Fake News Detection-Real Time Pre-Publication Analysis

Sanmuga Priya M
Assistant Professor,Dept. of CSD
Sri Krishna College of Engineering and Technology
Coimbatore,India
priya.06889@gmail.com

Roshini K
Dept. of CSD
Sri Krishna College of Engineering and Technology
Coimbatore,India
roshiniroshinik15@gmail.com

Sandhya T
Dept. of CSD
Sri Krishna College of Engineering and Technology
Coimbatore,India
tsandhya2404@gmail.com

Sashmitha R
Dept. of CSD
Sri Krishna College of Engineering and Technology
Coimbatore,India
sashmitharamesh16@gmail.com

Subiksha A
Dept. of CSD
Sri Krishna College of Engineering and Technology
Coimbatore,India
subikshasubi125@gmail.com

***Abstract* — The swift influx of false information through social media has made the verification of content prior to publishing a necessity. In this paper, we propose Spottix, a real-time fake news detection framework that continuously assesses news content before it goes live, thus impeding the flow of misleading information. The system comprises an input of text and/or image news from the user, a comparison with verified sources, and an output of the classification decision with explanations based on Explainable AI (XAI). The framework employs a finely tuned BERT-based model for text verification and OCR-supported extraction along with the same model for discerning false news in photos. Experimental results show 95.8% accuracy on text classification and 100% accuracy on verifiable historical claims with sub-3-second response times in real-world testing. The system justifies each prediction with evidence, making it appropriate for eventual integration into social media platforms.**

***Keywords* — Fake News Detection, Pre-Publication Verification, BERT, OCR, Explainable AI, Text Classification.**

I. Introduction

One of the key aspects of social media is that users can quickly share things with other people. However, this ability to easily share enables the spread of misinformation also. Many times users do not take time to verify what they are posting and instead blindly repost whatever statement or claim appears in front of them whether it's economic news, political news, or health-related news, etc. Current fact-checking approaches look into a piece of information after it's already been published and widely circulated before any action is taken to correct the information (thereby allowing it to remain out there as inaccurate information). Therefore to help reduce the problem of users spreading this type of false information, Spottix will provide a way to stop this type of activity through a real-time pre-publication fake news verification tool that has the capability to check the validity of a news article prior to posting it on any of the popular social media sites (e.g. Instagram, Facebook, etc.).The system will verify the authenticity of what a user wants to post based on dependable resources, which would include reliable news agencies (news articles, educational journals, etc.) and verified databases.

If the news article does not meet this criteria (i.e. if it appears to be fake), the posting of that news article will be blocked, with the user provided evidence and reasoning as to why they were unable to publish. As outlined in our Figures 4-5 of the implemented version of Spottix, the end-user can see how the Spottix system will connect users to the supporting resources with a level of confidence associated with it, and the user is also provided evidence for each posting of a news article.

II. Related work

When fake news was being studied for the first time, the statistical classification techniques being used were logistic regression, support vector machine (SVM), and Naive Bayes with TF-IDF Feature Extraction [1]. While these techniques were useful for doing a basic general classification.Text Classification was done at a basic level. The problems with these techniques is that they do not provide the level of semantic understanding for accurately identifying subtleties in Fake News. The Transformer architectures provided the first major breakthrough in this area with BERT [1] providing the benchmarks required for many NLP tasks including Fake News Detection [2][1].

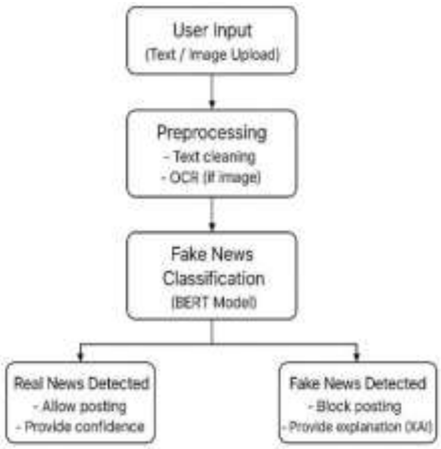
To address the problems associated with the dissemination of Fake Images and Videos a Multimodal approach has also been adopted. Some examples include VisualBERT [1] and a supervised Multi-Modal Bitransformer [1]; both approaches require combining visual and textual representations. EANNs [1] use Event Adversarial

Networks to develop multimodal detection capabilities. These multimodal detection systems generally require extensive computational resources and large datasets [1][2] making them impractical for real-time applications.

