This paper was downloaded from

Charles Sturt University

https://researchoutput.csu.edu.au

Accepted manuscript for the following conference paper.

Paper title: COVIDFakeExplainer: An explainable machine learning based web application for detecting COVID-19 fake news

Authors: Dylan Warman & Muhammad Ashad Kabir

Conference title: 2023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)

Pages: 7

Theme: N/A

Conference dates: 4th December - 6th December 2023

Conference location: Yanuca Island, Fiji

Copyright 2023 IEEE. Published in the 2023 IEEE Asia-Pacific Conference on Computer Science and Engineering (CSDE) 4th-6th December 2023 in Fiji. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works, must be obtained from the IEEE. Contact: Manager, Copyrights and Permissions / IEEE Service Center / 445 Hoes Lane / P.O. Box 1331 / Piscataway, NJ 08855-1331, USA. Telephone: + Intl. 908-562-3966

COVIDFakeExplainer: An Explainable Machine Learning based Web Application for Detecting COVID-19 Fake News

Dylan Warman
School of Computing, Mathematics and Engineering
Charles Sturt University
NSW, Australia
dwarman@csu.edu.au

Muhammad Ashad Kabir
School of Computing, Mathematics and Engineering
Charles Sturt University
NSW, Australia
akabir@csu.edu.au

arXiv:2310.13890v1 [cs.SI] 21 Oct 2023

***Abstract*—Fake news has emerged as a critical global issue, magnified by the COVID-19 pandemic, underscoring the need for effective preventive tools. Leveraging machine learning, including deep learning techniques, offers promise in combatting fake news. This paper goes beyond by establishing BERT as the superior model for fake news detection and demonstrates its utility as a tool to empower the general populace. We have implemented a browser extension, enhanced with explainability features, enabling real-time identification of fake news and delivering easily interpretable explanations. To achieve this, we have employed two publicly available datasets and created seven distinct data configurations to evaluate three prominent machine learning architectures. Our comprehensive experiments affirm BERT's exceptional accuracy in detecting COVID-19-related fake news. Furthermore, we have integrated an explainability component into the BERT model and deployed it as a service through Amazon's cloud API hosting (AWS). We have developed a browser extension that interfaces with the API, allowing users to select and transmit data from web pages, receiving an intelligible classification in return. This paper presents a practical end-to-end solution, highlighting the feasibility of constructing a holistic system for fake news detection, which can significantly benefit society.**

***Index Terms*—COVID-19, machine learning, deep learning, fake news, explainability, web application, chrome extension**

I. Introduction

Fake news is known by many interchangeable names, with the main two being the terms "Fake news" itself and "Misinformation" [1], [2]. These terms are used to mean false or misleading information shared to deceive an individual, group, or population into believing something that is clearly not true, often with political motivations with the intention of damaging public trust. In the context of this article, we refer to all of this as fake news [3].

Social networks provide a platform that millions of people around the world use to communicate and share information on a daily basis. However, especially at a time of global crisis such as during the COVID-19 pandemic, the amount of Fake news, being shared is staggering. Global statistics indicate that 74% people are very concerned about the amount of fake news they have seen during the pandemic [4], and furthermore, studies have shown that more than 50% of all social media users have spread fake news knowingly or unknowingly [5].

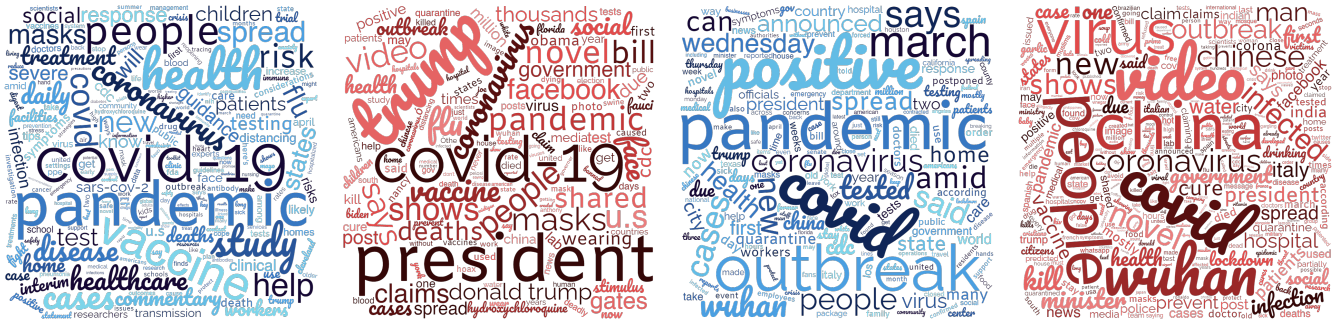
Further research shows that fake news is not more likely to be shared by robots or artificial intelligence, people were found to be more likely to spread fake news [6]. A potential reason for sharing fake news is identified as our cognitive biases, more specifically our memory biases, and a phenomenon known as the "false memory effect" [7]. Additionally, there has been shown to be a large disconnect between what people believe and what types of fake news they will share, furthering the idea that people share this information irrationally [8].

