ENHANCING HATE SPEECH DETECTION IN TAGALOG AND TAGLISH TEXT: AN ENSEMBLE LEARNING FRAMEWORK USING BERNOULLI NAIVE BAYES, LSTM, AND mBERT

A Thesis

Presented to

The Faculty of the College of Computer Studies

Silliman University

In Partial Fulfillment

of the Requirements for the Degree

Bachelor of Science in Computer Science

NATHAN ANGELO B. CRUZ
ELLYZA MARI J. PAPAS
SAM LEONEIL G. SALA

December 2023

ABSTRACT

The rise of social media and online platforms has brought to light the prominence and harmful effects of hate speech. To combat this, these platforms have employed systems to detect hate speech. The efforts thereof have been primarily focused on more commonly spoken languages, consequently leaving languages like Tagalog, as well as the online vernacular for Filipinos, Taglish, being underexplored. With that in mind, this thesis aims to address this gap by creating and evaluating three models each using the Soft Voting, Hard Voting, and Stacking ensemble learning methods to detect hate speech in Tagalog or Taglish text. The models will be composed of the Bernoulli Naive Bayes, Long Short-Term Memory (LSTM), and multilingual Bidirectional Encoder Representations from Transformers (mBERT) learners. The performance of these models will be compared using metrics such as accuracy, precision, recall, and F1-score to determine which model performs the best. By doing so, this research seeks to shed light on the effectiveness of different ensemble learning methods and learners in detecting hate speech in Tagalog or Taglish. Ultimately, this study aims to contribute to the development of more comprehensive and inclusive hate speech detection systems on social media and online platforms.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF EQUATIONS

CHAPTER 1: INTRODUCTION

BACKGROUND OF THE PROBLEM

In the present day of this digital-driven world filled with ever-evolving computer technologies, there have been an increasing number of challenges alongside the benefits that have been introduced into this modern age. One specific challenge is the proliferation of hate speech in social media and online platforms [1]. Discriminatory, violent, vilifying, or prejudicial speech directed at groups or individuals, in the form of online slander, has the potential to degrade, divide, or propagate harm, defining what is referred to as hate speech [2]. While numerous systems have been created to detect hate speech, they are primarily tailored to widely spoken languages, which is disadvantageous to less commonly spoken languages, such as Tagalog. This could be problematic, as the Philippines is known for being the “social media capital of the world” [3], and certain demographics in the country like young adults are especially vulnerable to being victims of hate speech; these victims end up feeling threatened, angry, and embarrassed, further contributing to their exclusion from online platforms [4]. The importance of detection lies in its role as the foundational step in developing a feature that can filter or block hate speech on online platforms. Moreover, it is necessary that detection is perceived by users as reliable; users would rather “contend with an instance of hate speech than have an innocent user punished for a non-hateful post [5].”

Natural Language Processing (NLP) has become prominent during this period, with a wide variety of approaches having been developed to tackle problems like this.

These approaches have shown remarkable capabilities in understanding and generating text across multiple widely spoken languages [6]. However, there is still a lack of development for hate speech in less spoken languages like Tagalog, thus posing a significant risk of online harassment and discrimination as hate speech differs from culture to culture [7] [8] [9]. Likewise is true for Taglish, which has become the de facto vernacular for Filipinos on the Internet [10].

Recognizing the existing gaps in hate speech detection for less commonly used languages, this study is dedicated to devising a method specifically tailored for Tagalog and Taglish. This approach employs an ensemble system that integrates three distinct components: Bernoulli Naive Bayes, Long Short-Term Memory (LSTM), and multilingual BERT (mBERT). Bernoulli Naive Bayes, a probabilistic classifier, operates on the principles of Bayes' theorem, offering an effective means to model and classify binary data. LSTM, on the other hand, stands as a specialized recurrent neural network architecture, adept at capturing sequential dependencies within data, making it particularly suited for language-related tasks. Lastly, mBERT, a versatile multilingual Transformer model, possesses the capability to understand and process text across a multitude of languages. Each of them has shown promising results when used for hate speech detection; it is suggested that an ensemble of the three is worth exploring [7].

An ensemble learning strategy involves the combination of multiple learners to solve problems. That is, the output from each individual learner is considered in making the final result. Hence, the learners must be diverse to fully benefit from such a framework [11]. Ensemble learning was chosen for its demonstrably superior performance compared to any individual one for this task [12]. In particular, it was noted to be more adaptive in cross-dataset environments [13] and “help reduce the number of false positives and false negatives, which are familiar challenges in hate speech detection [14].”