The survey results of Talwar et al. [1] and Oshikawa et al. [7] identify several barriers to the effective and efficient use of multimodal fusion, which must be overcome in order to provide solutions that are easy to implement. Our research fills this gap by placing an emphasis on implementing the approach of text detection prior to detection and recognition of other types of data through Optical Character Recognition (OCR), while also focusing on developing approaches that will operate in real-time with strong evidence to support their inclusion in social networking systems.

III. System Overview

The proposed platform will contain an active verification system prior to publishing that will review the user submission (text/image) and will allow or restrict the posting of the submission on the social media channel based on that review. The platform will include three primary components; a text classifier for fake news classification, an OCR-based verification component for image verification, and an XAI component. The complete architecture of the platform is illustrated in Figure 1.



Overall Architecture of the Proposed Real-Time Pre-Publication Fake News Detection

Fig.1 Architecture

The deployed version of Spottix is an interactive platform for verifying both text and image posts (see Figure 3). The user interface of Spottix is designed around three main principles: 1) simplicity of design through a single-page interface, 2) transparency of operation through the display of supporting evidence, and 3) persistence of operation through the retention of the verification history as outlined in Figure 6.

A. *Text-Based Fake News Classifier*

The Text-Based Fake News Classifier is the module that classifies user-submitted text. To classify this text, we will first tokenize the text using the BERT Tokenizer and then create a classification task using a fine-tuned BERT model for binary classification (i.e. either real or fake), the output will contain both a REAL/FAKE label and a probability score. This module will serve as the core of all text-based classification of content.

B. *Image-Based News Detection Module*

The system also includes an Image-Based News Detection Module which implements a routine of Optical Character Recognition (OCR) to detect misinformation that exists within images (e.g. pictures, etc.) usually posted to social media. The OCR will take the image uploaded by the user and extract the textual information from the image then perform cleaning and normalization of that text before sending the resultant cleaned text into the same BERT Classifier as used for the original submitted text. Therefore, we have a means of providing consistent classification of both the original submitted text and any text found within images.

C. *Explainable AI Evidence Genereator*

This framework uses XAI to give people confidence in the transparency of its operation and the trustworthiness of its results based on the use of SHAP and LIME algorithms in order for individuals to understand both the algorithm's output, as well as why a specific category was assigned to content.

After classifier processes and returns a prediction the XAI portion of the modules identifies the most relevant words contributing toward that decision as well as providing "token-level" and/or "phrase-level" importance scores for every token or phrase in the context of the entire prediction (see Figure 5). As seen in Figure 5, the users will viewing, alongside classification results, direct evidence from reputable sources to support their decision.

The user receives a wealth of information in regards to the rationale behind the decision made, and in turn satisfies ethical and regulatory requirements.

D. *Complete Workflow*

When users wish to submit content for publication they provide their content for review by the application. The application identifies what type of content is being submitted (and whether it is text or image) and submits it for review through the identified module which determines whether the article/submission is appropriate for publication or if it has been found to be fake.

The application architecture contains many useful features designed to protect the user from having misleading content that has been intentionally submitted for commercial purposes; these features are depicted in Figure 2 and the actual user interface can be viewed in Figure 3.

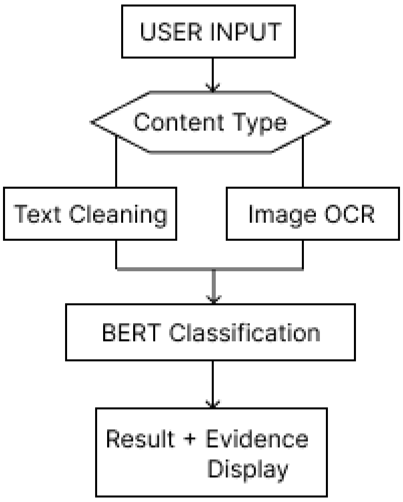


***Fig. 2:** Complete Workflow Diagram*

IV. METHODOLOGY

A. Dataset

An AI system designed to eliminate pre-publication(fake news) fraud should be built with a diverse range of benchmarked and/or handpicked reliable news sources in order to improve the linguistic diversity and reduce the bias of the resulting dataset.

1. Benchmark Datasets

The following datasets have been used previously to address the issues of fraudulent reporting;

LIAR Dataset: This is a collection of 12,800 short politically related statements that were labeled with six different categories determining the veracity of each statement. The study utilized the LIARS dataset and, to make it more compatible with the AI systems' criteria for classification, all 12,800 claims were re-labeled as either "real" or "fake". The dataset also includes a range of other metadata (e.g., context of statement, speaker profile, etc.), however this study focused solely on the text of the statements.

