ITT — Is That True? Towards More Inclusive Tools for Fake News Detection

Denis Tavares da Silva[1], **Gustavo Grandisolli Zwicker**[1], **Robson Bonidia**[1]

[1] Department of Data Science – Faculdade de Tecnologia de Ourinhos
Av. Vitalina Marcusso, 1440, Campus Universitário – Ourinhos – SP – Brazil

denistvrs95@gmail.com, gustavogzwicker@gmail.com,

rpbonidia@gmail.com

Abstract. *With the rise of technology, access to information has become easier than ever. However, this convenience has introduced a significant challenge: the spread of fake news. The post-pandemic scenario has further exacerbated this issue, creating a society more susceptible to encountering fake news or biased and manipulated articles in their daily lives. This phenomenon has become a global problem, posing one of the main threats to various sectors, including politics, education, and the environment. Consequently, Artificial Intelligence (AI) solutions have been developed to classify fake news from true articles. Nevertheless, these solutions have limited reach and room for improvement. To support and increase the use of AI for combating fake news and its spread, we propose an open AI platform, called Is That True (ITT). ITT aims to (1) classify and indicate the likelihood of a statement being truthful or not, using Machine Learning (ML) techniques; (2) promote and popularize the practice of fact-checking; and (3) provide a tool to facilitate this process, making the public less vulnerable to the impacts of fake news. At this moment, we chose to work in both Portuguese and English, given that in Brazil, fake news is a massive problem that is disseminated constantly, and also given the reach of English-based news in a worldwide context. This platform holds significant potential to support global initiatives to combat fake news and its influence on society.*

1. Introduction

According to [McCreadie and Rice 1999], information forms the foundation of societies, enabling continuous human growth and advancement. The accessibility of information affects various facets of our existence, including economic prosperity, privacy protection, decision-making, and policy formulation, as well as our everyday routines [McCreadie and Rice 1999]. Given the importance that information has in our daily lives, the ability to choose trustworthy sources is crucial to individual development of critical thinking.

With the emergence of technologies and their incessant evolution, accessing information has become considerably easier than in the past. However, this accessibility, along with the post-pandemic scenario, has ushered in a society that is more susceptible to encountering fake news or biased and manipulated articles in their daily lives [Aimeur et al. 2023]. Furthermore, the use of social media platforms is progressively increasing. As of 2021, there were over 4.26 billion social media users, and this figure is projected to escalate to approximately 6 billion by 2027 [Dixon 2023]. Consequently, the reach, and impact of information spread throughout digital platforms have reached unprecedented dimensions [Akhtar et al. 2023]. The COVID-19 pandemic showed the risk of fake news in creating panic and preventing people from accessing reliable information on social media, mobile messaging apps, video hosting services, and websites [Balakrishnan et al. 2022, Beauvais 2022]. This massive spread of fake news during the pandemic has been dubbed the "infodemic" [Zarocostas 2020], prompting the World Health Organization to call for the development of international fact-checking organizations [Beauvais 2022].

Moreover, this phenomenon is becoming a worldwide problem, considered one of the main threats to several areas, such as politics, education, environment, health, and science [Aimeur et al. 2023, Raza Shaina 2022]. According to [Di Domenico et al. 2021, Lahby et al. 2022], there are various classes of fake news, such as misinformation, disinformation, rumors, satirical news, fabricated reviews, opinion-driven spam, deceptive advertisements, and conspiracy theories, making its detection challenging. In an extensive literature review [Egelhofer and Lecheler 2019, Ahmad et al. 2022], authors delineate three fundamental dimensions of fake news: (1) diminished factual accuracy (encompassing aspects such as false associations, deceptive content, contrived narratives, distorted context, and impersonation); (2) adherence to journalistic structure (comprising elements like headlines, text, body, and visual components); and (3) the underlying intention to deceive (including motivations of a political/ideological nature, financial motivations, and those aimed at entertainment or provocation).