There are several common methods in ensemble learning, including boosting (converting weak learners into strong learners by training base learners repeatedly, then weighing and combining them) and bagging (aggregating sampled subsets of a dataset, then training base learners and combining them) [11]. To take advantage of the aforementioned heterogeneous components, the researchers have decided to focus on stacking and voting. Stacking utilizes heterogenous base learners each trained on a partitioned data set whose predictions are used by a meta learner—Logistic Regression in this study, which is most suitable for binary text classification [15]—to generate the final output [16]. On the other hand, voting determines the output based on the majority result among learners; the voting mechanism can be either soft, which “determine[s] the class with the highest probability by averaging the individual values of the inducers”, or hard, wherein it “[sums] the votes for crisp class labels from the other inducers and [predicts] the class with the most votes [15].”

The learners are described in further detail below.

mBERT is a variant of BERT that was pre-trained on a large corpus of multilingual data in a self-supervised fashion. It predicts the probability of the input being classified for all the classes, making it a probabilistic classifier. mBERT can be used for either masked language modeling or next sentence prediction [17].

LSTM is a type of recurrent neural network (RNN) that can process data sequentially and keep its hidden state through time. It is designed to deal with the vanishing gradient problem present in traditional RNNs [18]. LSTM networks are capable of learning long-term dependencies between time steps of data, making them suitable for tasks such as text classification based on time series data.

Bernoulli Naive Bayes is a variant of Naive Bayes that works on Bernoulli distribution. It accepts features only as binary values like true or false, yes or no, or success or failure. It is used when features are binary. So instead of using the frequency of the word, there are discrete features in 1s and 0s that represent the presence or absence of a feature [19].

Together, the three learners form an ensemble geared towards enhancing the detection of hate speech in the Tagalog language, addressing a critical gap in online content detection for less commonly spoken languages. While existing solutions may work well for widely spoken languages, they often fall short when it comes to languages like Tagalog due to the lack of available data for training and fine-tuning these models.

STATEMENT OF THE PROBLEM

There have been many studies focusing on the detection of hate speech using machine learning using various approaches. There are shallow methods, which use traditionally encoded words and shallow classifiers. Typically, they predict through word occurrence rather than context [7] and perform better on datasets with less records [20]. Examples of these include Support Vector Machines (SVM) and Naive Bayes.

Meanwhile, the deep learning methods, as the name suggests, utilize deep neural networks. They generally score better across the board at the cost of computational efficiency; they also greatly benefit from being trained on a large amount of data [15]. This is further split into two: word embeddings-based methods that vectorize distributed representations of words which are combined with classifiers like recurrent neural networks (RNN) and LSTM, and transformers-based methods such as BERT which use an attention mechanism to refer to data located elsewhere [21]. Each of these approaches have their own strengths and weaknesses, consequently leading to variety in how they ultimately carry out the task. This lends credence to the development of an ensemble learning framework utilizing all three approaches.

Additionally, the main language being used in hate speech detection studies have focused on the English language. In contrast, low resource languages like Tagalog have been explored significantly less. Consequently, this necessitates research that deals with the unique challenges posed by a certain language. An issue that arises with respect to Tagalog is that it is not enough to simply consider that language alone; the vernacular commonly spoken by Filipinos online includes Taglish as well, hence its inclusion in this study.

Thus, the researchers would like to fill this gap in knowledge for Tagalog and Taglish hate speech detection by proposing an approach using an ensemble system composed of Bernoulli Naive Bayes, LSTM, and mBERT.

OBJECTIVES

The objective of this study is to develop an ensemble learning framework combining the Bernoulli Naive Bayes, LSTM, and mBERT models to detect hate speech in Tagalog or Taglish text. Three models will be developed using different ensemble learning methods, and they will be evaluated using a hate speech dataset in Tagalog and Taglish to determine which performs the best in each metric.

SCOPE AND LIMITATIONS

The scope and limitations of the study are as follows:

- This study focuses on NLP and does not examine or analyze video or audio output. All collected data will align with this scope.
- The data used for this study must be in Tagalog or Taglish languages.
- The proposed model will be an ensemble framework composed of the mBERT, Bernoulli Naive Bayes, and LSTM models.
- Classification of hate speech text will be classified into two categories: 0 (non-hate speech) or 1 (hate speech).
- The models will be trained and tested on a dataset which is a combination of the 2016 and 2022 Philippine Presidential Elections Hate Speech Dataset [8] [22], alongside a dataset obtained from the Reddit web platform, specifically focusing on comments in subreddits about the Philippines and Filipino interests.

- The evaluation of the ensemble model will be based on several performance measures, including accuracy, precision, recall, and F1-score.

SIGNIFICANCE

The development of an accurate hate speech detection system for the Tagalog and Taglish languages can have a profound societal impact. Given that these languages are prevalent in the Philippines, a comprehensive detection system would influence a significant portion of the population. Hate speech, often dismissed as mere words on a screen, carries real-world weight. Hate speech can quickly harm individuals, causing increased stress, anxiety, and depression. Additionally, it promotes discord and division within communities, creating an atmosphere of hostility and confrontation that can potentially escalate to physical violence.