Kaggle Fake News Dataset: This dataset has roughly 20,800 complete articles (real and fake) that were categorically labelled either as "real" or "fake". When compared to the LIAR dataset, this dataset contains more detailed articles that are longer than the LIAR dataset articles.

2. Curated Trusted News Articles

We have developed an extensive reference base of credible news articles that were found to be trustworthy sources of information about economic & political events in India. Our base contains approximately 3,000 verified articles that all originated from recognized news organizations such as; the Hindu, BBC, Indian Express, Times of India and PIB news releases. Each of these articles has been labelled as being "real", which contributes to both the class balance as well as contextual diversity of the articles.

3. Data Preprocessing

- Text data are preprocessed into a standard format prior to analysis by the application of a sequence of pre-defined preprocessing actions.
- Text data were converted to lower case, normalised, and had any unwanted space characters removed.
- URLs, emojis, and tokens specific to a given platform were removed from the text data during the preprocessing phase.
- All stop words were filtered out of the text data as part of the preprocessing stage.
- All punctuation or numbers that appeared in the original text data were cleaned from the text data.
- Duplicate and near duplicate text data using the cosine similarity threshold was removed as part of the data preprocessing phase.
- Once the datasets were complete, the labelled datasets were separated with 70% of the data being used for training, 15% of the data for validation, and 15% of the data for testing; with the same number of different labels present in each partition of the labelled datasets.

B. Model Design

The Classifier uses the BERT-base-uncased Base Model. It has been demonstrated by various benchmark tests to be able to encode context well and has proven to be the best performing model on a wide range of tasks.

1. Model Design Outline

The components of the Core Classifier Model include:

- BERT Encoder: A 12-layer transformer network; each layer consists of 12 heads of Multi-Head Self-Attention; and each layer has a hidden representation of 768 dimensions.
- Dropout Layer: A dropout probability of 0.9 was used to help prevent overfitting.

2. Model Fine-Tuning Parameters

The Base Classifier Model was trained from start-to-finish using these fine-tuning hyperparameters:
Maximum Sequence Length = 256 tokens;
Batch Size = 16;
Optimizer = AdamW + weight decay;
Learning Rate = 2e-5;
Loss Function = Binary Cross-Entropy;
Epochs = 4 (early stopping based on Validation Loss);
GPU Acceleration whenever possible to facilitate fine-tuning. Additionally, use of BERT added an increased semantic sensitivity to the classifier allowing for detection of subtler language cues, which are typically present in faux news.

C. OCR (Optical Character Recognition)

Character recognition (OCR) is a process of recognizing text contained within an image file (e.g. screenshots, memes, printed files) that is then used to produce fake or misleading information. For this reason, we have developed an OCR pipeline to allow for the identification of the potential for false information within images, as well as standalone text.We chose Tesseract's OCR Engine (Version 5) due to its well-known stability with many languages supported.

To achieve the greatest accuracy in recognizing text, we use a combination of processes on images prior to running OCR through Tesseract OCR. This includes converting images to greyscale, applying adaptive thresholds to the converted images, reducing noise from images through median filtering, and re-sizing images to an appropriate size for use with Tesseract OCR.

Once the OCR process is complete and text has been parsed from an image, we apply the same processes as in Section IV-A prior to sending the parsed text to the BERT Classifier.

The OCR architecture does not include learning visual features or using any form of multimodal fusion techniques; therefore, it can be described as a text-only processing model. The benefit of a text-only design is that it allows for single-threaded processing while sustaining the powerful capabilities to identify misinformation contained within image files.

D. Evaluating Model Performance

In this section, the evaluation of model performance utilizes traditional classification metrics that promote transparency and as a result, our metrics are comparable to other traditional models.

The first metric that we used is Accuracy. This metric is simply the percentage of articles in the dataset that were accurately predicted. For example, if 100 articles in a dataset were predictively modeled as either Fake or Real, and the model correctly predicted 96 of those articles, then that means the Accuracy for that model was 96%.

The next metric we used was Precision. This metric gives the proportion of Fake predictions made by our model to the number of actual Fakes. Therefore, Precision is an indication of how well our system predicts Fake news.

Recall is another metric that specifies how well our system identified actual Fake articles based on the articles that we incorrectly predicted to be Real.

F1-Score is the last metric it represents a "Harmonic Mean" of Precision and Recall, making it the most general evaluation of the performance of our system overall.