Regarding this, many studies have been dedicated to combating fake news using Artificial Intelligence (AI), specifically Machine Learning (ML) and Natural Language Processing (NLP). In this context, ML can be used to develop automated systems that analyze online content for signs of misinformation. For example, in current review articles [Varma et al. 2021, Ahmad et al. 2022, Akhtar et al. 2023], authors point to various strategies currently being used, such as sentiment and emotion analysis, font classification, textual entailment features, automated fact-checking, and others. Nevertheless, the composition and stylistic nuances in the writing of new content and articles exhibit variations contingent upon nations, regions, fields, and origins, limiting the results of the studies due to the uniform data used in the analysis [Varma et al. 2021, Zeng et al. 2021, Mishra et al. 2022]. Furthermore, techniques of disinformation are perpetually advancing, necessitating continuous adaptation of detection methodologies [Mohawesh et al. 2023]. Also, there is an absence of user-friendly systems or platforms that allow end users to effectively apply and benefit from the findings of these studies, e.g., platforms publicly available.

To build a society based on truth, fact-checking plays a pivotal role. Fact-checking involves critically evaluating information and verifying its accuracy before accepting and sharing it. Both individuals and institutions must prioritize fact-checking to ensure that the information they consume and disseminate is reliable and evidence-based. Moreover, educating the public about critical thinking, media literacy, and verifying information is crucial in the fight against fake news. Also, fostering a culture of transparency can aid in restoring trust in information sources, and offering accessible services to simplify the fact-checking process, e.g., user-friendly platforms, is a vital step towards a future where fake news can be effectively combated. Considering this, we propose a web application called Is That True? (ITT), where the public can access a fact-checking tool capable of identifying vulnerabilities in a text input or article, providing a probability that indicates the likelihood of the inputted data being fake news.

Taking this into account, ITT holds a significant promise as a tool for aiding the public in combating fake news. Additionally, by suggesting sources to fact-check the information, it can also contribute to enhancing public awareness and encouraging higher participation in a fact-checking culture. This, in turn, can promote responsible consumption of information on social media and other online platforms. At the current stage, we chose to work based on Portuguese and English datasets. The reason for this initiative is rooted in the fact that, as Portuguese is our native language, we are better equipped to catch nuances and work towards quick solutions in the training or dataset selection process. Additionally, we have witnessed the significant impact that fake news had in our country, especially during the pandemic period.

We also chose English as a secondary training language in this testing phase, given the global reach that news in English has on the internet. So we thought it could be interesting to compare the results and see how different, or similar, are the context of the fake news spread locally, nationally, and globally. Furthermore, our objective is not to determine truth, but to assist individuals in selecting the most reliable sources.

Considering this, we hypothesize that it is possible, although challenging, to create a tool that uses state-of-the-art techniques to detect inconsistencies in written sentences. Through this, the tool can detect fake news, alert the public, and thereby diminish the influence of disinformation in our daily lives. Our solution also aligns with the International Grand Committee (IGC) on Disinformation and the European Commission's report on combating fake news and online disinformation. Also, this solution is part of a comprehensive suite dedicated to generating AI that prioritizes positive outcomes and has a meaningful societal impact, connected to AutoAI-Pandemics[1], which was selected as one of the most promising proposals (out of 221 submissions) in a global competition, held by the Global South Artificial Intelligence for Pandemic and Epidemic Preparedness and Response Network (AI4PEP). Finally, this solution won Falling Walls Lab Brazil 2023, competing among the 100 best ideas in the world[2].

2. Related Works

Contemplating ways to enhance performance in the deficiencies and gaps present in current state-of-the-art technologies, our attention turned to exploring existing studies addressing similar challenges. Our objective was to assess the efficacy of their solutions and analyze the overall effectiveness of their methodologies in addressing the identified issues. These issues include the dependence on the quality of available datasets and the necessity to check for any inconsistencies or inherent biases within the chosen datasets, which could lead to detrimental classification results that do not fully reflect reality.

Table 1 presents a comparative analysis of prevalent techniques employed in the existing literature across various studies for classifying Fake News. Moreover, it highlights the distinctive features of our proposed solution, *ITT*, setting it apart from these

[1] http://autoaipandemics.icmc.usp.br/

[2] https://falling-walls.com/discover/videos/breaking-the-wall-of-fake-news-detection/

established methods. This comparative examination presents the contributions that our approach makes in the domain of Fake News classification.

