JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

1. INTRODUCTION

Lately, there has been a ton of deception news circling in the Indonesian news media. Hoax is information or news that contains facts that are not true or uncertain [1]. The hoax that is going around the Indonesian media is very bad and makes a lot of people think the wrong things. It says that the Indonesian media has become a place where many Indonesians can get the latest news. The hoax was due to the rapid spread of fake news in the Indonesian media, which provided an opportunity to reach out to customers in the informal community who had not seen the news recently [2].

Indonesian social media such as Twitter, Instagram, and Facebook are increasingly popular among the public [3]. Especially Twitter. Nowadays, most Indonesians use Twitter to obtain various data, particularly the most recent news. False information or news, whose integrity is only sometimes apparent, can be easily identified due to easy access. Every year, the development of Twitter clients accelerates. Comparatively, the number of users is anticipated to rise by 26%, reaching an average of 192 million in the fourth quarter of 2021 [4]. Accordingly, the utilization of tweets on Twitter is expanding, which can have both positive and unfortunate results. Due to the ease with which people interact on social media, particularly Twitter, it is difficult for most people to distinguish hoaxes from non-hoaxes due to the sheer volume of hoaxes that circulate. One of the side effects of social media is this.

In the study of detecting fraud on Twitter, SVM and other techniques were utilized. SVM is a calculation based on features that do a great job of grouping text with a lot of information about the text as an element. Although the word vector is unique in that the related words in a sentence can differ, most of the output sequence messages can be extracted directly, and the message report consists of many redundant parts. Hyperplanes that can delimit locales into subsets are created via SVM estimation. A hyperplane is a shape that separates two classes and considers the distance between their closest components. [5]. The best ratio, as determined by benchmark system accuracy tests, is 90:10, with a value of 78.33 percent. Highlights influencing the phony news class incorporate the Twitter retweet and URL elements and backing highlights. Twitter fake news can be identified with the help of the SVM classification. Create knowledge, explicitly setting off and battling [6]. Additionally, LSTM, or Long Short-Term Memory, is an iterative network design that avoids issues with long-term dependency by using memory cells and gate banks. LSTM is altered to defeat the deficiencies of RNNs in that they can't foresee words, recall long-put-away data, or eliminate information that is not generally required [7]. despite the fact that CNN is a deep-learning model that computers frequently use. Nonetheless, as Kim's exploration has shown, the CNN model can likewise be utilized to arrange sentences [8]. Convolutions, or close associations, connect multiple neurons to a single neuron in the resulting layer in CNNs. Consequently, CNNs can separate spatial data from pictures. Its compositionality and immutability of location distinguish CNN [9]. LSTM-CNN in identifying COVID-19-related fraud. To find the best model bounds, we did some research. With 16 unitary layers, LSTM-CNN can achieve 79.71 percent accuracy in experiments that combine regularizers and dropouts. [4].

Connecting the impacts of running LSTM and the IndoBERT strategy utilizing Twitter-isolated datasets on the Covid for area coercion. LSTM accomplishes a typical precision of 87.54 percent considering experimental outcomes. In addition, the tests indicate that the IndoBERT model is accurate on average to 92.07 percent [7]. Therefore, the IndoBERT model outperforms the LSTM model in deception detection tasks and has been demonstrated to offer superior average accuracy results.

Word2Vec and IndoBERT perform exceptionally well in various fraud detection tests that place an emphasis on word expansion and deep learning calculations. IndoBERT can deliver superior results than other strategies. Consequently, Twitter data on public opinion regarding news will be used in research that combines Word2Vec and IndoBERT.

IndoBERT is a model that follows the BERT Base setup [10]. The number of datasets used in IndoBERT is limited because it uses more datasets than there are different strategies. In IndoBERT, fake news is detected by utilizing transfer learning from pretraining transformer models like customized native pretraining BERT, multilingual pretraining mBERT, and monolingual pretraining IndoBERT [11]. According to the findings of the research, mBERT base-cased finetuned had an accuracy of 97.93 percent; As a result, the model outperforms the competition [12]. Moreover, this last undertaking proposition, other than utilizing IndoBERT, likewise utilizes Word2Vec. Word2Vec doesn't cover vector spaces, embedding, analogies, similarity metrics, etc. first, last, or best. Yet, word2vec is basic and available [13]. Training a log-bilinear model based on a jump-gram or continuous bag-of-words (cbow) architecture, such as implemented in word2vec and fastText, is the standard method for learning word representation. [14][15]. The word2vec representation has been widely used in NLP pipelines to improve performance. Their impressive ability to transfer to new problems suggests they collect vital statistics about three training sets [16]. Then, hoax detection research using Word2Vec with IndoBERT is still rare. Therefore, this research will focus on detecting hoaxes using Word2Vec with IndoBERT by using Indonesian news data on Twitter from public opinion on news circulating in Indonesia. Besides that, this research can reduce and educate the public about hoax news circulating.

Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**