2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) | 979-8-3315-0494-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/WI-IAT62293.2024.00111

Machine Learning-Based Fake News Detection on Social Media

Li Zeng and Xiaoci Tao
Information Retrieval & Knowledge Management Research Lab
YorK University, Canada
lz1202@yorku.ca, txc0411@yorku.ca

***Abstract*—In the age of data, artificial intelligence has become an effective tool for detecting fake news, enabling a swift response to correct misinformation during critical events such as elections, health crises, and global disasters. This case study provides a detailed analysis of news shared by nine media organizations on Facebook in the crucial week leading up to the 2016 U.S. election. By employing detailed fact-checking and advanced data mining techniques, we investigate the linguistic characteristics of fake news articles to develop models capable of accurately distinguishing between fake and authentic news. Our approach includes an exploratory data analysis (EDA) to identify patterns and linguistic features, followed by the application of multiple machine learning algorithms for classification. The results yield valuable insights into the defining characteristics of fake news and demonstrate the potential of machine learning to enhance the detection and filtering of misinformation.**

***Index Terms*—Fake news detection, Machine learning, Exploratory data analysis (EDA), Classification algorithms, 2016 U.S. election.**

I. Introduction

In recent years, social media has become increasingly integral to both professional and personal lives. Breaking news is often first reported on these platforms rather than traditional media channels [1], which has led to the widespread dissemination of unverified information. Users frequently share news without verifying its accuracy, contributing to the spread of fabricated and inaccurate content, such as rumors, hoaxes, and misinformation [2]. This proliferation of fake news undermines social trust, informed decision-making, and can significantly shape political landscapes, influence public opinion, and erode the democratic process. The growing urgency to detect and combat fake news calls for reliable, real-time detection systems.

A. Background

The rise of advanced language models and search engines has made information retrieval easier and more accurate [3]. However, the diversity of media formats available on social platforms—ranging from text to video and audio—complicates the detection of fake news [4]. This influx of information, coupled with the ability of fake news to impact politics, economics, and culture, has garnered the attention of researchers across fields such as information technology, social sciences, and computer science [5]. These disciplines aim to better understand the spread, origins, and detection of fake news.

B. Significance of fake news detection

Fake news definitions online nowadays tends to be intrusive and diverse in terms of topics, styles and platforms [6]. According to Wikipedia (Fake news), Fake news or information disorder is false or misleading information, misinformation including disinformation, propaganda, and hoaxes, presented as news. According to Stanford University's definition, fake news is "the news articles that are intentionally and verifiably false, and could mislead readers (Detecting fake news with nlp)". There are other authors' definitions of Fake News, however in this paper, the proposed definition is that "Fake news refers to any false information or false stories that are published on the internet, to mislead readers purposely."

Fake news characteristics are used to clearly understand the scope and variety of online fake information, one study showed some important aspects for defining fake news [7]. "Fake News" contains four major components: Creator and Spreader, Target Victims, News Content, and Social Context. Anyone can be a fake news creator/spreader or target victims with the wide use of social media. In context of the news content, the study proposed that each news item consists of physical news content and non-physical news content. Non-physical content is the main kernel of fake news since it contains all the important ideas, feelings and views that the authors want to pass to the readers. The social context is an influential indicator of the distribution patterns of both false news and real news. Sentiment polarity is another important feature of non-physical content for fake news. In order to make their news persuasive, authors often express strong positive or negative feelings in the text body [8].

Fake news poses a significant threat to societal trust and belief systems, particularly evident in events like the 2016 U.S. presidential election. During this campaign, a multitude of Russian fake accounts inundated social media platforms with misinformation, targeting supporters of Hillary Clinton with false claims such as "Hillary was sick", "Hillary was a criminal", "Obama had a secret army" are some such examples. This deliberate dissemination of falsehood had the potential to sway public opinion and undermine confidence in authoritative figures and institutions. The aftermath of the election saw discussions attributing Donald Trump's victory to the proliferation of fake news, further highlighting its detrimental impact on democratic processes [9]. Today, the

979-8-3315-0494-6/24/$31.00 ©2024 IEEE
DOI 10.1109/WI-IAT62293.2024.00111

pervasiveness of fake news continues to exert damaging effects across various domains, including politics, technology, finance, and societal trust. Addressing this urgent issue requires the development of reliable, real-time systems for detecting and combating fake news.

