DETECTING MISINFORMATION USING BERT: A LANGUAGE MODEL APPROACH TO ANALYZING FAKE NEWS NARRATIVES

Undergraduate Thesis
Submitted to the Faculty of the
Department of Computer Studies
Cavite State University – Imus Campus
City of Imus, Cavite

In partial fulfillment
of the requirements for the degree
Bachelor of Science in Computer Science

VON PHILIPPE C. ACERO
LOUISE MARK A. BANDOJA
CZAR JOHN V. VILLAREAL
January 2025

DETECTING MISINFORMATION USING BERT: A LANGUAGE MODEL APPROACH TO ANALYZING FAKE NEWS NARRATIVES

INTRODUCTION

Background of the Study

In the modern era, news and media platforms have become a primary source of information for individuals worldwide and for a long time have been viewed as a reliable source for understanding current trends and happenings within a community. The rise of these online platforms has allowed the rapid and widespread sharing of information, often without being validated, which allows fake news to spread. As there aren't gatekeepers that validate the information, fake news on the internet spreads faster and wider compared to factual information(Wu et al., 2022). As of 2022, approximately 75% of the Filipinos are using the Internet, potentially exposing them to fake news and topics on the internet.

As we keep seeing misinformation to be one great challenge, the importance of advanced technologies to detect fake news has increased great deal. Among the most promising techniques in recent years has been the BERT (Bidirectional Encoder Representations from Transformers), an advanced language model designed to understand the complexities of natural languages. BERT's enormous pre-training on a huge amount of text data lets it configure at the level of words and sentences, a makes it a very productive language model for applications like fake news detection. Fine-tuning it on domain-specific datasets, news articles in Tagalog, for example, allows such knowledge of how to detect slightly deceptive linguistic cues, deceptive narratives, and inconsistency in conclusion typical of fake news.

While BERT has shown great promise in detecting fake news, it faces particular challenges when it comes to distinguishing between genuine misinformation and satire. Satirical news, which often employs humor, exaggeration, and irony to critique societal issues, can closely resemble real news stories. Fake

news and satirical news are similar but the motivation is what makes them different(Das and Clark, 2019).

This makes it difficult for machine learning models like BERT to differentiate between the two. Unlike traditional fake news, which is typically designed to mislead or deceive readers, satirical content is intended to entertain or provoke thought through a comedic lens. However, this subtlety in tone and intent can easily confuse automated systems.

Statement of the Problem

While various machine learning models have been developed to detect fake news, their effectiveness in the specific context of Filipino media remains underexplored. Pre-trained models like BERT have shown promise in natural language processing tasks, but their performance in distinguishing between fake news and satire in the Filipino context is not well understood. Moreover, fine-tuning these models to handle context-specific challenges, such as the nuances of Tagalog and English language usage, presents an additional layer of complexity.

This study seeks to address the following problems:

1. Performance Evaluation: *How effectively can pre-trained BERT models detect fake news narratives within Filipino news media and social media?*
2. Distinguishing Satire from Fake News: *What are the primary challenges in differentiating satirical content from genuine misinformation in the context of Filipino discourse, and how can these challenges be addressed?*
3. Model Fine-Tuning: *How can BERT models be fine-tuned to improve their accuracy and reliability in context-specific fake news detection within Filipino media?*

By addressing these questions, this research aims to contribute to the development of more robust and culturally sensitive tools for misinformation detection in the Filipino media landscape.

Objectives of the Study

The primary objective of this study is to explore how pre-trained BERT models can be fine-tuned to detect fake news narratives in Filipino news media and social media, while also addressing the challenge of distinguishing between satirical content and genuine misinformation in the context of Filipino discourse.

Specifically, the study aims to:

1. Evaluate pre-trained BERT models performance in fake news detection.
2. Explore challenges in differentiating fake news from satire in Filipino News and Social Media
3. Fine-Tune BERT for Context-Specific Fake News Detection in Filipino Media

Time and Place of the Study

The study “DETECTING MISINFORMATION USING BERT: A LANGUAGE MODEL APPROACH TO ANALYZING FAKE NEWS NARRATIVES” will be conducted at Cavite State University – Imus Campus from December 2024 to January 2025.

