I. INTRODUCTION

The rapid development of technology has greatly facilitated public access to social media. In recent years, the number of social media users in Indonesia has continuously increased, with Twitter (now X), YouTube, and Instagram being widely utilized platforms [1][2][3]. The large number of social media users often leads to uncontrolled communication, where many individuals use harsh language or hate speech [4]. The presence of social media has transformed what was originally freedom of expression into freedom to hate. Therefore, to address this issue, it is crucial to develop classification systems capable of automatically and adaptively detecting hate speech [5][6]. However, one of the primary challenges in developing such systems is data imbalance, where the volume of hate speech data (minority class) is significantly smaller than that of non-hate speech data (majority class) [7][4]. This data imbalance can cause classification models to be biased towards the majority class, potentially failing to detect the true context of texts that contain hate speech.

Various previous studies have proposed deep learning models to address this problem of hate speech detection. In this research, the Bidirectional Long Short-Term Memory (BiLSTM) model was chosen due to its ability to capture word sequence context in two directions [8][9][10][11], which is crucial for recognizing the complex textual meaning of hate speech. Furthermore, FastText word embeddings are utilized because of their capability to leverage sub-word information (n-grams) [12][13][14][15], making them effective in handling non-standard vocabulary in the Indonesian language [16].

To tackle the data imbalance challenge, this study implements and compares three oversampling methods with distinct characteristics. Random Oversampling works by randomly duplicating minority class data [7]; this simple approach serves as a baseline for comparison. SMOTE (Synthetic Minority Over-sampling Technique) generates synthetic data based on the nearest neighbors of minority samples [7], while ADASYN (Adaptive Synthetic Sampling) is more adaptive as it focuses on minority samples that are harder for the model to learn [17][18]. While individual studies have explored these components, a systematic and comparative analysis of how these specific oversampling techniques perform across varying degrees of data imbalance, when integrated with a BiLSTM-FastText architecture for Indonesian hate speech detection, and crucially, an explicit comparison of their performance outcomes, remains an underexplored area in the current literature.

The dataset used in this study is the Indonesian Hate Speech Superset, a publicly accessible dataset available on the Hugging Face platform [19]. This dataset consists of 14,306 Indonesian social media comments that have been binary-labeled as hate speech (label 1) and non-hate speech (label 0), collected from social media platforms Twitter (X), YouTube, and Instagram [19][20]. This dataset was chosen for its large scale, open access, and its representative of hate speech forms commonly found on Indonesian social media.