Table 1. Studies used for Fake News detection

Authors	ML Algorithm	Dataset
[Della Vedova et al. 2018]	Naïve Bayes	Facebook (collected by the authors), PolitiFact and BuzzFeed
[Ashraf et al. 2021]	Random Forest, Multi-Layer Perceptron	CLEF2021
[Linmei Hu and Wu 2022]	Bidirectional Encoder Representations from Transformers (BERT-text)	BuzzFeedNews, BuzzFace, CoAID, FakeCovid, CHECKED, among others
[Yahan Ke 2022]	Support Vector Machine, Decision Tree	Million News Headlines, Fake and real news, Getting Real about Fake News (Kaggle)
[Al Asaad and Erascu 2018]	Bag of Words, Term Frequency and Inverse Term Frequency (TF-IDF)	fake-real-news-dataset (George McIntire)
[Mohawesh et al. 2023]	Multilingual Bidirectional Encoder Representations, Multilingual-Fake Model	TALLIP
[Mishra et al. 2022]	Support Vector Machine, Logistic Regression	LIAR, FEVER, Factify

As demonstrated, numerous studies have investigated the potential of ML in classifying fake news. Nevertheless, a notable limitation in these investigations is the little focus on the semantic aspects of the data, without deepening the pragmatic field. According to [Yuan et al. 2023], all existing methods have their limitations, and an urgent problem we need to solve to improve fake news detection is how best to combine and improve them. Pragmatic analysis is an interesting alternative because it enables the model to understand subtle nuances of the text, making it more adept at classifying fake news, especially when applied to real and current news. As concurred by [Yan Li and Liu 2021], pragmatic analysis is identified as the most challenging linguistic dimension in NLP due to the extensive context and deep linguistic understanding it demands.

About this, [Yan Li and Liu 2021] says that pragmatic insertion studies only represent about 8% of the total, and only when we advance further in this field will we be able to build a model capable of understanding subtleties and intentions, which is essential for analyzing fake news in the real world. Additionally, there's a growing need for methods capable of detecting fake news in its early stages of dissemination. The current state-of-the-art solutions are often limited to identifying fake news 12 hours or more after its creation, which, according to [Raza Shaina 2022], may be delayed due to the rapid spread of misinformation. Therefore, we believe that ITT can be a strong ally in helping to strengthen the culture of fact-checking in the population.

3. Research Plan – Technologies, Methods, and Algorithms

3.1. Dataset Selection

Having thoroughly reviewed the literature on methods for classifying fake news, we look for robust and reliable databases to develop a model and rigorously test our findings through various techniques. We decided to start our project with the well-known Portuguese-based Corpus, named Fake.Br [Santos et al. 2018]. This decision was based on the dataset's broad recognition and detailed examination by numerous researchers. Furthermore, the collaborative creation of this dataset by the University of São Paulo (USP) and the Federal University of São Carlos (UFSCAR) at the Interinstitutional Center for Computational Linguistics (NILC) enhanced its credibility. It is important to highlight that, despite its reliability, Fake.Br is subject to a temporal bias, as it contains articles collected from January 2016 to January 2018.

The creators of the dataset highlighted the painstaking and largely manual data collection process necessary to ensure the quality of the corpus and its effectiveness in distinguishing between true and false articles. The corpus includes 7,200 news, split equally with 3,600 true news stories and 3,600 fake news, all in plain text format, with each article saved in a separate file. Nevertheless, After some exploration and experimentation, we saw the need to expand the database with short-length true labeled data. For this, small specific true data and large false data were included from a more recent database, FakeTrueBR [Chavarro et al. 2023]. Subsequently, we expanded our focus to include English-language fake news classification. Employing the same metrics used for selecting the Fake.Br Corpus, we chose the LIAR dataset [Wang 2017] for our model training, which consists of 12,836 short statements labeled according to their credibility.

3.2. Exploratory Data Analysis and Preprocessing

After determining the datasets for training, our focus shifted to understanding fundamental aspects of the data, such as:

- How is the data distributed within the datasets?
- What is the quality of the available data?
- What characteristics are essential for analyzing the data?
- Are there any corrections needed before utilizing the data?