Scope and Limitations of the Study

The research focuses on fine-tuning pre-trained BERT models to detect fake news narratives within Filipino news media and social media. It will primarily target content in Tagalog and English, using pre-trained BERT models designed for these languages. If a Tagalog-specific model is unavailable, a broader Filipino-language model will be used. The study will analyze news articles from mainstream Filipino media and user-generated content from social media platforms like Facebook and X(formerly known as Twitter). It will differentiate between satire, fake news, and real news by identifying narrative and linguistic patterns. The research will adapt BERT models to handle linguistic nuances, such as idiomatic expressions and code-switching, and will evaluate model performance using metrics like precision,

recall, F1 score, and accuracy. Additionally, a comparative analysis of Tagalog-specific and broader Filipino-language models will be conducted.

The research will be limited to content in Tagalog and English, excluding other Filipino languages and dialects due to time and resource constraints. The study's success will depend on the availability and quality of pre-trained BERT models and labeled datasets for fake news, satire, and real news in the target languages. The diversity of social media content may be constrained by platform-specific and demographic biases, potentially affecting the generalizability of findings. Furthermore, the model may encounter challenges in understanding deeply contextual or culturally specific references common in Filipino satire and fake news. The scope of experimentation will also be limited by computational resources and the short timeframe, focusing on a specific aspect of fake news detection rather than a comprehensive analysis.

Definition of Terms

- BERT or Bidirectional Encoder Representation from Transformers: transformer-based machine learning model developed by Google, designed to understand the context of words in a sentence.
- Fake News: false or misleading information presented as news, often created to deceive the public or manipulate opinions.
- Satire: genre of literature and media that uses humor, irony, exaggeration, or ridicule to criticize or highlight societal issues, often through fictitious or exaggerated narratives that are not meant to be taken literally.
- Real News: information and reports that are fact-based, verified, and produced by credible news organizations following journalistic standards.
- Framing Theory: theory that explores how the way information is presented (framed) influences audiences' perception and interpretation.

- Pre-trained Model: A machine learning model that has been trained on a large dataset before being fine-tuned for a specific task.
- Fine-Tuning: The process of adapting a pre-trained model to a specific task by training it further on a smaller, task-specific dataset.
- Social Media: Online platforms where users create, share, and engage with content.
- News Media: Organizations and platforms that produce and distribute news content, including newspapers, television channels, and online news websites.

Theoretical Framework of the Study

This study is grounded in Framing Theory, which serves as the primary lens for analyzing the various forms of information—fake news, satire, and real news—within the context of Filipino news media and social media. Framing theory, as proposed by Erving Goffman (1974) and later developed by Robert Entman (1993), posits that media outlets and communicators "frame" news stories in particular ways to influence the perception of an issue, event, or narrative. This framing process involves selecting certain aspects of a story to highlight, while downplaying others, thereby shaping how the audience interprets and understands the message.

Framing Theory will be employed to analyze the different frames present in fake news, satire, and real news. By identifying framing cues—such as exaggerated claims, emotional appeals, and humorous exaggerations—the study will distinguish between fake news and satire. Framing cues will also be important in identifying how real news frames stories to ensure they are understood as factual and balanced. This approach will involve the use of pre-trained BERT models, which will be fine-tuned to recognize these framing strategies and identify misinformation, satire, and news content in both Tagalog and English.

In conjunction with Framing Theory, Deception Detection Theory will be incorporated into the study to focus on identifying deceptive content. Deception Detection Theory, which looks at linguistic cues and psychological indicators of dishonesty, will provide a framework to detect fake news through textual patterns that often signal deception. These indicators include:

- Contradictory statements: Deceptive content may include inconsistencies in the narrative or factual inaccuracies.
- Hedging language: Phrases like "perhaps," "maybe," or "some say" suggest uncertainty or evasion.
- Overly emotional or manipulative language: Deceptive content often uses extreme emotional appeals to sway the reader's opinion.
- Exaggerated claims: Deceptive narratives often exaggerate details to appear more convincing or sensational.
- Irrelevant details: Fake news and deceptive stories might introduce irrelevant or tangential information to divert attention from the core issues.

