

Towards a News Authenticity Predictor (NAP AI) [†]

Arif Wali ^{1,*}, Stelios Kapetanakis ² and Giacomo Nalli ¹ 

¹ Computer Science, Science and Technology, Middlesex University London, London NW4 4BT, UK; g.nalli@mdx.ac.uk

² Distributed Labs, Distributed Analytics Solutions, 30 Churchill Pl, London E14 5RE, UK

* Correspondence: aw1197@live.mdx.ac.uk

[†] Presented at the 6th International Electronic Conference on Applied Sciences, 9–11 December 2025; Available online: <https://sciforum.net/event/ASEC2025>.

Abstract

The rapid spread of misinformation on social media has emerged as a major societal issue. Over 40% of British social media news-sharers admitted they had shared inaccurate or fake news. The extensive distribution of false information causes public trust deterioration while modifying public opinions and potentially destabilizing social and political systems. There are profound challenges due to this hard-to-detect, hard-to-stop reality and the financials and societal implications are remarkable. As an attempt to limit the challenges created from misinformation this paper introduces some preliminary work on detection of fake news and verification of their reliability based on online content. Large language models (LLMs) are being used along with natural language processing (NLP) techniques to evaluate news articles through their linguistic and contextual characteristics. Several models are compared on how they can typically identify typical indicators of misinformation through the analysis of extensive verified datasets to develop an ability to classify content as authentic or fabricated. This work has been through thorough testing to determine its operational effectiveness and dependability after completion. We present a relatively easy-to-use tool which enables a wide range of people also for those without a background in computer science to easily verify news accuracy before sharing or trusting it. This work could help to stop false information from spreading while promoting fact-based discussions and improving digital literacy skills. The research demonstrates how technology fights the fake news crisis to create an informed digital environment which supports public conversation protection and information integrity in the modern digital age.

Keywords: fake news detection; LLM; natural language processing; misinformation; spot unverified claims

1. Introduction

The internet revolutionizes information sharing and access by creating faster and broader communication channels that reach more people than ever before [1]. However, this rapid increase in online material was followed by gradual decrease and even elimination of verification mechanisms, resulting in high misinformation spread [2]. Most platforms currently lack any dependable system to verify the accuracy of shared content. Political groups along with conspiracy theorists and people who want personal benefits exploit this information gap to distribute manipulated content which supports their goals [3]. Recent studies show that fake news is included in the content shared on social media platforms [4]. This demonstrates the serious extent of misinformation since it shows how deeply entrenched false information has become in modern digital society.



Academic Editor: Lucia Billeci

Published: 24 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Different groups, especially political teams and thought-based movements, have created social media squads. Their specific aim is to craft and share stories that match their objectives. Many times, this means intentionally spreading false news or unproven theories. This causes confusion, creates divisions, and tricks people emotionally. The average person has few means to confirm the honesty of the content they come across. The circulation of misinformation and conspiracy theories confuses, divides, and emotionally manipulates the public [5]. This lack of clearness hurts public trust, warps engaging in democracy, and shakes up society's agreement on fact and truth [6].

This work presents News Authenticity Predictor (NAP AI, will refer to it also as BiasBreak), a system to detect false news, supported by artificial intelligence as a solution to the pressing problem of fake news. Using natural language processing (NLP) and machine learning, BiasBreak can examine the text patterns as well as further details of online content to evaluate its trustworthiness. This work is structured as follows: Section 2 covers the literature review, Section 3 describes the followed methodology, Section 4 presents the promising results whilst Section 5 accommodates the conclusion and future work.

2. Literature Review and Related Work

Internet allows for fast access of multi resources for any given search or enquiry. However, this easy access creates a pressing issue, the one of confirming whether the encountered information is true or not. In the digital age, it is increasingly hard to tell if information is believable, untrue, or deliberately false. Misinformation spread mainly through social media due to how quickly and widely content is shared [3].

Recent studies show how teenagers struggle to differentiate real and fake content online [7]. This shows a remarkable weakness in digital smarts, especially with the younger folks. Most users often lack the tools or knowledge needed to fairly judge what they read or share. The vast amount of content created every minute is jaw dropping and cannot be checked by experts. Old or limited detection tools also cannot handle its complexity and rapid frequency.

Efforts are being made by major tech platforms to tackle this issue. Companies such as Google and Microsoft are constantly working on improving their algorithms to suppress misleading content in search results. Some platforms also offer manual complaint and spam reporting forms, enabling users to flag deceptive information [8]. Similarly, social media giants like Facebook (now Meta) and Twitter (now X) are exploring AI-based solutions to detect, flag, and reduce the visibility of fake news on their platforms [9]. Despite these initiatives, studies continue to suggest that approximately most of the online content shared through social platforms is either misleading or outright false [10].