To answer these questions, we conducted an Exploratory Data Analysis (EDA) on LIAR and Fake.Br. The results were intriguing, revealing an unexpected issue that warrants further exploration. The articles classified as true, have in their totality, generally more than 400 words, but in the contrary direction, the articles classified as fake, generally have less than 25 words. That is a problematic aspect because it can teach our model a quality that is not necessarily truthful: **That short texts are inherently false, and longer texts, are inherently true.**

Nevertheless, we would like to emphasize that the prevalence of short fake news is understandable. In Brazil, fake news is primarily spread through social media platforms such as WhatsApp, Facebook, and Telegram. For these media to spread quickly, the content needs to be short, use strong language, and lead readers to make quick judgments based more on emotion than on facts. Consequently, most examples of fake news available on the internet tend to be brief and shocking, while true news often provides context and aims to explain events more reasonably and factually. This fact shows that there is an urgent need to work towards more diversified datasets, that can diminish inherent biases so that Fake News classification can reflect reality more closely, and in doing so, be able to help the public in the fight against misinformation

Thereby, after conducting our EDA and gaining insights into the quality of the data, we moved forward to preprocess the data, aiming to enhance its quality and prepare it for model training. This preprocessing step is crucial as it can significantly impact the performance and accuracy of the models we intend to build. The objective is to refine the raw data, address any inconsistencies, and extract meaningful features that contribute to the robustness of our models. To assess the effectiveness of our preprocessing techniques, we utilized word clouds to visualize the word frequency distribution in both the LIAR and Fake.Br before and after preprocessing. Word clouds provide a graphical representation of the most frequent words in a corpus.

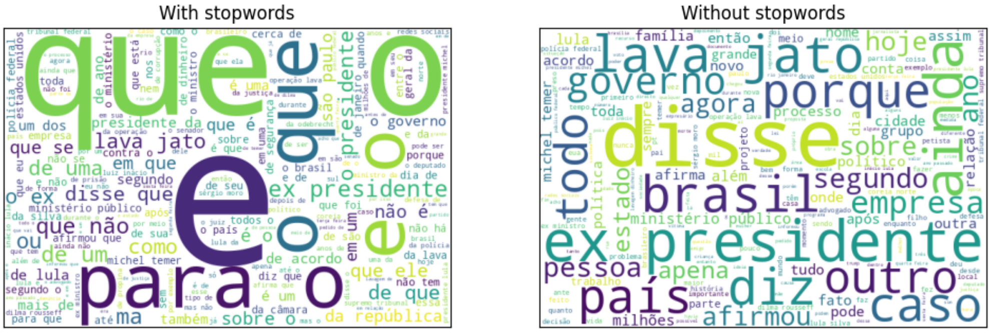


Figure 1. Wordclouds showing the most frequent words in the database

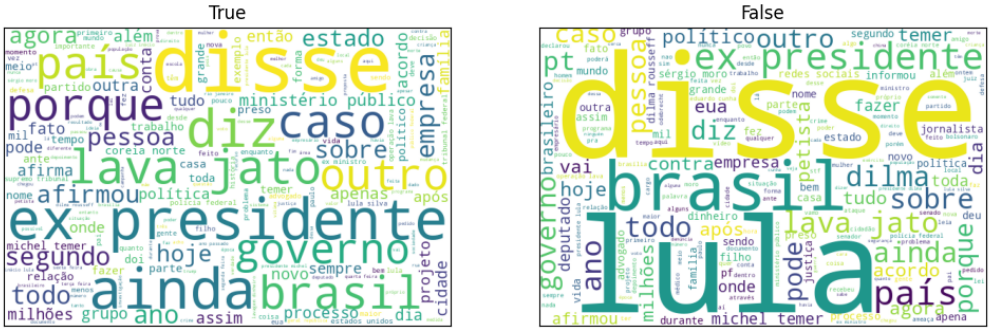


Figure 2. Wordclouds showing the most frequent words by target

The shift in word frequency before and after preprocessing is notable. On the right, before the pruning, we can see that the bigger words (that appear with a higher frequency in the text), are general terms such as 'a', 'or', 'that', and 'no', but after the preprocessing technique being applied, we can see some more interesting words appearing. This highlights the refinement achieved through techniques such as stop word removal and other text-cleaning processes, which are vital for reducing the dimensionality of the data. Finally, for the numerical representation of words, we opted for Global Vector (GloVe). According to [Pennington et al. 2014], GloVe's ability to capture linear substructures in the data and efficiently handle global statistics makes it stand out.

