298.

Göksu, M., Cavus, N., Cavus, A., and Karagozlu, D. 2020. "Fake News Detection on Social Networks with Cloud Computing: Advantages and Disadvantages," I*nternational Journal of Advanced Science and Technology*, (29:7), pp. 2137–2150.

Jain, A., and Kasbe, A. 2018. "*Fake News Detection*," *Proceedings of IEEE Intern. Students' Conference on Electrical, Electronics and Computer Science* (SCEECS), pp. 1-5.

Keita, Z. 2021, "Text data representation with one-hot encoding, Tf-Idf, Count Vectors, Co-occurrence Vectors and Word2Vec" [Online]. Available: https://towardsdatascience.com/text-data-representation-with-one-hot-encoding-tf-idf-count-vectors-co-occurrence-vectors-and-f1bccbd98bef. Last accessed: Sept 7, 2023.

Khivasara, Y., Khare, Y., and Bhadane, T. 2020. "Fake News Detection System using Web-Extension," *Proceedings of 2020 IEEE Pune Section International Conference* (PuneCon), pp. 119-123, doi: 10.1109/PuneCon50868.2020.9362384.

Lazer, D.M.J., Baum, M.A., Benkler, Y., Berinsky, A. J., Greenhill, K.M., Menczer, F., Metzger, M.J., Nyhan, B., Pennycook, G., Rothschild, D., and Schudson, M. 2018. "The science of fake news," *Science*, (359:6380), pp. 1094-1096.

Nieminen, S., and Rapeli, L., 2018. "Fighting Misperceptions and Doubting Journalists' Objectivity: A Review of Fact-checking Literature," *Political Studies Review*, (17:3), pp. 296-309. doi: 10.1177/1478929918786852.

Paschalides, D., Christodoulou, C., Andreou, R., Pallis, G., Dikaiakos, M.D., Kornilakis, A., and Markatos, E. 2019. "Check-It: A plugin for Detecting and Reducing the Spread of Fake News and Misinformation on the Web," *Proceedings of 2019 IEEE/WIC/ACM International Conference on Web Intelligence* (WI), pp. 298-302.

Pew Research center. 2023. "Digital News Fact Sheet," *Pew Research Center's Journalism Project*. [Online]. Available: https://www.pewresearch.org/journalism/fact-sheet/digital-news/. Last accessed: May 15, 2023.

Phippen, A., Bond, E., and E. Buck, E. 2021. "Effective strategies for information literacy education: Combatting "fake news" and empowering critical thinking," *Future Directions in Digital Information*, Chandos Publishing, pp. 39-53.

Radhakrishnan, S., Erbis, S., Isaacs, J.A., and Kamarthi, S. 2017. "Novel keyword co-occurrence network-based methods to foster systematic reviews of scientific literature," *PLoS ONE*, (12:3), pp. e0172778, doi: 10.1371/journal.pone.0172778.

Reality Defender. 2023. "Reality Defender". [Online]. Available: https://www.realitydefender.ai. Last accessed: Jan. 28, 2023.

Rohman, M.A., Khairani, D., Hulliyah, K., Arini, P., Riswandi, and Lakoni, I. 2021. "Systematic Literature Review on Methods used in Classification and Fake News Detection in Indonesian," *Proceedings of 9th Intern. Conference on Cyber and IT Service Management* (CITSM), pp. 1-4.

Saini, G.K., Lievens, F., and Srivastava, M. 2022. "Employer and internal branding research: a bibliometric analysis of 25 years," *Journal of Product & Brand Management*, (31:8), pp. 1196-1221, doi: 10.1108/JPBM-06-2021-3526.

Securing Cloud Functions. 2023. "Cloud Functions Documentation," *Google Cloud*. [Online]. Available: https://cloud.google.com/functions/docs/securing. Last accessed: June 16, 2023.

Song, C., Ning, N., Zhang, Y., and Wu, B. 2021. "A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks," I*nformation Processing & Management*, (58:1), p. 102437, doi: 10.1016/j.ipm.2020.102437.

Trujillo, C.M., and Long, T.M. 2018. "Document co-citation analysis to enhance transdisciplinary research," *Science Advances*, (4:1), pp. e1701130, doi: 10.1126/sciadv.1701130.

The Factual - Unbiased News, Trending Topics - The Factual. [Online]. Available: https://www.thefactual.com/?lp=new. Last accessed: May 16, 2023.

van Eck, N.J., and Waltman, L. 2010. "Software survey: VOSviewer, a computer program for bibliometric mapping," *Scientometrics*, (84:2), pp. 423-538, doi: 10.1007/s11192-009-0146-3.

Vinhas, O., and Bastos, M. 2022. "Fact-Checking Misinformation: Eight Notes on Consensus Reality," *Journalism Studies*, (23:4), pp. 448-468.

Yangming L., and Yunlai Z. 2022. "An Ensemble Learning Approach for COVID-19 Fact Verification," P*roceedings of 3rd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, pp. 383-387.

Wei, F., and Zhang, G. 2020. "A document co-citation analysis method for investigating emerging trends and new developments: a case of twenty-four leading business journals," *Science Advances*, (4:1).

Zare-Farashbandi, F., Geraei, E., and Siamaki, S. 2014. "Study of co-authorship network of papers in the Journal of Research in Medical Sciences using social network analysis," *Journal of Research in Medical Sciences*, (19:1), pp. 41-46.