There are tools like Snopes [11] and FactCheck [12] that check online posts for authenticity. However, their process takes time since human experts should do the checking. Scalability is also an issue since they cannot verify multi sources at once. There are some sophisticated verification tools available, but they do not always get updated and do not always learn as well as they should.

This work presents an AI-powered tool, called News Authenticity Predictor, as an approach to address the challenge of fake news by using indicators for misinformation, biasness, and unverified claims in the provided content.

3. Methods and Methodology

The goal of this work is to develop a News Authenticity Predictor by utilizing large language models (LLMs) and sophisticated natural language processing (NLP) techniques. This section presents data collection and pre-processing, dataset selection, and used algorithms.

3.1. Data Collection and Preprocessing

Along with real-time news accumulated from multiple online sources, a varied dataset of verified real and fake news articles, biased and unbiased content, clickbait titles, and specific claims with and without references has been sourced from various portals. To eliminate unnecessary or redundant information, collected articles will go through deduplication, language detection, and metadata extraction. The preprocessing pipeline includes the following two processes:

Process 1: Tokenization and Lemmatization: breaking text into meaningful units and reducing words to their base form for consistency.

Process 2: Vectorization and Feature Engineering: The primary method involves leveraging transformer-based contextual embeddings generated directly by the BERT model. This approach was chosen over traditional techniques like TF-IDF or Word2Vec to achieve a more nuanced and accurate text representation for the final classification task.

3.2. Dataset Selection

The dataset used for training and evaluation combined multiple publicly available misinformation datasets with newly collected examples:

- Authentic news articles from reputable outlets such as BBC, Reuters, and The Guardian.
- Fake news articles identified by fact-checking organizations such as Snopes [11] and FactCheck.org [12].
- Biased content exhibiting strong partisan framing.
- Clickbait headlines designed to attract attention but are often misleading.
- Specific factual claims with or without credible references.

Data were collected using automated API processes and supplemented with manually curated entries from existing datasets (e.g., LIAR, FakeNewsNet).

3.3. Text Processing

Preparing text for BERT differs significantly from traditional NLP preprocessing. Instead of aggressively cleaning the text by removing stopwords or punctuation, we leverage a specialized BERT tokenizer (specifically, the WordPiece tokenizer). This process was essential for aligning the input data with the format the model was pre-trained on.

3.4. Feature Generation via BERT Embeddings

Traditional feature extraction techniques like TF-IDF were superseded by a more potent, integrated approach for a complex model like BERT. The feature extractor is the model itself. To produce contextualized embeddings, the input text was processed by passing it through BERT's pre-trained layers. For every token in the input, these were dense numerical vectors whose value is derived from the words around it. This procedure enables the model to:

- Understand complex linguistic patterns, sentence structure, and nuance much better than word-frequency-based methods;
- Accurately capture the semantic meaning of words in their context, differentiating between different uses of the same word.

The final classification layer of the model uses these rich embeddings as input and refines it to determine whether a news article is authentic or fraudulent. The best features are produced for the classification task thanks to this end-to-end method.

3.5. Model Selection

The BERT (Bidirectional Encoder Representations from Transformers) model was chosen as the central component of the classification engine for the News Authenticity Predictor. Since misinformation is complex and frequently subtle, it calls for a model with deep contextual linguistic understanding, which is why BERT is the best option for this work. In misinformation detection tasks, the original BERT architecture [13] has been widely used, and numerous studies have reported higher accuracy than previous models. It is especially good at spotting the narrative framing and subtle linguistic clues typical of fake news due to its bidirectional attention mechanism. Studies comparing different methods for spotting fake news [14] have found that transformer-based models like BERT tend to do a better job than older techniques like Support Vector Machines (SVMs) or Logistic Regression with TF-IDF features. BERT's ability to understand the context of words helps it achieve higher accuracy scores. While other versions like RoBERTa and DistilBERT also perform well on standard tests, BERT strikes a good balance between keeping things accurate and not requiring excessive computational resources.

Model selection was based on its literature strengths in contextual understanding, Transfer Learning, performance and Feature Generation. Specifically:

- Contextual Understanding: Unlike models that process text in a single direction, BERT's bidirectional nature allows it to understand the full context of a word by examining the words that come before and after it. Jahan et al. [14] and Devlin et al. [15] demonstrate that deep bidirectional pre-training in BERT, which allows both left and right context, yields substantial gains over prior unidirectional LSTM/transformer language models on diverse benchmarks such as GLUE and SQuAD.
- Transfer Learning: This work uses a pre-trained BERT model that Google trained on a sizable corpus of text, such as the BookCorpus and Wikipedia. Jahan et al. [14] support that such approach provides human comprehension of language. This pre-trained model is further hyper-tuned using the dataset we covered in previous sections. With this method, we can attain accuracy without having computational costs associated with starting from scratch to train such a large model.
- High performance: For a variety of natural language processing (NLP) tasks, such as text classification, BERT and its derivatives continuously stand as the industry standard. By choosing BERT, this work strives for maximum accuracy, going beyond straightforward keyword matching to a more thorough content analysis.
- End-to-end Feature Generation: The feature engineering pipeline is made simpler by BERT. Gomes, et al. [13] show that the model internally creates rich, contextualized word embeddings (numerical representations) for the input text rather than depending only on TF-IDF. By capturing intricate linguistic patterns that are far more informative than word frequency counts, these embeddings enable the model to learn the subtle characteristics that set trustworthy news apart from not true material.