C. Motivation and Contribution

Given the serious impact that fake news has on societal stability and trust, there is a pressing need for effective mechanisms to detect and predict fake news. This work focuses on classifying fake news from BuzzFeed News, specifically analyzing news from the 2016 U.S. presidential election week. We begin with exploratory data analysis (EDA) to identify patterns and key features of the dataset. These insights provide the foundation for comparing language patterns between fake and real news articles. Furthermore, we evaluate different machine learning classifiers to assess their ability to distinguish between genuine and fake news, offering a discussion on future research directions in this domain.

II. Literature review

The pervasive spread of fake news and its significant societal consequences have driven a substantial body of scholarly research aimed at addressing this urgent issue. Recent studies have focused heavily on the integration of machine learning and natural language processing (NLP) techniques as powerful tools for combating fake news. Researchers have explored the nuances of language, seeking to uncover subtle distinctions between real and fake news articles, which can enhance detection accuracy.

One notable study [10] provides a comprehensive overview of fake news research, focusing on theoretical frameworks, detection strategies, and unresolved challenges. The authors discuss various theoretical perspectives surrounding fake news, including social psychology theories and computational models, which provide a critical aspect of developing precise detection models. These frameworks are vital in understanding how individuals engage with fake news and how this understanding can inform the design of detection models.

A key component in the development of effective fake news detection systems is the availability of robust evaluation datasets. An article [11] conduct a survey of evaluation datasets used in fake news detection research, categorizing them based on size, language, and domain. Their analysis underscores the need for diverse and representative datasets to ensure the robustness and generalizability of detection models. This focus on dataset diversity is essential for creating models that can perform well across different contexts and platforms.

Further contributing to this area of research, another paper [12] explored the use of language-independent features in fake news detection across multiple languages. Their study demonstrated the potential for cross-lingual detection models, emphasizing the importance of shared linguistic features in identifying fake news. This work points to the possibility of developing more universal fake news detection systems that can operate effectively across different languages and regions, addressing a significant gap in the current literature.

These studies collectively underscore the evolving landscape of fake news detection, as the field increasingly leans towards sophisticated computational methods. The integration of linguistic features and advanced machine learning models is indicative of a paradigm shift in the quest for a more resilient defense against the proliferation of fabricated information in the digital age. These developments serve as a compelling backdrop for the current research, which seeks to push the boundaries of fake news detection by harnessing cutting-edge machine learning techniques.

III. Methodology and Experimental

This chapter outlines the methodology employed in the case study on Machine Learning-Based Fake News Detection. Given the complexities and the multifaceted nature of fake news dissemination, especially during significant events like the 2016 U.S. presidential election, our approach combines meticulous data collection, exploratory data analysis (EDA), and the application of various machine learning classifiers [13], [14]. The methodology is structured to not only identify linguistic patterns and features unique to fake news but also to compare the accuracy of fake news detection.

A. Dataset

The dataset used in this study is from the Kaggle website and features news reported by nine different news organizations in the week leading up to the 2016 U.S. election. The content of the data has been certified by Buzzfeed to be authentic. The dataset is divided into two categories: fake news and real news. Each category includes 91 entries and 12 different attributes. After checking, the variables are described as strings and there are no missing values in the dataset which indicates that the dataset is clean and well structured for further analysis. The attributes which include ID, title, text, source, images, and videos [15], [16].

- **id:** A unique identifier for the webpage of the news article, indicating its authenticity.
- **title:** The headline designed to capture reader's interest, is closely related to the news topic's essence.
- **text:** The article's main body, detailing the news story and often emphasizing and elaborating on a central claim.
- **source:** The author or publisher of the news piece.
- **images:** Visual elements that support the article's content, aiding in story framing.
- **videos:** Video content embedded in the news article, including video clips of the news story or related footage.