The magnitude of this problem has been underscored by the World Health Organization, which categorizes the proliferation of fake news during the COVID-19 pandemic as an "Infodemic" [9]. They explicitly emphasize that an Infodemic can pose as much risk to public health and security as the virus itself.

Given the clarity the World Health Organisation provides on the enormity of this issue, along with the consideration of how easily people with the best intentions can share fake news, the development of a tool that can provide a person with an understanding of what they are reading with regard to its legitimacy and validity is vital. Furthermore, considering how fake news can be used maliciously with the intent to harm public trust and cause unrest, it is important that people have the ability to protect themselves to prevent a continued deterioration of the public's understanding of the pandemic among other topics.

Significant research has been conducted in this area recently, with machine learning (ML) approaches being the most widely implemented [2]. Incorporating explainability [10] could further allow us to trace the predictions generated by these ML approaches, in particular deep learning (DL) models which are considered as "black box", and provide a contextual

©2023 IEEE. This manuscript has been accepted to publish in the 10th IEEE Asia-Pacific Conference on Computer Science and Data Engineering (IEEE CSDE 2023). Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.



(a) CoAID – real news (b) CoAID – fake news (c) C19-Rumor – real news (d) C19-Rumor – fake news

Fig. 1: Word clouds generated from COVID-19 news datasets

understanding of why a particular classification is made. Therefore, explainability in fake news detection could significantly increase user trust [11] and would provide a concrete understanding to the user why an article is fake and could potentially lead to a reduction in the sharing of fake news.

This paper aims to develop an explainability tool/application that is embedded directly into a Google Chrome extension. In particular, this paper makes the following three major contributions:

- We have conducted an extensive empirical evaluation of state-of-the-art ML algorithms to train a fake news classification model using two different datasets with seven configurations.
- We have identified the most prominent explainability techniques and discussed their suitability for developing a web-based application.
- We have implemented a web application as a Google Chrome extension using the best-performed ML model and the most suitable explainability technique to demonstrate the suitability and usefulness of our approach.

II. Related Work

The current landscape of tools for detecting fake news reveals several limitations and gaps in addressing the critical need for accessible and explainable solutions. CoVerifi [12] is a functional application that provides accurate classifications and human generation scores for COVID-19-related news articles, but it lacks genuine explainability techniques, leaving users without a comprehensive understanding of the reasoning behind the classification. However, a more fundamental issue highlighted by dEFEND [13] is the scarcity of tools accessible to end-users for detecting fake news.

A notable example of such limitations is FakerFact [14], a Chrome extension that analyzes and verifies fake news by URL. Although it offers classification percentages in various areas, it falls short of providing direct explainability. Similarly, SEMiNExt [15] analyses user search terms for potential fake news content without explaining or classifying actual news articles. While there are tools [11] like xFake [16] that demonstrate the potential for sentiment and linguistic analysis with explainable outputs, they often have limitations, like being restricted to specific websites, as seen in xFake's compatibility with PolitiFact [17] only.

Despite these limitations, tools like Bunyip [18] provide promising visual explainability outputs and classifications for human-generated text, showing that similar applications can be developed. However, these tools do not directly address traditional fake news detection, although they serve as proof of concept for the feasibility of creating user-friendly and explainable applications.

