1. INTRODUCTION

In recent years, there has been a rapid increase in the use of social media such as Facebook, Twitter, Instagram, and others. Based on WeAreSocial data in 2021, one of Indonesia's most used social media platforms is Twitter, with a value of 63.6% of users [1]. On Twitter, users can freely tweet, upload photos, and share information on their accounts, including creating tweets containing hate speech.

According to Paula and Sérgio [2], hate speech is a language that attacks or demeans that incites violence or hatred against groups based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur in a variety of linguistic styles, including subtly or with humor. One of the hate speeches that has received much attention is those directed at public officials, religious leaders, and public figures [3]. The Indonesian National Police [4] reported that hate speech cases dominated pornographic content reports from April 2020 to July 2021. Namely, there were around 473 cases, including provocative cases, hate content, and hate speech. The Indonesia National Police stated that combating cybercriminals is not simple, so infrastructure and personnel are required [5]. The impact of the problem of hate speech on social media can become considerable and widespread, with harmful implications for people, communities, and society if it is not addressed appropriately. This necessitates the development of an automatic detection tool for hate speech in the Indonesian language so that law enforcement can detect the spread of hate speech. Therefore, to solve the issue of hate speech, one of the solutions is to detect hate speech on social media using deep learning, which is part of machine learning and functions to train computers about basic human instincts.

Research [6] experimented with toxic comment classification by taking two datasets, the Google Jigsaw and Twitter datasets [7]. The purpose of the research [6] was to compare deep learning and propose an ensemble for individual classifiers in the F1-score. Deep learning carried out includes using LSTM-FastText and LSTM-GloVe. From the comparison results, LSTM-GloVe obtained an F1-score of 78.1 and was ahead of LSTM-FastText by 77.8%. In comparison, the ensemble is far superior, with an F1-score of 79.3%. However, the error in the analysis of this study is that the ensemble results identified subtasks that were difficult to classify as toxic comments because there needed to be more consistent label quality. In addition, unsolved challenges occur due to needing more training data with very special or rare vocabulary.

In research [8], [9] participated in the Hate Speech Detection on Social Networks organized by VLSP Shared 2019 with the aim of detecting Vietnamese social media text according to predetermined labels. The label consists of a pre-label dataset and an unlabeled dataset for comments or social media posts. The downside of such datasets is that the language is a low resource for natural language processing. Research [8] pre-processed and built machine learning models to classify comments or posts. Using two-word embeddings as a comparison, we get the best word embeddings, which are FastText and GloVe. The models used include SVM, Logistic Regression, GRU, and Bidirectional-LSTM. When experimenting with word embedding GloVe with the baomoi.vn.model dataset.txt, accuracy, precision, memory, and F1-score levels of 93.26%, 90.74%, 50.30%, and 53.62%, respectively, were obtained in the training dataset. Similar to word embedding, FastText gets accuracy, precision, memory, and F1-scores of 95.67%, 85.61%, 67.36%, and 73.84%, respectively, on the training dataset. And combining Bi-LSTM with FastText will bring better results. So, for the dataset of VLSP Shared 2019, it gets an F1-Score of 71.43%. Deep learning methods based on the Bi-GRU-LSTM-CNN classifier with word embedding FastText as pre-training are used in research [9]. The study obtained an F1-score of 70.576%.

Research by [10] conducted a hate speech detection experiment on Twitter with a dataset of 16 thousand tweets that other studies have annotated. For word embedding, use random embedding and GloVe. As for optimization, it uses 'Adam' for CNN and LSTM and 'RMS-Prop' for FastText. Word embedding learned from deep neural network (DNN) models, combined with the Gradient Boosted Decision tree(GBDT), yielded the best accuracy value for LSTM-Random Embedding-GBDT with an F1-score of 93% compared to FastText-GloVe F1-score of 82.9% and LSTM-GloVe of 80.8%. In addition, similar words obtained using DNN learning embeddings clearly show "resentment" towards the target words, which generally are not seen in similar words obtained using GloVe.

This research uses hate speech with elements of blasphemy, provocation, and incitement. The blasphemy element was added because the Holywings company posted posters on social media with free liquor for the end of the day named Muhammad and Maria [11]. In addition to Holywings, Netflix aired the film The Umbrella Academy 3, in which Lafaz Allah appears to be written on the floor, and one of the characters is standing on the same floor [12]. Companies originating from China sell bikinis with the design of Quranic verses [13]. Later, the insult of the Prophet Muhammad was also carried out by the Bharatiya Janata Party (BJP) in India during a television debate on the Gyanvapi Mosque. The calls for a boycott of Holywings, Netflix, China, and India. The element of provocation and

incitement was taken from the foundation for distributing donations for Aksi Cepat Tanggap (ACT) for allegedly misappropriating people's funds [14] and followed by the Ministry of Social Affairs, which revoked ACT because it was judged that there was a violation of the Minister of Social Affairs Regulation [15]. It became trending on Twitter with #KamiPercayaACT and #JanganPercayaACT.

Based on the existing problems, this research will discuss the performance value in the classification of hate speech identification on Twitter using LSTM-FastText and LSTM-GloVe. As well as the value of the influence of unbalanced and balanced data on the LSTM-FastText and LSTM-GloVe methods. The purpose of the problem is to find out the results of comparing the accuracy, precision, recall, and F1-score values of the LSTM model with word embedding GloVe and LSTM with word embedding FastText in classifying hate speech text. In addition, to analyze the value of the influence of unbalanced and balanced data on the LSTM-FastText and LSTM-GloVe methods.