B. Text processing

Data preprocessing is a key step in the machine learning workflow, which directly affects the performance of the model, training efficiency and generalization ability. This part mainly relies on the `preprocess_corpus()` function. to clean the

dataset and do the transformation and integration for different attributes, the specific steps are as follows:

1) **Lowercasing:** Converts all text to lowercase to ensure uniformity across the corpus.
2) **Removing Numbers:** Deletes any numeric characters, as they are typically not relevant to the analysis of news authenticity.
3) **Eliminating Punctuation and Special Characters:** Removes punctuation and specific special characters (e.g., '<', '...') that are irrelevant for text analysis.
4) **Excluding English Stopwords:** Removes common English stopwords to highlight more meaningful words within the text.
5) **Removing News Source Names:** Excludes common news source names from the corpus to prevent bias in the analysis.
6) **Applying Stemming:** Reduces words to their root forms, facilitating a more consistent analytical approach.
7) **Removing Extra Whitespaces:** Cleans up the corpus by eliminating superfluous whitespaces.

The implementation of these text cleaning steps is encapsulated within the `clean_text` and `preprocess_corpus` functions, as outlined in the provided code snippet. This meticulous preprocessing ensures the standardization of the text data, preparing it for detailed analytical procedures.

Following the cleaning process with the `preprocess_corpus()` function, the analysis advances to identifying words that are distinctly associated with either 'real' or 'fake' news categories. A chi-square test is conducted to determine the statistical significance of the occurrence of specific words within each category. This step aims to uncover patterns or indicators that could assist in distinguishing between real and fake news.

Besides, we implemented a comprehensive URL normalization process to minimize potential discrepancies that could arise from inconsistent URL formatting. This step was crucial in enhancing the reliability of our analysis, allowing for a more accurate evaluation of the credibility of the news source, free from the biases introduced by URL inconsistencies.

In addition, when there is an image or video link under the entry, the corresponding attribute will be assigned a value of 1, and conversely if there is no link under the attribute it will be assigned a value of 0. BEesides, the original format (URL) of images and videos has been innovatively transformed into categorical variables, making abstract content easier to handle. By integrating disparate datasets, introducing essential variables for differentiation, refining variable representations, and conducting advanced textual analysis, a comprehensive dataset is prepared. This dataset is primed for in-depth analysis in subsequent phases of the study, aimed at unraveling the nuances of news authenticity.

C. Exploratory Data Analysis (EDA)

Our EDA focused on understanding the distribution of key features within the dataset, identifying patterns and characteristics unique to fake and real news articles. This involved analyzing term frequencies in news titles and bodies, examining title lengths, and exploring the presence of specific words and phrases (unigrams and bigrams) that might indicate the authenticity of a news article [17], [18].

D. Machine Learning Models for Classification

In this study, we explored several machine learning models to develop a robust classifier capable of distinguishing between real and fake news effectively. The following models were employed:

1) Multi-Layer Perceptron (MLP): The model is compiled using the 'RMSprop' [19] optimizer and the binary cross-entropy loss function. Accuracy is chosen as the evaluation metric. Training follows 5-fold cross-validation, reducing the learning rate on a plateau to optimize performance. Performance metrics, including loss, accuracy, precision, recall, and F1 score, are reported for each fold, with averages calculated across all folds.

2) Naive Bayes Classifier: Naive Bayes is a probabilistic classifier based on Bayes' theorem, assuming feature independence. It is trained using the naive Bayes function from the e1071 package in R, performing well in many text classification tasks [20]. The accuracy of the Naive Bayes model is assessed by comparing the predicted class labels with the true class labels on the test data, achieving an accuracy of approximately 54%.

3) Random Forest Classifier: The Random Forest classifier is an ensemble method that constructs multiple decision trees during training. By averaging the predictions of multiple decision trees, Random Forest reduces the risk of overfitting, improving generalization to unseen data [21]. In this implementation, 500 decision trees are trained, with a subset of features used at each split. The out-of-bag (OOB) error rate is approximately 26.47%, and final predictions are made through majority voting among the trees.

4) Logistic Regression Classifier: Logistic Regression is a linear model used for binary classification, making it well-suited for fake news detection, which involves determining the truth or falsity of news items. Additionally, it outputs probabilistic assessments, which help quantify uncertainty in predictions [22]. In this implementation, logistic regression is performed using the glmnet function with elastic-net regularization to prevent overfitting and improve generalization. Predictions are made using the logistic function, and class labels are assigned based on a threshold (usually 0.5). Model accuracy is evaluated by comparing predicted labels with actual labels on the test data.