In summary, the existing tools fall short of providing comprehensive and user-friendly solutions for detecting fake news, particularly identifying COVID-19-related fake news with explainability. There is a significant gap between the state-of-the-art studies on machine learning and explainability techniques for fake news, and the ideal end-user tools that offer both accurate classifications and clear interpretable explanations. While the above-discussed tools demonstrate progress, they underscore the necessity for a holistic approach that provides a user-friendly experience coupled with meaningful explainability to empower users in identifying and understanding fake news.

III. Materials and Methods

A. Dataset Pre-processing and Configurations

We have used the CoAID (Covid-19 heAlthcare mIsinformation Dataset) [19] with 5216 total news items and the C19-Rumor (A COVID-19 Rumor Dataset) dataset [20] with a total of 4129 news items. Both datasets have dedicated segments for news headlines and news articles about COVID-19. We selected these two datasets given their spread of real and fake news complement each other, with CoAID heavily weighted towards true or real news and C19-Rumor heavily weighted towards false or fake news. This alternate weighting allows us to test the datasets individually, with augmentation, and in combination, to assess the impacts of different class weightings on the data.

Fig. 1 illustrates the word cloud of real and fake news for both datasets. It allows for a direct comparison between the two datasets. From these word clouds, we can deduce that while COVID-19 and its variants are the predominant terms

TABLE I: Dataset Configurations

Dataset		Split	Class		Total	Configuration
			True (Real)	False (Fake)		
CoAID	Original	Training (70%)	2426	634	3060	C1
		Validation (20%)	681	193	874	
		Testing (10%)	349	89	438	
	Augmented	Training	2419	2419	4838	C2
		Validation	688	688	1376	
		Testing	349	349	698	
C19-Rumor	Original	Training	452	2137	2589	C3
		Validation	145	595	740	
		Testing	62	308	370	
	Augmented	Training	2117	2117	4234	C4
		Validation	615	615	1230	
		Testing	308	308	616	
Cross	CoAID train and validation C19-Rumor test	Training	2779	2779	5558	C5
		Validation	677	677	1354	
		Testing	659	659	1318	
	C19-Rumor train and validation CoAID test	Training	2443	2443	4886	C6
		Validation	597	597	1194	
		Testing	916	916	1832	
Merged		Training	2870	2779	5649	C7
		Validation	836	778	1614	
		Testing	409	399	808	

used within the CoAID dataset, the key terms within the C19-Rumor dataset are more aligned with words such as pandemic, outbreak, China, and Wuhan. This observation suggests that the CoAID dataset, given its focus on health and medical-related news, is unlikely to contain information related to the outbreak, videos, or Wuhan specifically.

We created seven configurations using the two named datasets for an extensive experimental evaluation. These configurations, as well as the training, validation, and test splits, are outlined in Table I. The objectives of configurations C1 and C3 are to establish baselines for each of the two datasets. Configurations C2 and C4 aim to assess whether data augmentation, particularly considering the limited dataset size, offers any advantages in terms of accuracy or if it adversely affects classification. C5 and C6 are employed to investigate two aspects: firstly, whether the datasets are representative of each other, and secondly, how well a model developed using 2020 COVID-19 fake news performs on news stories from 2021, and vice versa. Finally, the purposes of C7 (merged dataset) are to improve model robustness, increase generalisation, and mitigate the limitations associated with individual datasets such as imbalances in class distribution, lack of coverage for specific topics, or biases in data collection, which can ultimately enhance the accuracy and effectiveness of fake news detection systems.

B. Machine Learning Techniques

In this study, we employed a baseline CNN [21] model, and two advanced models, BERT [22] and Bi-LSTM [12] which have demonstrated high efficiency in fake news detection according to prior research [11], [12], [23]. Both BERT and Bi-LSTM have shown strong performance with small datasets [24], [25], which is crucial given the small size of our sourced datasets. In contrast, a CNN model is a more basic algorithm type compared to the first two, providing a valuable baseline to assess what a simpler algorithm can achieve with the same dataset.