4. Experimental Results

Our next step involved training several models on the above-mentioned datasets to evaluate how effectively they could classify real information based on the features and information gleaned from the data. We initially focused on the Portuguese corpus, Fake.Br. In line with insights from the literature analysis, we opted to construct the model using Gradient Boosting (XGBoost), Gated Recurrent Unit (GRU), and Long Short-Term Memory (LSTM) techniques. These methods were chosen because prominent works in the literature consistently highlight their effectiveness in Fake News classification, often yielding promising results. The figure below illustrates the performance of these different algorithms across three key metrics: Recall, Precision, and Accuracy.

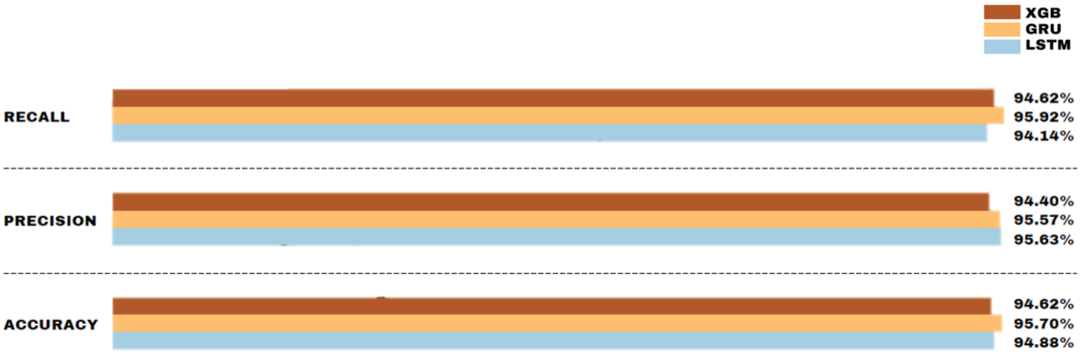


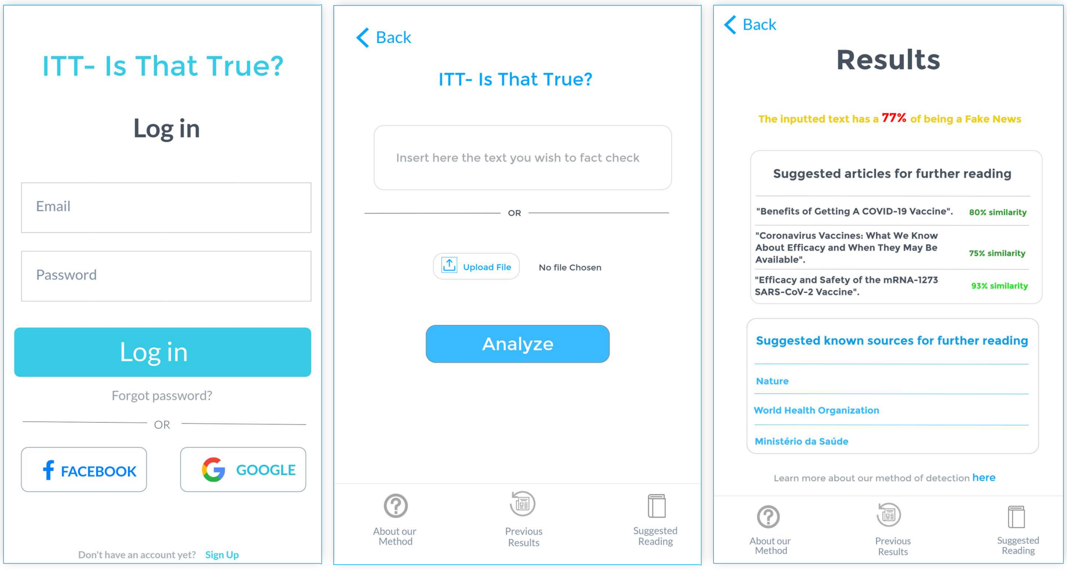
Figure 3. Graphic displaying the performance of the chosen methods for the Fake.Br Corpus