IV. Results and Interpretation

Our comprehensive analysis encompassed various machine learning models to distinguish between real and fake news articles. We assessed the performance of these models based on accuracy, precision, recall, and F1 score [23]. Below, we present our key findings:

A. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial component in research, as it provides reliable guidance for subsequent hypothesis validation and model construction. Previous studies have highlighted that the dissemination of fake news is significantly biased, with sources predominantly concentrated in non-mainstream and social media platforms. This dissemination pattern differs markedly from that of real news. Therefore, the primary objective of conducting EDA in our study is to identify the differences between fake and real news by examining their sources, content complexity, and distinctive characteristics.

We first examine the sources of all news items. As illustrated in Figure 1, both real and fake news sources are displayed. A closer analysis of fake news sources reveals that outlets such as rightwingnews.com and eaglerising.com show a clear preference for fake news, with the number of fake news articles surpassing that of real news. In contrast, politi.co and cnn.it publish the most real news. It is also important to note that some sources of fake news were not labeled. Nevertheless, we retained this portion of the data as it offers additional insights. While the sources of some fake news remain unidentified, all real news originates from reputable outlets familiar to the general public.

Additionally, it is essential to compare whether a given source produces both real and fake news. As shown in Figure 2, we identified eight common sources that publish both types of content, with fake news often being more prevalent. Sources like rightwingnews.com and eaglerising.com clearly favor fake news, whereas freedomdaily.com almost exclusively produces fake news with minimal real news reporting. Conversely, sources such as occupydemocrats.com and conservativebyte.com exhibit a more balanced or smaller distribution between real and fake news.

Figures 3 and 4 show the distribution of the most discriminating words in news headlines and body text across real and fake news, providing important clues to key parts of the exploratory data analysis and informing further diagnosis of real and fake news using AI models.

Figure 3 illustrates the word frequency distribution of the 20 most discriminating words in news headlines. The blue bar represents the word frequency in fake news, while the red bars represent real news. The results indicate that the word “Trump” is prevalent in both fake and real news, but its frequency in fake news is significantly higher, appearing nearly 40 times. This suggests that Trump-related topics are frequently leveraged in fake news, likely due to their controversial nature, which tends to attract more attention. Other words such as “Hillary” and ”Muslim” also appear more often in fake news, highlighting a preference for political and religious topics. In general, fake news tends to use more inflammatory terms (e.g., “bomb,” “refuge”, “shoot”), while real news employs more neutral terms like “debate” and “first”. This difference suggests that fake news often relies on emotionally charged language to capture audience attention.

Further analysis of the 30 most frequent words in the body of news articles shows a more concentrated distribution compared to headlines, especially when comparing fake and real news. As in the headlines, “Trump” and “Hillary” were also the most frequently mentioned keywords in both fake and real news. In the body text, “Trump” has a word frequency of up to 600 times, showing that Trump-related stories are the main carriers of false information. Moreover, the analysis reveals distinct differences in the emotional tone and stance words used in fake versus real news. For instance, neutral words like “said,” “think,” and “like” are more common in real news, reflecting its tendency to present facts and logical arguments. In contrast, fake news frequently uses emotionally manipulative terms such as “bomb” and “refugee.” Additionally, temporal words like “Monday” and “Wednesday” appear more frequently in real news, likely because real news often references specific, verifiable timeframes, whereas fake news