BERT [22] holds immense promise for fake news detection due to its ability to comprehend the nuances of language and context. By pre-training on a massive corpus of text, BERT becomes adept at understanding the subtle linguistic cues that often distinguish fake news from genuine content. Its bidirectional architecture allows it to capture relationships between words, making it highly effective in discerning the contextual intricacies that fake news articles often employ to deceive readers. Additionally, BERT's fine-tuning capability enables it to adapt to specific datasets, thereby enhancing its accuracy in identifying misleading or fabricated information. As fake news continues to pose a significant challenge, BERT's natural language processing prowess positions it as a valuable tool in the ongoing fight against misinformation and disinformation.

Bi-LSTM [12], on the other hand, is a recurrent neural network, that specializes in capturing sequential dependencies in text. This makes it particularly effective at discerning subtle linguistic patterns within shorter pieces of text and excels at analyzing the structural flow of information within news articles.

C. Explainable Techniques

SHAP (SHapley Additive exPlanations) [26] and LIME (Local Interpretable Model-Agnostic Explanations) [27] are two popular explainable machine learning techniques used in the context of fake news detection to provide insights into model predictions and make the decision-making process more transparent. In this study, we employed SHAP as it holds several advantages over LIME when it comes to explaining the predictions of machine learning models. One key advantage is the global interpretability that SHAP offers. Unlike LIME, which provides local explanations for individual predictions, SHAP calculates feature importance consistently

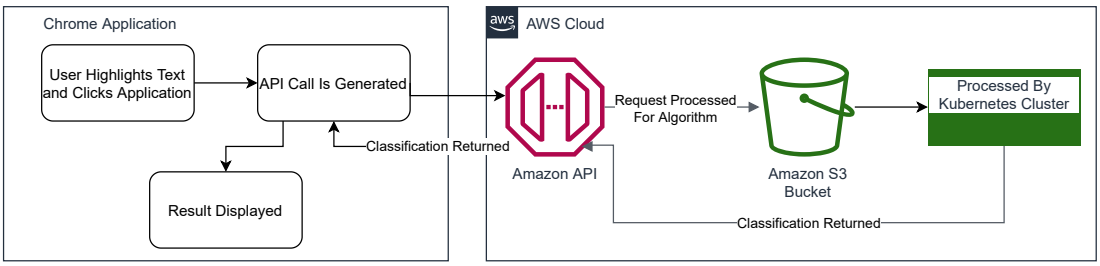


Fig. 2: Web application architecture

across all possible feature combinations [26]. This means that SHAP gives a holistic view of how each feature impacts model predictions across the entire dataset, allowing for a more comprehensive understanding of the model's behavior. Additionally, SHAP is grounded in cooperative game theory, providing a mathematically rigorous framework for explaining the contributions of each feature. This makes SHAP particularly useful when it needs to identify overarching patterns and relationships within data, which is often crucial in applications like fake news detection, where understanding global linguistic and contextual patterns is essential for model transparency and improvement.

D. Web Application Architecture

Our web application's architecture, depicted in Fig. 2, comprises two main components: a server component hosted on the AWS (Amazon Web Services) cloud and a client component implemented as a Chrome plugin.

The server for our application is hosted on an S3 (Simple Storage Service) Bucket within AWS[1]. This bucket serves as a storage location, providing access to the files required to run the algorithm. The server hosts a trained machine learning model (Section III-B) and the request handler for the application. To publish this on AWS and generate the necessary components for the API, we used Cortex[2]. Cortex is a free application that employs Docker images to automatically create Kubernetes clusters on AWS. A Kubernetes cluster consists of a set of functions from AWS that facilitate communication among the nodes, each representing a specific feature or function. This approach allows us to generate an API and a public access point for the application without the need to manually configure each feature individually. Once the API is accessible, it is connected using JavaScript within the client component (i.e., Chrome extension). This involves sending a request to the previously uploaded request handler within the S3 bucket, which generates a prediction and a set of force values representing explainability. These results are then returned to the Chrome extension.

The client component (Chrome plugin) is built using HTML, with JavaScript providing the essential functionality for transmitting and processing text. JavaScript processes a request by reading the user-selected data (highlighted text) from the active webpage when the user clicks the application icon. Subsequently, the extension converts the response (classification result with an explanation) from the server into a readable format, applies CSS formatting to provide color-coding for the explainability element, and presents the results to the user.