Upon analysis, it becomes evident that the GRU algorithm outperformed the other two across all metrics. Having sufficiently explored the Portuguese dataset, we proceeded to test the model using real news. Interestingly, the model demonstrated effectiveness with political articles and longer texts, possibly influenced by an inherent bias in the dataset originating from a political scenario in Brazil. However, when confronted with shorter texts (200 words or fewer), the model struggled and failed to accurately classify whether the input was Fake News or not.

4.1. ITT Prototype

In line with our goal to contribute to societal improvement, we aim to offer tools that facilitate fact-checking and identify reliable information sources. To this end, we developed our web application designed to help users discern the veracity of information. ITT is able to assess the likelihood of information being false, while also providing reports highlighting areas where similarities are detected with known cases of fake news. This approach empowers individuals to not only more easily evaluate reliable information, but also to critically evaluate news. For a visual representation of ITT, see Figure **??**.

Our initial concept comprises a login screen, an analysis screen, and a results screen. These interfaces serve to guide users through the process of inputting data into our model, receiving recommendations for further reading, and exploring highlighted areas of concern within the text that our analysis has flagged as potentially containing fake news. While we have developed a prototype capable of classifying texts in both Portuguese and English with some degree of success, as detailed extensively in this study, ITT is not yet



The inputted text has a 77% of being a Fake News

"Benefits of Getting A COVID-19 Vaccine".	80% similarity
"Coronavirus Vaccines: What We Know About Efficacy and When They May Be Available".	75% similarity
"Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine".	93% similarity

Figure 4. Concept of how the user interface will look to the end-user

ready for public release. Numerous refinements and implementations are still required to ensure that the tool can reliably handle the classification of fake news in real-world scenarios. We wish to achieve a level of confidence in our work that assures us of its ability to contribute positively to society.

5. Conclusions

As previously emphasized, data quality is fundamental in building and training a reliable and robust ML model. Unfortunately, a notable challenge arises of high-quality, recent, and well-balanced labeled data that covers the necessary variability in the represented classes. This scarcity becomes a significant obstacle in effectively classifying fake news. Furthermore, a critical aspect that deserves attention is the imbalance present in both the LIAR and Fake.Br Corpus. Notably, the data classified as false in these datasets tend to have shorter text lengths. This presents a potential pitfall, as the model may develop an inherent bias favoring shorter texts, influencing its classifications when deployed in real-world scenarios.

Furthermore, it is worth noting that while semantic models are widely used in efforts to classify fake news, they are inherently limited in their ability to comprehensively represent reality. Recognizing these theoretical limitations, there is a pressing need to advance models capable of working with and interpreting data based on pragmatic cues embedded in the texts. Only through the development of such models can we genuinely combat and identify what constitutes fake news. This underscores the importance of evolving beyond semantic approaches and delving into the nuances of language and context to enhance the accuracy and reliability of fake news classification models.

References

Ahmad, T., Aliaga Lazarte, E. A., and Mirjalili, S. (2022). A systematic literature review on fake news in the covid-19 pandemic: Can ai propose a solution? *Applied Sciences*, 12(24):12727.

Aimeur, E., Amri, S., and Brassard, G. (2023). Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30.

Akhtar, P., Ghouri, A. M., Khan, H. U. R., Amin ul Haq, M., Awan, U., Zahoor, N., Khan, Z., and Ashraf, A. (2023). Detecting fake news and disinformation using artificial intelligence and machine learning to avoid supply chain disruptions. *Annals of Operations Research*, 327(2):633–657.

Al Asaad, B. and Erascu, M. (2018). A tool for fake news detection. *2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pages 379–386.