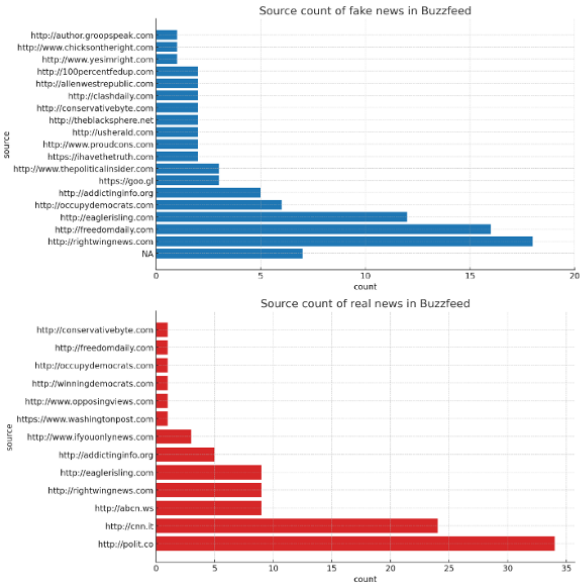

Fig. 1. Statistical bar chart of real and fake news sources

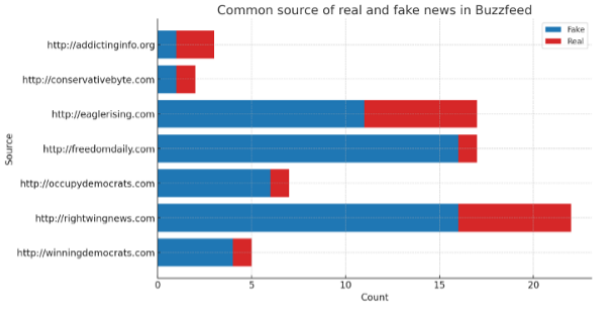

Fig. 2. Comparison chart of common real and fake news sources

tends to rely on vague, unverifiable statements.

Finally, we conducted an in-depth analysis of the most frequent bigrams in the news text, comparing their occurrence in both real and fake news. This analysis highlights significant linguistic differences between the two types of content. Notably, “Donald Trump” is the most common double-word phrase in real news, while “Hillary Clinton” is a high-frequency phrase in fake news. This suggests that while news about these political figures is common across both types of stories, the contexts and narratives surrounding them differ. Some phrases, such as “Young Adults,” “Clinton Foundation,” and “Down Hawkins,” are present in both real and fake news but are more frequent in fake news, indicating that fake news may exaggerate or distort certain topics to appeal to specific audiences or evoke emotional reactions. On the other hand, phrases like “Barack Obama,” “New York,” and “United States” are much more common in real news, suggesting that real news is more focused on actual events, people, and places. The lower frequency of terms like “Story Continued,” “Barack Obama,” and “Book 101” in fake news further suggests that real news is more likely to cover complex or in-depth topics, while fake news often focuses on simpler, attention-grabbing subjects.

B. Model Performance

This section evaluates four classical machine learning classifier models, including Simple Bayes, Random Forest, Logistic Regression, and Multi-Layer Perceptron (MLP), to identify fake news. These models are compared and evaluated using the main performance metrics (Accuracy, Precision, Recall, F1-Score) presented in Table 1 and the confusion matrix and accuracy scores of the four classifiers are extracted. By combining these evaluation metrics, it facilitates a comprehensive understanding of the performance of each model in handling the task of fake news detection.

The Multilayer Perceptron (MLP) performs the best overall among the four models, achieving an accuracy of 90.9%, a precision of 0.833, a recall of 0.556, and an F1 score of 0.667. These results indicate that MLP is able to effectively capture the subtle differences between fake and real news, and its high precision rate and overall F1-Score make it show strong potential in text categorization tasks despite its relatively low recall rate. Therefore, MLP is able to strike a good balance between precision and recall, and shows strong robustness especially when dealing with the task of detecting false news.

Random Forest also shows good performance, with an accuracy of 83.6% and an F1 score of 0.608, indicating that it has a balanced performance in capturing positive and negative samples. The Random Forest model slightly outperforms the logistic regression in terms of the balance between precision and recall.

The logistic regression also performs well with 76.4% accuracy and an F1 score of 0.519, demonstrating its stability in the task of detecting false news. The model performs reasonably well in terms of recall (0.538), but is relatively weak in terms of precision (0.5).

The performance of the plain Bayesian model is weaker, with an F1 score of only 0.387 and an accuracy of 65.4%, indicating that the model has some limitations when dealing with fake news detection. Since the assumption of plain Bayes is premised on the independence between features, and the features in actual linguistic data tend to have strong dependencies, the model performs poorly in capturing the subtle differences in fake news. Moreover, the performance of plain Bayes is limited when confronted with complex text categorization problems, although it performs moderately well in certain simple categorization tasks.