3.6. Model Development and Fine-Tuning

Our methodology is based on a mixed method that uses both internal linguistic analysis and external evidence gathering to figure out if a news article is real. This process has the following two main parts: Initial Analysis with a Fine-Tuned LLM and External Verification using a Search API.

Initial Analysis with a Fine-Tuned LLM: A core large language model (e.g., BERT, Llama) is fine-tuned on our labeled dataset of credible and fake news. This model is trained to recognize the linguistic patterns, stylistic cues, and semantic structures commonly associated with misinformation. This initial analysis provides a preliminary assessment of the content's likely authenticity based purely on its inherent characteristics.

External Verification using a Search API (Retrieval-Augmentation): For claims that require external validation, the system employs a retrieval-augmentation mechanism. The system programmatically queries a search engine using a third-party service like SerpApi. SerpApi is used to accumulate Google Search results in real-time, gathering snippets from other news outlets, fact-checking sites, and reputable sources related to the key claims in the article. The retrieved search results are then fed back to our language model as additional context and the LLM analyzes this new evidence, cross-checking the original article's claims against the external sources to make a more informed and evidence-based final classification.

This two-step procedure produces a strong and dynamic fact-checking engine by utilizing both a refined model and real-time web search results. PyTorch (version 2.6.0) and Tensor-Flow (version 2.19.0), which use the Hugging Face library (version 4.51.3) for model management, are the main development frameworks used for training and deployment.

Architectural Roadmap and Continuous Learning

The architecture is made to possibly integrate models through external APIs, like those offered by the Google AI Platform or Groq (for high-speed inference). This makes it possible to scale in the future and use more cutting-edge models without having to host them locally. The system also has a feature that allows users to report inaccurate classifications. To ensure that the model consistently adjusts to new types of misinformation, this feedback is gathered and used to produce better versions of the training dataset for subsequent iterations.

3.7. Testing and Evaluation

Even though BERT is a powerful large language model, for this work, we fine-tuned it specifically for a simple, two-way task: figuring out if news is real or fake. We added a little extra layer called a classification head on top of BERT's [CLS] token output. This layer gives us a probability for each class, so we can see how likely it is that the news is either real or fake. We then evaluated how our model does using common metrics like accuracy, precision, recall, and F1-score, based on the predictions from that classification head. These are standard in fake news detection studies [10] and make it easy to compare how well our approach stacks up against other traditional machine learning methods. A multifaceted evaluation approach used to thoroughly confirm the model's efficacy and dependability including Standard Performance Metrics, Cross-Validation and User Testing.

4. Experimental Results

BiasBreak uses a Fine-tuned BERT model in conjunction with an external verification mechanism, which is thoroughly evaluated in this section. A public testing phase was used to validate the model's real-world usability after it was tested on the test set to determine its efficacy on unseen data.

4.1. Model Performance Metrics

A standard set of classification metrics was used to assess the model's accuracy in classifying article content. The outcomes demonstrate how effective the deep learning strategy is.

- Accuracy: The model achieved an overall accuracy of 98.03%. This score represents the percentage of all articles correctly classified;
- Precision: The precision for the "fake" class was 98.15%. This measures the model's reliability when it flags an article as fake;
- Recall: The recall for the "fake" class was 98.15%. This measures the model's ability to find all the fake articles in the dataset;

- F1-Score: the F1-score, the harmonic mean of precision and recall, was 98.15%, providing a single, balanced measure of the model's overall performance.

4.2. Confusion Matrix Analysis

We have created a confusion matrix (Table 1) to provide a deeper insight into the model's pre-diction behavior.

Table 1. Confusion matrix of BiasBreak results.

Category	Predicted Real	Predicted Fake
Actual Real	511	11
Actual Fake	585	11

This analysis shows that:

- True Negatives (511): The model correctly identified 511 real news articles as real;
- True Positives (585): The model correctly identified 585 fake news articles as fake;
- False Positives (11): Eleven real articles were misclassified as fake, representing false alarms where legitimate content could be wrongly flagged;
- False Negatives (11): Eleven fake articles were misclassified as real, representing missed detections where misinformation slipped through undetected.