Ashraf, N., Butt, S., Sidorov, G., and Gelbukh, A. F. (2021). Cic at checkthat! 2021: Fake news detection using machine learning and data augmentation. pages 446–454.

Balakrishnan, V., Ng, W. Z., Soo, M. C., Han, G. J., and Lee, C. J. (2022). Infodemic and fake news–a comprehensive overview of its global magnitude during the covid-19 pandemic in 2021: A scoping review. *International Journal of Disaster Risk Reduction*, 78:103144.

Beauvais, C. (2022). Fake news: Why do we believe it? *Joint bone spine*, 89(4):105371.

Chavarro, J. P., Carvalho, J. T., Portela, T. T., and Silva, J. C. (2023). Faketruebr: Um corpus brasileiro de notícias falsas. *Escola Regional de Banco de Dados (ERBD), 18.*, pages 108–117.

Della Vedova, M. L., Tacchini, E., Moret, S., Ballarin, G., DiPierro, M., and De Alfaro, L. (2018). Automatic online fake news detection combining content and social signals. pages 272–279.

Di Domenico, G., Sit, J., Ishizaka, A., and Nunan, D. (2021). Fake news, social media and marketing: A systematic review. *Journal of Business Research*, 124:329–341.

Dixon, S. (2023). Number of social media users worldwide from 2017 to 2027. 202027.

Egelhofer, J. L. and Lecheler, S. (2019). Fake news as a two-dimensional phenomenon: A framework and research agenda. *Annals of the International Communication Association*, 43(2):97–116.

Lahby, M., Aqil, S., Yafooz, W. M., and Abakarim, Y. (2022). Online fake news detection using machine learning techniques: A systematic mapping study. *Combating Fake News with Computational Intelligence Techniques*, pages 3–37.

Linmei Hu, Siqi Wei, Z. Z. and Wu, B. (2022). Deep learning for fake news detection: A comprehensive survey. *AI Open*, 3:133–155.

McCreadie, M. and Rice, R. E. (1999). Trends in analyzing access to information. part i: cross-disciplinary conceptualizations of access. *Information Processing & Management*, 35(1):45–76.

Mishra, S., Suryavardan, S., Bhaskar, A., Chopra, P., Reganti, A., Patwa, P., Das, A., Chakraborty, T., Sheth, A., Ekbal, A., et al. (2022). Factify: A multi-modal fact verification dataset.

Mohawesh, R., Maqsood, S., and Althebyan, Q. (2023). Multilingual deep learning framework for fake news detection using capsule neural network. *Journal of Intelligent Information Systems*, pages 1–17.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. pages 1532–1543.

Raza Shaina, D. C. (2022). Fake news detection based on news content and social contexts: a transformer-based approach. *International Journal of Data Science and Analytics*, 13.

Santos, R. L. S., Monteiro, R. A., and Pardo, T. A. S. (2018). The fake . br corpus-a corpus of fake news for brazilian portuguese.

Varma, R., Verma, Y., Vijayvargiya, P., and Churi, P. P. (2021). A systematic survey on deep learning and machine learning approaches of fake news detection in the pre-and post-covid-19 pandemic. *International Journal of Intelligent Computing and Cybernetics*, 14(4):617–646.

Wang, W. Y. (2017). "liar, liar pants on fire": A new benchmark dataset for fake news detection. pages 422–426.

Yahan Ke, Ruyi Qu, X. L. (2022). Classification of fake news headline based on neural networks. *CoRR*.

Yan Li, M. a. T. and Liu, D. (2021). From semantics to pragmatics: where is can lead in natural language processing (nlp) research. *European Journal of Information Systems*, pages 569–590.

Yuan, L., Jiang, H., Shen, H., Shi, L., and Cheng, N. (2023). Sustainable development of information dissemination: A review of current fake news detection research and practice. *Systems*, 11.

Zarocostas, J. (2020). How to fight an infodemic. *The lancet*, 395(10225):676.

Zeng, X., Abumansour, A. S., and Zubiaga, A. (2021). Automated fact-checking: A survey. *Language and Linguistics Compass*, 15(10):e12438.