TABLE I
PERFORMANCE COMPARISON OF VARIOUS CLASSIFIERS

Classifiers	Accuracy	Precision	Recall	F1-Score
Naive Bayes	0.654	0.429	0.353	0.387
Random Forest	0.836	0.583	0.636	0.608
Logistic Regression	0.764	0.5	0.538	0.519
Multi Layer Perceptron	0.909	0.833	0.556	0.667

V. CONCLUSIONS

In this project, we conducted detailed exploratory data analysis on a Buzzfeed dataset containing both real and fake

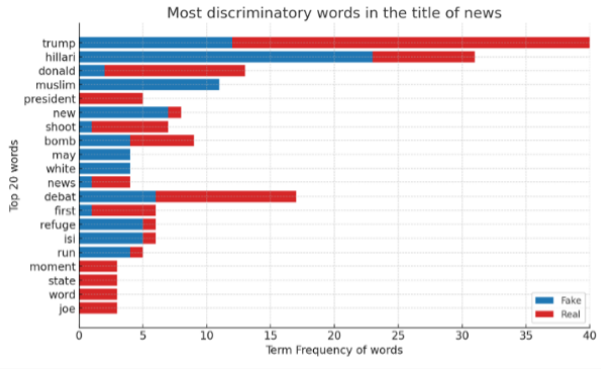

Fig. 3. High Frequency Words in Fake and Real News Titles

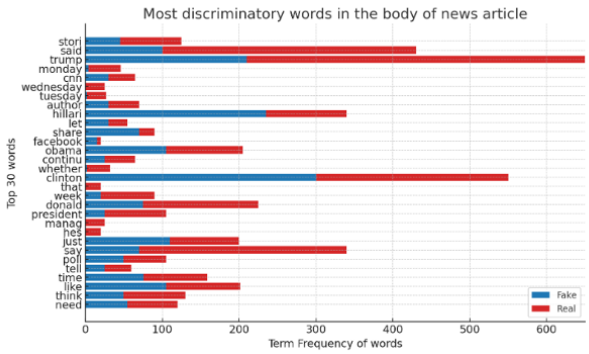

Fig. 4. High Frequency Words in Fake and Real News Bodies

news. Our analysis included generating multiple plots for all variables within each news category. we meticulously explored both unigrams and bigrams to identify distinctive words and phrases commonly found in fake news articles, focusing on both titles and bodies of text. For unigrams, we employed a comprehensive text cleaning process. This included converting all text to lowercase, removing numbers, punctuation, selected special characters (like '<', '...'), and filtering out English stopwords and common news source names to diminish noise. Furthermore, we applied stemming to reduce words to their root forms, thereby simplifying our analysis to more generic language patterns.

For bigram analysis, we took a different approach, recognizing the importance of preserving specific word combinations that may convey more complex meanings, especially in the context of manipulative language often found in fake news. As a result, we opted not to apply the same rigorous cleaning process. We refrained from removing stopwords and stemming in order to retain the linguistic nuances that could be essential in identifying tactics employed in fake news. This decision allowed us to capture a more accurate representation of how language is manipulated in these articles, providing a deeper understanding of the subtleties behind fake news narratives.

Future research in fake news detection could benefit from exploring more advanced machine learning, deep learning, and AI techniques. For instance, integrating transformer-based models like BERT [24] or GPT [25] could enhance the ability to understand context and semantics in fake news. Moreover, a detailed comparison of experimental runtimes [26] has been proposed to better evaluate the computational efficiency of various approaches. These directions aim to optimize detection accuracy, efficiency, and applicability. Additionally, incorporating multi-modal approaches—analyzing not only text but also images, videos, and metadata—could provide a more holistic framework for fake news detection. Another promising direction is the development of explainable AI (XAI) models [27], which would provide insights into how and why certain articles are classified as fake or real, increasing transparency and trust in automated detection systems.

VI. Acknowledgment

This research is supported in part by the research grant from Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

[1] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD explorations newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[2] R. V. Darekar and A. P. Dhande, "Emotion recognition from speech signals using dcnn with hybrid ga-gwo algorithm," *Multimedia Research*, vol. 2, no. 4, pp. 12–22, 2019.

[3] Y. Huang and J. Huang, "A survey on retrieval-augmented text generation for large language models," *CoRR*, vol. abs/2404.10981, 2024.