This theory will guide the identification and classification of fake news through linguistic markers associated with deception. By training BERT on these deception cues, the model will be able to differentiate between news that intentionally misleads and content that is humorous or satirical but not intended to deceive.

The study will also focus on understanding how these theories apply to Filipino media, considering cultural nuances and linguistic features specific to the Filipino context. A bilingual corpus (Tagalog and English) will be developed, containing news content from mainstream outlets, independent media, and social media platforms. This will allow the BERT model to identify linguistic and cultural frames specific to Filipino media, such as the use of idiomatic expressions or rhetorical devices that may signal specific frames related to Filipino identity, politics, or social issues.

Filipino satire presents a unique challenge for the detection model, as satire often uses exaggerated claims and sarcasm that could resemble deceptive or sensational content. Therefore, special attention will be paid to the linguistic and cultural markers that distinguish Filipino satire from fake news and real news. For example, Filipino satire often involves playful sarcasm, hyperbole, and irony, which are not intended to deceive, but rather to entertain or critique. The study will focus on differentiating these elements from deceptive narratives that intentionally mislead the audience.

In addition, the study will evaluate how well the model handles the complexities of bilingual content and how framing techniques might differ between Tagalog and English. The model will be fine-tuned to recognize framing strategies in both languages, ensuring its effectiveness across diverse types of Filipino media content.

To assess the performance of the fine-tuned BERT model, standard NLP metrics—such as accuracy, precision, recall, and F1-score—will be employed. The study will also explore the limitations of the model, particularly when it comes to handling cultural ambiguities or subtle language nuances that may not be fully captured by pre-trained models alone.

By integrating Framing Theory with Deception Detection Theory, this study aims to provide a more comprehensive approach to detecting and understanding misinformation in Filipino media. The combined theoretical framework will enable the model to recognize not only the framing techniques that influence how news is perceived but also the linguistic cues that indicate deception. Ultimately, this research will contribute to the field of automated news analysis and help address the challenges posed by misinformation, satire, and the blurred lines between them in the digital age.

REVIEW OF RELATED LITERATURE/STUDIES

This chapter discusses some of the recent technologies and discovered techniques in dealing with fake news. The focus is on the advancements in natural language processing (NLP), machine learning, and artificial intelligence (AI) that have been used or trained for detecting misinformation and/or analyzing disinformation in online narratives.

Fake News and Traditional Methods of Fake News Detection

Fake news has been around for a long time, even during the ancient history where, fake news has been used for several purposes. Balachandran, Roberts and Leal-Schuman (2023) argue that fake news has intended purpose of swaying its readers in order to change their perspectives or beliefs. Fake news is inherently persuasive, Siar (2021), stated several reasons on why people believe in fake news based on cognitive psychology and behavioral research. There are several factors, one of the cognitive factors, specifically confirmation bias, when people have a belief tends to support that belief while ignoring or dismissing the facts don't support that belief (Beauvais, 2022).

Traditionally, there are already existing several approaches when it comes to identifying fake news, before machine learning, people can do manual fact-checking with the goal being, to provide an accurate and unbiased analysis to correct misperceptions; verify the factual accuracy. There are three fake news detection model approaches: knowledge-based, modality-based, and features-based, these are according to Hamed, Aziz, and Yakub (2023), these three traditional methods as discovered in their review, have significant limitations specifically, scalability and speed which are crucial factors as fake news spreads rapidly on social media. This implies that there's a need for advanced techniques, and methods that can handle large datasets, and complex patterns.

Approaches in Fake News Detection: Machine Learning

To address the need for advanced techniques and methods, the usage of machine learning has become relevant. There have been various machine learning algorithms in tackling fake news detection with different strengths and weaknesses.

A method which combines multiple algorithms to improve classification accuracy, Ahmad et al. (2020) highlights that ensemble methods have demonstrated the ability in detecting fake news across different social media platforms, the adaptability of the model in various datasets and contexts. The comparison between machine learning algorithms such as Naïve Bayes and Decision Trees, as discussed by Dinesh and Rajendran (2021), discovered that decision trees algorithm is better than Naïve Bayes in terms of consistency and accuracy when it comes to political news classification.