[1]https://aws.amazon.com/

[2]https://github.com/cortexlabs/cortex

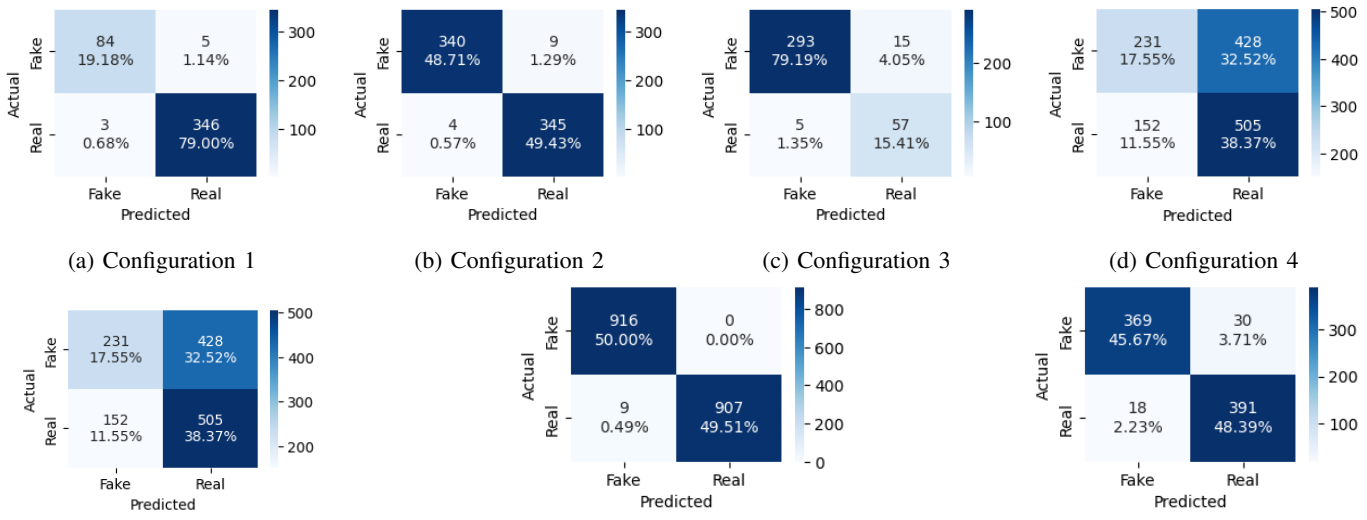
TABLE II: Classification Results

Config.	Model name	Precision	Recall	F1	Accuracy
C1	BERT	98.16	98.17	98.17	**98.17**
	Bi-LSTM	92.77	92.92	92.81	92.92
	CNN	93.72	93.84	93.74	93.84
C2	BERT	98.15	98.14	98.14	**98.14**
	Bi-LSTM	85.85	81.09	80.47	81.09
	CNN	86.55	83.24	82.85	83.24
C3	BERT	95.11	94.59	94.75	**94.59**
	Bi-LSTM	87.42	88.38	87.14	88.38
	CNN	88.30	88.38	88.34	88.38
C4	BERT	76.19	75.97	75.92	**75.97**
	Bi-LSTM	75.41	72.73	71.99	71.99
	CNN	73.14	61.80	56.47	61.80
C5	BERT	56.23	54.78	51.98	54.78
	Bi-LSTM	51.28	50.91	47.12	50.91
	CNN	57.06	55.84	53.85	**55.84**
C6	BERT	99.41	99.40	99.40	**99.40**
	Bi-LSTM	50.00	50.00	42.44	50.00
	CNN	46.95	49.13	38.07	49.13
C7	BERT	94.10	94.06	94.06	**94.06**
	Bi-LSTM	86.76	86.76	86.76	86.76
	CNN	88.07	88.00	87.98	88.00

IV. Results and Discussion

A. ML Models Evaluation

The results of running the three models (BERT, Bi-LSTM, and CNN) for seven configurations (C1 to C7) are presented in Table II. For configuration C1, all three models consistently perform well, achieving the highest average and displaying the most consistent scores among all the tests. However, it is important to consider whether this high accuracy might be influenced by the imbalanced nature of the dataset. This consideration is quickly addressed by examining Fig. 3a, which displays the BERT confusion matrix. Here, only five fake labels were misclassified compared to the three from the real set. This suggests that the dataset's imbalanced nature is unlikely to have significantly influenced the accuracy. Instead, it appears that the dataset contains well-separated data that the models were able to effectively distinguish.



