Abstract

The rapid adoption of the internet in Indonesia, with over 200 million active users as of January 2022, has dramatically transformed information dissemination, particularly through social media and online platforms. These platforms, while democratizing information sharing, have also become hotbeds for the spread of misinformation and hoaxes, significantly impacting the political landscape, as seen in the Jakarta gubernatorial election from late 2016 to April 2017. Research by the Indonesian Telematics Society (MASTEL) revealed a high prevalence of hoax content, predominantly socio-political, underscoring the critical need to address this misinformation and hoaxes challenge. This research delves into the challenge of detecting hoaxes in Indonesian political news, particularly focusing on the classification of news as factual or hoax in the presence of class imbalances within datasets. The dataset exhibits a significant class imbalance with 6,947 articles identified as hoaxes and 20,945 as non-hoaxes, Utilizing the IndoBERT model, a specialized variant of the BERT framework pre-trained on the Indonesian language, the study aims to assess its effectiveness in discerning between factual and hoax news. This involves fine-tuning IndoBERT for specific text classification tasks and exploring the impact of various resampling techniques, such as Random Over Sampling and Random Under Sampling, to address class imbalances since the dataset, significantly imbalanced with 6,947 articles labeled as hoaxes and 20,945 as non-hoaxes, necessitated these approaches. The study's findings demonstrate the IndoBERT model's consistent accuracy across different resampling methods like Random Over Sampling (ROS) and Random Under Sampling (RUS), highlighting its effectiveness in handling imbalanced datasets produce the accuracy of hoax detection with the 98.2% accuracy, 97.5% Recall, 97.8% F1-score, and 97.2% Precision. This is particularly relevant for tasks like misinformation detection, where data imbalance is common. The success of IndoBERT, a language-specific BERT model, in text classification for the Indonesian language contributes to the understanding of BERT-based models in diverse linguistic contexts.

Keywords: Hoax Detection; IndoBERT; Imbalanced Data; Political News; BERT