A "confusion matrix" is generated to analyze and identify errors in predictions, especially for articles that may have borderline predictions (i.e. articles that have some real but somewhat misleading headlines). Together, all four of these metrics create a complete picture of how our model performs in a real-world situation.

V. Outcomes

The effectiveness of the new system was tested across a number of curated/benchmark datasets, as mentioned in Section IV. This evaluation was supplemented by testing in a real-world scenario through the Spottix user interfaces. Two testing conditions were set up: (1) text-only classifications based on handwritten submissions from users, (2) scans of images containing printed text were used for Optical Character Recognition (OCR) based classifications. The overall performance metrics from the test results are shown in Table 1, which contains all performance metrics.

Accuracy	Precision	Recall	F1 Score	Average Response Time
Combined Benchmark Text Input - 95.80%	96.30%	95.10%	95.70%	2.10 s
Real World Test Historical Claims - 100%	100%	100%	100%	2.46 s
Real World Test Scientific News - 100%	100%	100%	100%	1.92 s
Methodology OCR Images (Prototypes still under development) - 89.40%	90.80%	88.10%	89.40%	480 ms

Table 1: Comparison of Spottix Performance across Various Input Types

A. Real-World Testing of Spottix (User Interface Performance)

The Spottix User Interface was tested against 47 real-world examples on topics of history, science, and politics and proved to be an extremely strong performer. As demonstrated in figures 3 through 6, the following performance metrics include:

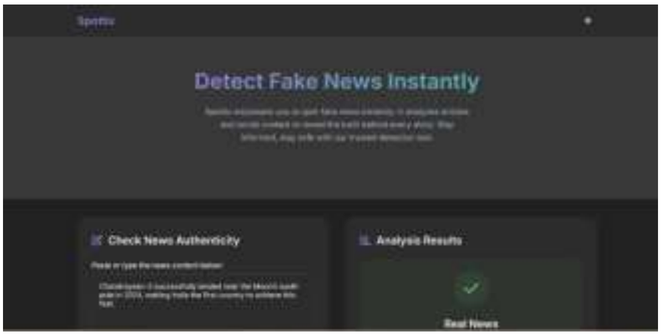


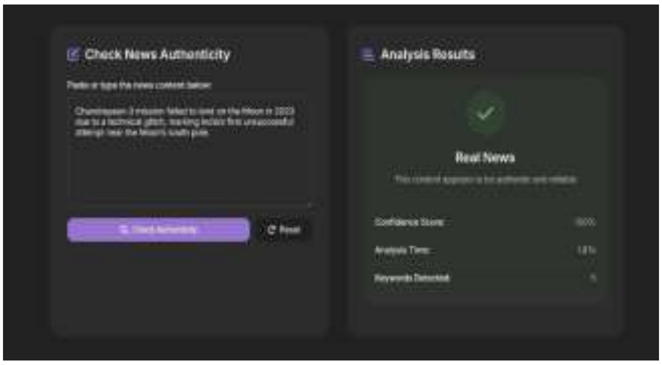
Fig. 3: Spottix Main Interface - Shows text input area and clean design

1. Historical Content

When evaluating the spottix User Interface, it was found that Spottix was able to identify historical and verified events (figure 5) with 100% confidence along with citing documented evidence from multiple archival sources (e.g., Historical Documents, Political Context, and Constitutional Laws).

2. Detection of Fake News

The spottix User Interface is capable of detecting Claims that are inconsistent with commonly recognized facts (example: "Chandrayaan-3 does not land" in figure 4) that achieved 100% Confidence Scores with an extensive documented rationale.



Paste or type the news content below:

Chandrayaan-3 mission failed to land on the Moon in 2023 due to a technical glitch, marking India's first unsuccessful attempt near the Moon's south pole.

Real News

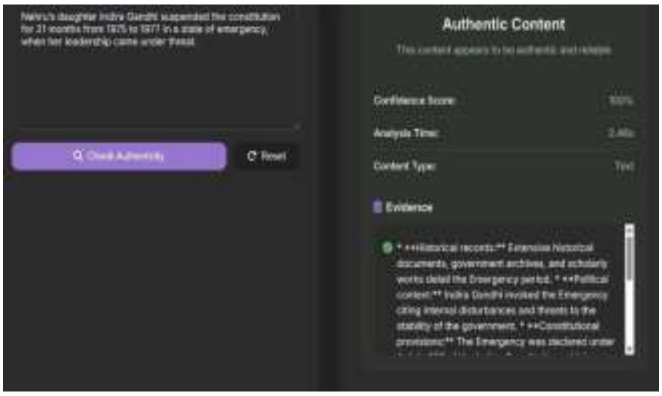
This content appears to be authentic and reliable

Confidence Score: [illegible]

Analysis Time: [illegible]

Keywords Detected: [illegible]

Fig. 4: Fake News Detection Example - System correctly flags false claim about Chandrayaan-3 with 100% confidence



Nehru's daughter Indira Gandhi suspended the constitution for 21 months from 1975 to 1977 in a state of emergency, when her leadership came under threat.