(a) Configuration 1 (b) Configuration 2 (c) Configuration 3 (d) Configuration 4

(e) Configuration 5 (f) Configuration 6 (g) Configuration 7

Fig. 3: Confusion matrix of the best model of each configuration

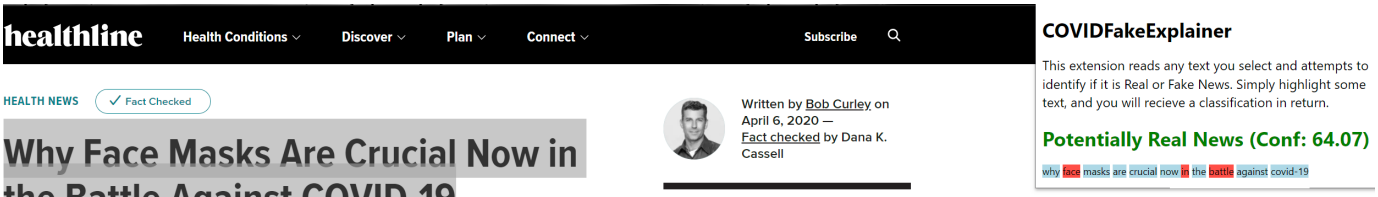
For configuration C2, it is worth noting that while BERT maintains a consistent score, with only a minor drop of 0.03%, both the Bi-LSTM and CNN models experience significant decreases in accuracy, exceeding 10%. This drop in accuracy suggests that oversampling is ineffective for this dataset, and it indicates that BERT is more resilient to the challenges posed by oversampled data. Fig. 3b displays the relationship between predicted labels and true labels, which closely resembles that of C1. This further confirms that the accuracy achieved in C1 was not solely due to the imbalanced dataset. Despite the slight reduction in accuracy, this configuration strengthens the case that BERT is the optimal model for this particular scenario.

For configuration C3, we observe lower overall accuracy compared to C1. This suggests that the examples in this dataset are more challenging to distinguish than those in the first dataset. More challenging data can be both a positive and negative aspect. On the one hand, it demonstrates that the algorithm can still differentiate between the examples to a significant extent. On the other hand, the lower accuracy indicates that the algorithms face greater difficulty in establishing connections between textual elements. Additionally, when we examine Fig. 3c, we notice that the optimal algorithm's accuracy varies mainly when classifying real data. This variance can be attributed to the overall scarcity of real news in this dataset, resulting in a 20.83% misclassification rate for real news.

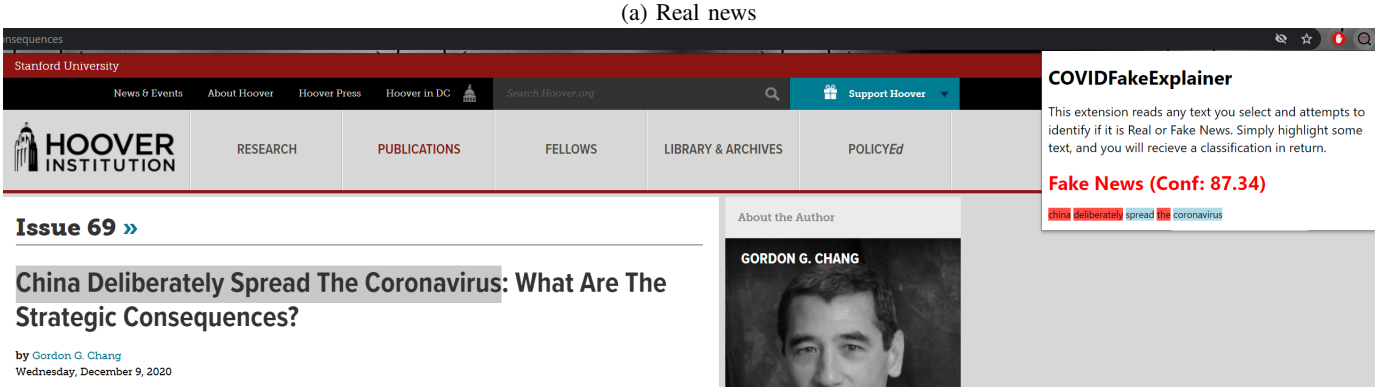
Interestingly, the trend observed in C2 continues in C4, with a significant decrease in accuracy across all three models. BERT experiences the most substantial accuracy drop when compared to C3, which can be partly attributed to its higher initial accuracy. Across all three models, the average drop in accuracy is 20.53%, a sharp increase compared to the average drop of 7.49% in C2. This significant increase in accuracy loss suggests that this second dataset is less robust than the first. When comparing Fig. 3d to Figure 3b, the general trend in accuracy is evident. However, Fig. 3d displays notably lower accuracy when dealing with what was initially the minority class. Despite the considerable accuracy drop, BERT outperforms the other models once again, strengthening the argument that it is the overall best model.