[4] Y. Huang and J. X. Huang, "Exploring chatgpt for next-generation information retrieval: Opportunities and challenges," *Web Intell.*, vol. 22, no. 1, pp. 31–44, 2024.

[5] X. Zhou, A. Jain, V. V. Phoha, and R. Zafarani, "Fake news early detection: A theory-driven model," *Digital Threats: Research and Practice*, vol. 1, no. 2, pp. 1–25, 2020.

[6] C.-C. Wang, "Fake news and related concepts: Definitions and recent research development," *Contemporary Management Research*, vol. 16, no. 3, pp. 145–174, 2020.

[7] X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Information Processing & Management*, vol. 57, p. 102025, 2020.

[8] H. Laroca, V. Rocio, and A. Cunha, "Does fake news have feelings?," *Procedia Computer Science*, vol. 239, pp. 2056–2064, 2024.

[9] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, 2017.

[10] R. Zafarani, X. Zhou, K. Shu, and H. Liu, "Fake news research: Theories, detection strategies, and open problems," in *Proceedings of the 12th ACM Conference on Web Science*, 2019.

[11] A. D'Ulizia *et al.*, "Fake news detection: A survey of evaluation datasets," *PeerJ Computer Science*, vol. 7, 2021.

[12] H. Abonizio, D. M. JI, T. GM, and B. J. S, "Language-independent fake news detection: English, portuguese, and spanish mutual features," *Future Internet*, 2020.

[13] K. Shu *et al.*, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, pp. 22–36, Jun 2017.

[14] K. Shu, S. Wang, and H. Liu, "Exploiting tri-relationship for fake news detection," *arXiv preprint arXiv:1712.07709*, Dec 2017.

[15] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "Fakenewsnet: A data repository with news content, social context, and dynamic information for studying fake news on social media," *arXiv preprint arXiv:1809.01286*, Sep 2018.

[16] P. Meel and D. K. Vishwakarma, "A temporal ensembling based semi-supervised convnet for the detection of fake news articles," *Expert Systems with Applications*, vol. 177, p. 115002, Oct 2021.

[17] L. I. Meng, L. I. Yanling, and L. I. N. Min, "Review of transfer learning for named entity recognition," *Journal of Frontiers of Computer Science Technology*, vol. 15, 2021.

[18] C. M. Tsai, "Stylometric fake news detection based on natural language processing using named entity recognition: In-domain and cross-domain analysis," *Electronics*, vol. 12, p. 3676, 2023.

[19] N. Ganapathy, Y. R. Veeranki, and R. Swaminathan, "Convolutional neural network based emotion classification using electrodermal activity signals and time-frequency features," *Expert Systems with Applications*, vol. 159, p. 113571, 2020.

[20] M. Granik and V. Mesyura, "Fake news detection using naive bayes classifier," in *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, pp. 900–903, 2017.

[21] R. Jehad and S. A.Yousif, "Fake news classification using random forest and decision tree (j48)," *Al-Nahrain Journal of Science*, 2020.

[22] J. M. Ogdol and B.-L. Samar, "Binary logistic regression based classifier for fake news," 06 2018.

[23] F. Peng, X. Huang, D. Schuurmans, and N. Cercone, "Investigating the relationship between word segmentation performance and retrieval performance in chinese ir," in *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002.

[24] J. Wang, J. X. Huang, X. Tu, J. Wang, A. J. Huang, M. T. R. Laskar, and A. Bhuiyan, "Utilizing bert for information retrieval: Survey, applications, resources, and challenges," *ACM Computing Surveys*, vol. 56, no. 7, pp. 1–33, 2024.

[25] M. T. R. Laskar, M. S. Bari, M. Rahman, M. A. H. Bhuiyan, S. Joty, and J. X. Huang, "A systematic study and comprehensive evaluation of chatgpt on benchmark datasets," *arXiv preprint arXiv:2305.18486*, 2023.

[26] C. Zhang, A. Gupta, X. Qin, and Y. Zhou, "A computational approach for real-time detection of fake news," *Expert Systems with Applications*, vol. 221, p. 119656, 2023.

[27] A. Athira, S. M. Kumar, and A. M. Chacko, "A systematic survey on explainable ai applied to fake news detection," *Engineering Applications of Artificial Intelligence*, vol. 122, p. 106087, 2023.