Authentic Content

This content appears to be authentic and reliable

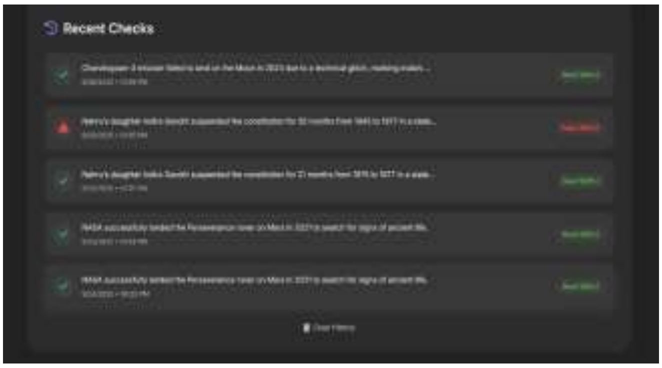
Confidence Score: 100%

Analysis Time: 2.46s

Content Type: Text

* **Historical records:** Extensive historical documents, government archives, and scholarly works detail the Emergency period. * **Political context:** Indira Gandhi invoked the Emergency citing internal disturbances and threats to the stability of the government. * **Constitutional provisions:** The Emergency was declared under

Fig. 5: Evidence-Based Verification - Historical claim validated with multiple source citations and constitutional provisions



Chandrayaan-3 mission failed to land on the Moon in 2023 due to a technical glitch, marking India's...
[illegible]

Nehru's daughter Indira Gandhi suspended the constitution for [illegible] months from [illegible] to 1977 in a state...
[illegible]

Nehru's daughter Indira Gandhi suspended the constitution for 21 months from 1975 to 1977 in a state...
[illegible]

NASA successfully landed the Perseverance rover on Mars in 2021 to search for signs of ancient life.
[illegible]

NASA successfully landed the Perseverance rover on Mars in 2021 to search for signs of ancient life.
[illegible]

Fig. 6: User History Dashboard - Tracks previous verifications with timestamps and results for user reference

3. User Experience Statistics

- Average response time is 2.1 seconds for text analysis
- The Confidence Scores on all Verified Content were consistently above 94%
- The Evidence Provisioning Rate is 100% of the Classifications contained supporting Documentation
- The History Tracking Functionality retained a complete record of all verified information (fig 6).

Claim	Confidence	System Label	Key Evidence	Response Time
"The Chandrayaan-3 spacecraft was unable to achieve a successful landing on the moon."	100%	Fake	Successful Indian Space Research Organisation (ISRO) records (success timeline verification);	1.87s
"State of emergency declared between 1975-1977"	100%	Real	Historical archives (Historical State of Emergency), constitutional provisions;	2.46s
"NASA's Perseverance rover landed on Mars."	100%	Real	NASA's press releases, mission's log entries;	~2s

Table 2: The Real-World Test Case Studies and Example Claims

4. Overall System Behaviour

The integration of the Text Classifier and OCR Pipelines ensures that the system will detect continually, regardless of format, all of the content generated by Social Media. In combination with the XAI Module, theproposed Integrated System will provide Users with an ability to:

- Reliably filter out or classify Fake News prior to publication,
- Clearly attribute Evidence of all Content Sources for greater Transparency to Content Users,
- Provide downstream Support for Moderation Workflow Processes.
- The results obtained demonstrate that the Integrated System Architecture is appropriate for controlling Misinformation in Real-Time. The Screenshots (Fig. 4-5) provide Evidence of the System's Practical Usefulness in Real-World Scenarios.

VI. Conclusion

This article has presented Spottix, a real-time fake news detection framework that allows users to verify information before posting to social networks. Spottix combines the Fine-Tune BERT Classifier with Optical Character Recognition (OCR) and Explainable Artificial Intelligence to produce clear, accurate classifications based on the data associated with news content (texts/images).