Examining the cross-dataset evaluation results depicted in Fig. 3e and Fig. 3f, we observe an interesting aspect in the results of C5 and C6. Models trained on the CoAID dataset performed poorly on the C19-Rumor dataset (Fig. 3e). Conversely, the best model trained on the C19-Rumor dataset excelled when tested on the CoAID dataset (Fig. 3f). This phenomenon suggests that the C19-Rumor dataset effectively represents the characteristics of the CoAID dataset, but the reverse is not necessarily true. This relationship is also partly reflected in the word cloud analysis presented in Fig. 1.

In C7, we consistently observe a high level of accuracy, reaffirming the primary objective of this configuration. This objective, centered around testing the models' ability to handle an expanded dataset covering various subcategories, has been convincingly validated. Once again, BERT outperforms the other models in terms of accuracy, as evident in Table II. Fig. 3g illustrates a minimal number of misclassified inputs. The significant advantage of witnessing C7 perform exceptionally well lies in its confirmation that any marginal reduction in accuracy can be attributed to the models' enhanced comprehension of the broader subject of COVID-19. This heightened understanding is a direct result of the merged dataset, which thoughtfully addresses issues like class distribution imbalances, topic coverage limitations, and data collection biases present in individual datasets. As a result, the merged dataset naturally enriches the pool of COVID-19-related data, ultimately fostering the accuracy and effectiveness of fake news detection systems.



(a) Real news



(b) Fake news

Fig. 4: Chrome extension is providing classification results with explanation for a selected text in webpage

B. Web Application Evaluation

Fig. 4 demonstrates a real-time use of our web application, encompassing the article content and highlighting the specific line within that article that is undergoing classification. Notably, there is currently no similar application except for CoVerifi [12], which necessitates copying the text and leaving the active page or site to generate a response. In contrast, our tool offers a straightforward highlight-and-click function, eliminating the need for copying or additional buttons to produce results, as required by tools that redirect to external pages. Furthermore, our tool grants users the flexibility to select any text they desire, providing complete control over the application's inputs.

V. Conclusion

This paper presents a comprehensive pipeline encompassing the entire process, from training machine learning models to prototype implementation, for the detection of fake news with explainability. Our study demonstrates the potential of combining machine learning and explainability techniques to create a web application tailored for detecting COVID-19-related fake news. Through an extensive empirical analysis, we evaluated the performance of three prominent machine learning algorithms for text classification across seven distinct configurations, employing two distinct datasets. The results indicate that BERT emerges as the optimal choice for COVID-19 fake news classification. Moreover, we critically examined the two leading explainability visualization techniques, offering insights into their respective advantages and limitations. Finally, we developed a prototype web application in the form of a Chrome extension. This approach is highly adaptable and can be extended beyond COVID-19 fake news detection, for instance, to classify text or misinformation in a variety of domains. The only requirement for this adaptation is the replacement of the model that serves as the foundation for the application.

In the future, we plan to further improve the application's robustness by incorporating a larger and more comprehensive dataset. Additionally, we intend to conduct thorough evaluations to assess the usability and performance of the application.

References

[1] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1–40, 2020.