The addition of Natural Language Processing (NLP) techniques in Machine Learning enhanced its capabilities. The discovery of the application of artificial neural networks in classifying fake news in Spanish has emphasized the importance of language-specific models detecting misinformation (Moreno-Vallejo et al., 2023). One of the significant findings of Guo, Schlichtkrull, and Vlachos (2022), is the identification of various machine learning models that have been developed for fact-checking tasks. The models used deep learning architectures, such as transformers, to understand and process natural language more effectively. The application of these advanced models allows better contextual understanding and semantic analysis of claims, which is crucial for accurate verification (Guo et al., 2022).

A Deeper Delve into Fake News Detection: Deep Learning

The recent advancements in Deep Learning (DL) techniques have shown promise in enhancing the accuracy and efficiency of fake news detection systems. Al-Tai, Nema, and Al-Sherbaz (2023), stated that Deep Learning introduces multilayer learning models by integrating graphs with suitable neural transformations. Convolutional Neural Networks (CNN) have also been used in fake news detection. CNN Model with margin loss which outperformed a traditional machine learning methods in classifying news articles (Goldani et al., 2021). This claim is further supported by the review of Al-Tai et al. (2023), a comprehensive review of deep learning strategies emphasizing the superiority of CNNs and Recurrent Neural Networks (RNNs) in detecting fake news over conventional techniques. RNNs, particularly Long Short-Term Memory (LSTM) networks, capture temporal dependencies, making them effective at understanding how news stories evolve and spread.

Hybrid models that combine deep learning techniques employed by Dong, Victor, and Qian(2020) study employed CNN with a two-path semi-supervised learning framework that uses both content and propagation data to improve data accuracy. Upadhayay and Behzadan (2022) in their study, proposing a hybrid approach that combines the analysis of news propagation patterns on social media with textual features, had suggested that the dissemination characteristics of fake news differ from genuine news. These models offer a more comprehensive approach to detecting fake news by leveraging both the content of the news and the patterns of its spread across social media platforms. Despite the promise of deep learning models for fake news detection, challenges like data imbalance, limited labeled datasets, and high computational costs persist.

Bidirectional Fake News Detection: BERT

BERT (Bidirectional Encoder Representations from Transformers) helps address challenges like data scarcity by using pre-trained models. Recent studies have shown that a fine-tuned BERT on specific datasets have shown significant improvement in metrics such as accuracy and F1 scores compared to older methods like Word2Vec and FastText (Kim et al., 2022). Combining BERT with other deep learning architectures has shown results, a study reported that combining BERT with LSTM (Long Short-Term Memory) networks yielded superior accuracy in detecting misinformation compared to using BERT alone (Taha et al., 2023). These advancements highlight BERT's effectiveness in fake news detection, with fine-tuning and hybrid models demonstrating improved performance.

The increasing prevalence of misinformation has underscored the need for effective fake news detection systems, particularly for languages with unique linguistic structures such as Tagalog. In response to this challenge Jiang et al. (2021) proposed a framework that leverages multi-source data to create robust pre-trained language models specifically for Tagalog, which can be fine-tuned for various NLP tasks, including fake news detection. This approach underscores the importance of language-specific models in accurately capturing the nuances of the Tagalog language, which is crucial for effective misinformation identification. A related study explored the localization of fake news detection through multitask transfer learning, emphasizing the adaptability of BERT-based models in handling the intricacies of the Tagalog language(Cruz et al.,2020). Their research demonstrates that by employing multitask learning strategies, the performance of fake news detection systems can be significantly improved, allowing for better generalization across different datasets and contexts.

REFERENCES

Ahmad, I., Yousaf, M., Yousaf, S., & Ahmad, M. (2020). Fake news detection using machine learning ensemble methods. Complexity, 2020, 1-11. https://doi.org/10.1155/2020/8885861

Al-Tai, M., Nema, B., & Al-Sherbaz, A. (2023). Deep learning for fake news detection: literature review. Al-Mustansiriyah Journal of Science, 34(2), 70-81. https://doi.org/10.23851/mjs.v34i2.1292

Balachandran, S., Roberts, M., & Leal-Schuman, E. (2023). “Alternative Facts”: The Origins of Fake News and its Implications. *Journal of Student Research*, *12*(3). https://doi.org/10.47611/jsrhs.v12i3.5966

Beauvais C. Fake news: Why do we believe it? Joint Bone Spine. 2022 Jul;89(4):105371. doi: 10.1016/j.jbspin.2022.105371. Epub 2022 Mar 4. PMID: 35257865; PMCID: PMC9548403.