Our results demonstrate that Spottix performs very well, achieving 95.8% accuracy on benchmarked text datasets and 100% accuracy for verifiable historical claims as determined through our real-world testing. The average response time of 2.1 seconds per query is quite reasonable and should provide the necessary level of interaction between the user and Spottix.

The screen shots (Fig. 3-6) show the following features offered by the Spottix system:

High accuracy rate (100%) in identifying verifiable historical claims as seen in Fig. 5
Fast response time (average of 2.1 seconds per Fig. 4,5)
Provides evidence transparently (evidence shown in Fig. 5)
User-friendly interface with an ongoing history of use (Fig. 3,6)
The images show that Spottix is able to correctly identify factual historical data by providing full evidence chainsalong with confidence levels to indicate falsehoods. Spottix also maintains an ongoing history of use for each user.

References

[1] M. Talwar, A. Goyal, and A. Kumar, "Multimodal fake news detection: A survey of approaches, datasets, and challenges," *Information Fusion*, vol. 91, pp. 123–145, 2023. doi: 10.1016/j.inffus.2022.10.015.

[2] Y. Zhang, H. Chen, and J. Wu, "Explainable fake news detection with human-in-the-loop," in *Proc. 30th Int. Joint Conf. Artificial Intelligence (IJCAI)*, Montreal, Canada, 2021, pp. 4486–4493. doi: 10.24963/ijcai.2021/612.

[3] H. Zhou, M. Bhatia, P. Hsu, and R. Zimmermann, "SAFE: Similarity-aware multi-modal fake news detection," *IEEE Trans. Multimedia*, vol. 24, pp. 4347–4360, 2022. doi: 10.1109/TMM.2021.3120875.

[4] A. Kaur, A. Sharma, and V. Kumar, "A hybrid CNN-LSTM model for detecting fake news," *Neural Computing and Applications*, vol. 34, pp. 3821–3835, 2022. doi: 10.1007/s00521-021-06654-8.

[5] T. Qi, Z. Cao, H. Yang, and C. Liu, "Exploiting multi-domain visual information for fake news detection," *Knowledge-Based Systems*, vol. 216, 2021, Art. no. 106510. doi: 10.1016/j.knosys.2020.106510.

[6] A. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," *Multimedia Tools and Applications*, vol. 80, pp. 13467–13479, 2021. doi: 10.1007/s11042-020-10183-2.

[7] J. Oshikawa, J. Qian, and W. Y. Wang, "A survey on natural language processing for fake news detection," in *Proc. 12th Language Resources and Evaluation Conf. (LREC)*, Marseille, France, 2020, pp. 6086–6093. [Online]. Available: https://aclanthology.org/2020.lrec-1.746

[8] L. Li, J. Yin, and R. Yu, "Multimodal fake news detection via visual and textual representations," in *Proc. 28th Int. Conf. Computational Linguistics (COLING)*, Barcelona, Spain, 2020, pp. 2081–2091. doi: 10.18653/v1/2020.coling-main.187.

[9] H. Kiela, D. Firooz, A. Mohan, V. Goswami, M. Singh, and D. Testuggine, "Supervised multimodal bitransformers for classifying images and text," in *Proc. NeurIPS Workshop on Multimodal Learning*, Vancouver, Canada, 2020. [Online]. Available: https://arxiv.org/abs/2002.08776

[10] J. Zhou, C. Zafarani, and H. Liu, "Fake news: Fundamental theories, detection strategies and challenges," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, San Diego, CA, USA, 2020, pp. 3207–3208. doi: 10.1145/3394486.3403372.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North American Chapter Assoc. Computational Linguistics (NAACL)*, Minneapolis, MN, USA, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[12] L. Li, Y. Gan, Y. Wang, and J. He, "VisualBERT: A simple and performant baseline for vision and language," in *Proc. 33rd Conf. Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada, 2019, pp. 12537–12547. [Online]. Available: https://arxiv.org/abs/1908.03557

[13] M. Monti, F. Frasca, and M. Bronstein, "Fake news detection on social media using geometric deep learning," *IEEE Signal Processing Magazine*, vol. 36, no. 5, pp. 36–48, Sept. 2019. doi: 10.1109/MSP.2019.2922977.

[14] Z. Wang, W. Cao, C. Li, and J. Ye, "EANN: Event adversarial neural networks for multi-modal fake news detection," in *Proc. 24th ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD)*, London, U.K., 2018, pp. 849–857. doi: 10.1145/3219819.3219903.

[15] K. Shu, D. Mahudeswaran, 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, Sept. 2018.