[2] A. Sanaullah, A. Das, A. Das, M. A. Kabir, and K. Shu, "Applications of machine learning for covid-19 misinformation: a systematic review," *Social Network Analysis and Mining*, vol. 12, no. 1, p. 94, 2022.

[3] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, and et al., "The science of fake news," *Science*, vol. 359, no. 6380, p. 1094–1096, 2018.

[4] A. Watson, "Problems with finding coronavirus news worldwide 2020," https://www.statista.com/statistics/1104506/coronavirus-news-opinions-worldwide/, 2020, accessed: 30 August, 2023.

[5] ——, "Sharing of made-up news on social networks in the u.s. 2020," https://www.statista.com/statistics/657111/fake-news-sharing-online/, 2023, accessed: 30 August, 2023.

[6] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, p. 1146–1151, 2018.

[7] M. A. Britt, J.-F. Rouet, D. Blaum, and K. Millis, "A reasoned approach to dealing with fake news," *Policy Insights from the Behavioral and Brain Sciences*, vol. 6, no. 1, p. 94–101, 2019.

[8] G. Pennycook and D. G. Rand, "The psychology of fake news," *Trends in Cognitive Sciences*, vol. 25, no. 5, p. 388–402, 2021.

[9] T. L. I. Diseases, "The covid-19 infodemic," *The Lancet Infectious Diseases*, vol. 20, no. 8, p. 875, 2020.

[10] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *CoRR*, vol. abs/1708.08296, 2017.

[11] J. Ayoub, X. J. Yang, and F. Zhou, "Combat COVID-19 infodemic using explainable natural language processing models," *Information Processing & Management*, p. 102569, mar 2021.

[12] N. L. Kolluri and D. Murthy, "CoVerifi: A COVID-19 news verification system," *Online Social Networks and Media*, vol. 22, 2021.

[13] K. Shu, S. Wang, L. Cui, D. Lee, and H. Liu, "dEFEND: Explainable Fake News Detection," *dl.acm.org*, pp. 395–405, jul 2019.

[14] Y. Ma, D. Towey, T. Yueh Chen, and Z. Quan Zhou, "Metamorphic testing of fake news detection software," *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, 2021.

[15] A. B. Shams, E. Hoque Apu, A. Rahman, M. M. Sarker Raihan, N. Siddika, R. B. Preo, M. R. Hussein, S. Mostari, and R. Kabir, "Web search engine misinformation notifier extension (seminext): A machine learning based approach during covid-19 pandemic," *Healthcare*, vol. 9, no. 2, p. 156, 2021.

[16] F. Yang, S. K. Pentyala, S. Mohseni, M. Du, H. Yuan, R. Linder, E. D. Ragan, S. Ji, and X. B. Hu, "Xfake: Explainable fake news detector with visualizations," *The World Wide Web Conference on - WWW '19*, 2019.

[17] Poynter Institute, "Politifact," https://www.politifact.com/, accessed: 30 August, 2023.

[18] R. Sawant, "Bunyip," https://awesomeopensource.com/project/CT83/Bunyip, accessed: 30 August, 2023.

[19] L. Cui and D. Lee, "CoAID: COVID-19 healthcare misinformation dataset," 2020.

[20] M. Cheng, S. Wang, X. Yan, T. Yang, W. Wang, Z. Huang, X. Xiao, S. Nazarian, and P. Bogdan, "A COVID-19 rumor dataset," *Frontiers in Psychology*, vol. 12, 2021.

[21] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Trans. Neural Netw.*, vol. 33, no. 12, pp. 6999–7019, 2022.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, p. 4171–4186.

[23] C. Colón-Ruiz and I. Segura-Bedmar, "Comparing deep learning architectures for sentiment analysis on drug reviews," *Journal of Biomedical Informatics*, vol. 110, 2020.

[24] J. Y. Khan, M. T. Khondaker, S. Afroz, G. Uddin, and A. Iqbal, "A benchmark study of machine learning models for online fake news detection," *Machine Learning with Applications*, vol. 4, 2021.

[25] A. Ezen-Can, "A comparison of LSTM and BERT for small corpus," *ArXiv*, vol. abs/2009.05451, 2020.

[26] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, p. 4768–4777.

[27] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, p. 1135–1144.