The fact that false positives and false negatives are about equal shows that the system is not biased toward labeling content as mainly real or fake. Basically, the model does a pretty good job at balancing things out. Out of 1118 articles, only 22 got mixed up about 2%, and those mistakes were spread evenly between real and fake stories. This balance means the system is not overly aggressive about marking stories as fake and still catches most of the fake news, so you can count on it being fairly accurate and trustworthy in real-world scenarios.

Findings from System Testing

The system consistently gave predictions that were pretty accurate, staying in line with the numbers we saw in Section 4.1. It took about 25 s on average to handle each request, which shows it is pretty good for real-life use. When we tested how tough it was, it handled different kinds of input well, although it sometimes struggled with paywalled articles where the full text could not be pulled in.

5. Conclusions

This work successfully developed and delivered a standout News Authenticity Predictor. It uses a mix of a Fine-tuned BERT system for deep language understanding and a SerpAPI checking tool. The aim was to spot and highlight misinformation and biasness in content, doing an excellent job at that. The results show adequate model performance achieving accuracy and F1 score > 90%. This work shows technically robust promise which needs to be evaluated further with real users and a wider source of information to prove its applicability. Future work aims to investigate this further in both directions.

Author Contributions: Conceptualization, A.W., S.K. and G.N.; methodology, A.W., S.K. and G.N.; software, A.W., S.K. and G.N.; validation, A.W., S.K. and G.N.; formal analysis, A.W., S.K. and G.N.; investigation, A.W., S.K. and G.N.; resources, A.W., S.K. and G.N.; data curation, A.W., S.K. and G.N.; writing—original draft preparation, A.W., S.K. and G.N.; writing—review and editing, A.W., S.K. and G.N.; visualization, A.W., S.K. and G.N.; supervision, A.W., S.K. and G.N.; project administration, A.W., S.K. and G.N.; funding acquisition, A.W., S.K. and G.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is unavailable due to privacy restrictions.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Oeldorf-Hirsch, A.; Sundar, S.S. Social and Technological Motivations for Online Photo Sharing. *J. Broadcast. Electron. Media* **2016**, *60*, 624–642. [CrossRef]
- Allcott, H.; Gentzkow, M. Social media and fake news in the 2016 election. *J. Econ. Perspect.* **2017**, *31*, 211–236. [CrossRef]
- Lazer, D.M.J.; Baum, M.A.; Benkler, Y.; Berinsky, A.J.; Greenhill, K.M.; Menczer, F.; Metzger, M.J.; Nyhan, B.; Pennycook, G.; Rothschild, D.; et al. The science of fake news. *Science* **2018**, *359*, 1094–1096. [CrossRef] [PubMed]
- Pierri, F.; Ceri, S. False news on social media: A data-driven survey. *ACM Sigmod. Record.* **2019**, *48*, 18–27. [CrossRef]
- Enders, A.M.; Uscinski, J.E.; Seelig, M.I.; Klostad, C.A.; Wuchty, S.; Funchion, J.R.; Murthi, M.N.; Premaratne, K.; Stoler, J. The relationship between social media use and beliefs in conspiracy theories and misinformation. *Political Behav.* **2023**, *45*, 781–804. [CrossRef] [PubMed]
- Tandoc, E.C.; Lim, Z.W.; Ling, R. Defining “fake news”: A typology of scholarly definitions. *Digit. J.* **2018**, *6*, 137–153. [CrossRef]
- McGrew, S.; Breakstone, J.; Ortega, T.; Smith, M.; Wineburg, S. Can students evaluate online sources? Learning from assessments of civic online reasoning. *Theory Res. Soc. Educ.* **2018**, *46*, 165–193. [CrossRef]
- Chakraborty, M.; Pal, S.; Pramanik, R.; Chowdary, C.R. Recent developments in social spam detection and combating techniques: A survey. *Inf. Process. Manag.* **2016**, *52*, 1053–1073. [CrossRef]
- Pennycook, G.; Rand, D.G. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 2521–2526. [CrossRef] [PubMed]
- Zhou, X.; Zafarani, R. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Comput. Surv.* **2020**, *53*, 1–40. [CrossRef]
- Snopes.com: Snopes.com 2025. Available online: <https://www.snopes.com> (accessed on 24 January 2026).
- FactCheck.org: FactCheck.org. 2025. Available online: <https://www.factcheck.org> (accessed on 24 January 2026).
- Gomes, L.d.S.; Torres, R.d.S.; Côrtes, M.L. BERT- and TF-IDF-based feature extraction for long-lived bug prediction in FLOSS: A comparative study. *Inf. Softw. Technol.* **2023**, *163*, 107217. [CrossRef]
- Jahan, M.S.; Khan, H.U.; Akbar, S.; Farooq, M.U.; Gul, S.; Amjad, A. Bidirectional Language Modeling: A Systematic Literature Review. *Sci. Program.* **2021**, *2021*, 6641832. [CrossRef]
- Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.