Bitesize, B. (2024, November 19). *A brief history of fake news - BBC Bitesize*. BBC Bitesize. https://www.bbc.co.uk/bitesize/articles/zwcgn9q

Cruz, J.C.B, Tan, J.A, and Cheng, C. 2020. Localization of Fake News Detection via Multitask Transfer Learning. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 2596–2604, Marseille, France. European Language Resources Association.

Das, D., & Clark, A. J. (2019). *Satire vs Fake News: You Can Tell by the Way They Say It. 2019 First International Conference on Transdisciplinary AI (TransAI).* doi:10.1109/transai46475.2019.00012

Dinesh, T. and Rajendran, D. (2021). Higher classification of fake political news using decision tree algorithm over naive bayes algorithm. Revista Gestão Inovação E Tecnologias, 11(2), 1084-1096. https://doi.org/10.47059/revistageintec.v11i2.1738

Dong, X., Victor, U., & Qian, L. (2020). Two-path deep semisupervised learning for timely fake news detection. Ieee Transactions on Computational Social Systems, 7(6), 1386-1398. https://doi.org/10.1109/tcss.2020.3027639

Goldani, M. H., Safabakhsh, R., & Momtazi, S. (2021). *Convolutional neural network with margin loss for fake news detection. Information Processing & Management, 58(1), 102418.* doi:10.1016/j.ipm.2020.102418

Guo, Z., Schlichtkrull, M., and Vlachos, A. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Hamed, S., Aziz, M.J. & Yaakub, M.R. (2023). A review of fake news detection approaches: A critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion. Heliyon. 9. e20382. 10.1016/j.heliyon.2023.e20382.

Kim, M., Kim, M., Kim, J., & Kim, K. (2022). Fine-tuning bert models to classify misinformation on garlic and covid-19 on twitter. International Journal of

Environmental Research and Public Health, 19(9), 5126. https://doi.org/10.3390/ijerph19095126

Merriam-Webster. (n.d.). Fact-check. In Merriam-Webster.com dictionary. Retrieved January 6, 2025, from https://www.merriam-webster.com/dictionary/fact-check

Moreno-Vallejo, P., Bastidas-Guacho, G., Moreno-Costales, P., & Chariguaman-Cuji, J. (2023). Fake news classification web service for spanish news by using artificial neural networks. International Journal of Advanced Computer Science and Applications, 14(3). https://doi.org/10.14569/ijacsa.2023.0140334

Siar, S. V. (2021, August). Fake news, its dangers, and how we can fight it. *Philippine Institute for Development Studies*. https://www.pids.gov.ph/publication/policy-notes/fake-news-its-dangers-and-how-we-can-fight-it

Taha, M., Zayed, H., Azer, M., & Gadallah, M. (2023). Automated covid-19 misinformation checking system using encoder representation with deep learning models. Iaes International Journal of Artificial Intelligence (Ij-Ai), 12(1), 488. https://doi.org/10.11591/ijai.v12.i1.pp488-495

Tiwari, V. (2021, December 14). BERT: The theory you need to know!! - Analytics Vidhya - Medium. *Medium*. https://medium.com/analytics-vidhya/bert-the-theory-you-need-to-know-ddd316794395

Upadhayay, B. and Behzadan, V. (2022). Hybrid deep learning model for fake news detection in social networks (student abstract). Proceedings of the Aaai Conference on Artificial Intelligence, 36(11), 13067-13068. https://doi.org/10.1609/aaai.v36i11.21670

What is fact-checking - Ballotpedia. (n.d.). Ballotpedia. https://ballotpedia.org/What_is_fact-checking

World Bank Open Data. (2022). World Bank Open Data. https://data.worldbank.org/country/philippines

Wu, Y., Ngai, E. W. T., Wu, P., & Wu, C. (2022). Fake news on the internet: A literature review, synthesis, and directions for future research. *Internet Research, 32*(5), 1662–1699. https://doi.org/10.1108/intr-05-2021-0294