The capacity to accurately identify hate speech in these languages serves as a sentinel, alerting authorities and platform moderators to inflammatory content before it can do more harm. It also serves as a deterrent, dissuading users from initially sharing such content. By diminishing the occurrence of hate speech on the Internet, it can initiate a positive chain reaction that positively impacts society overall.

This research aims to make a significant and innovative contribution to the field of hate speech detection. Most existing studies have focused on English and other widely-spoken languages, leaving a considerable gap in the literature concerning languages like Tagalog and Taglish. This study seeks to fill that void by introducing methodologies and frameworks specifically tailored for these languages. It serves as a pioneering work that not only addresses an unmet need but also enriches the global academic discourse on hate speech detection. The methodologies developed here can serve as a template for future research in underrepresented languages.

The study's focus on Tagalog and Taglish offers a unique lens through which to explore the complexities of hate speech, adding a new dimension to the academic literature. It aligns well with McLaughlin's advocacy for online regulation [23], while also addressing Williams' concerns about the tangible, real-world impact of online hate speech [24]. By focusing on these specific languages, this study adds depth and breadth to existing research, providing a more holistic understanding of the global hate speech landscape.

The Ensemble Learning Framework proposed in this study, which incorporates Bernoulli Naive Bayes, LSTM, and mBERT, has immediate and far-reaching practical applications. In the governmental context, the framework offers an invaluable tool for monitoring public discourse on social media and other online platforms. This holds clear policy significance, furnishing lawmakers with essential information to develop effective regulations addressing online hate speech. It also provides law enforcement with a more streamlined method to detect and take legal action against hate speech, thereby enhancing online safety for all users.

For social media platforms and other online communities, the ensemble framework presents a robust and scalable solution for content moderation. The use of Bernoulli Naive Bayes, known for its simplicity and effectiveness, combined with the sequence understanding capabilities of LSTM and the multilingual strengths of mBERT, makes this a particularly potent tool. It can significantly improve the user experience by creating a safer and more inclusive online environment. This, in turn, can increase user engagement and retention, offering a competitive edge for platforms that prioritize user safety.

DEFINITION OF TERMS

Bernoulli Naive Bayes. A type of Naive Bayes classifier that works on the binary concept, that determines whether the term occurs in a document or not [20].

Ensemble learning. A research area in data mining that addresses challenges posed by complex data, including imbalanced, high-dimensional, and noisy data. It seeks to create an efficient knowledge discovery and mining model by integrating data fusion, modeling, and mining within a unified framework [25].

Hate Speech. Any form of communication that discriminates, vilifies, or incites violence or prejudice against groups or individuals based on attributes such as race, religion, ethnicity, gender, or other characteristics [2]. Note that this is not the absolute definition for hate speech as no such thing exists. Rather, this is the definition that the researchers have agreed best fits the data to be used and the themes of the study.

Logistic Regression. Logistic regression is a statistical method for predicting the probability of a binary outcome based on a set of features. In the context of natural language processing (NLP), it is commonly used for text classification tasks, such as spam filtering, sentiment analysis, and topic modeling.

Long Short-Term Memory. A recurrent neural network (RNN) that is specifically designed to learn long-term dependencies in sequential data. RNNs are a type of neural network that are able to learn from sequential data, such as text or time series data [26].

mBERT (Multilingual BERT). A variant of the BERT model which is designed to understand and generate multiple languages, facilitating multilingual tasks including, but not limited to, hate speech detection [6].

Model. A mathematical representation or algorithmic structure that is trained on data to make predictions, classify information, or perform specific tasks without being explicitly programmed. Models are a fundamental component of machine learning and are used in various applications such as image recognition, natural language processing, and recommendation systems [27].

Reddit. A social news aggregation, web content rating, and discussion website that is characterized by its large and diverse user base, its relatively permissive content policies, and its anonymity features [28].

Subreddit. Is a discussion forum on the social media platform Reddit. Subreddits are organized by topic, and users can subscribe to subreddits that interest them [29].

Stratified k-fold cross-validation. A resampling procedure used to evaluate the performance of machine learning models. It is a variation of the standard k-fold cross-validation, but it is specifically designed to handle imbalanced datasets, which are common in NLP tasks.

Tagalog. The principal language spoken in the Philippines, characterized by its regional variations and rich cultural intricacies [30]. Tagalog serves as the cornerstone upon which Filipino has been constructed, with Filipino representing the organic progression of Tagalog.

Taglish. A linguistic phenomenon characterized by the frequent and systematic alternation and blending of English and Tagalog (or Filipino) in speech and communication. This code-switching practice extends beyond basic borrowing of words and involves distinct rules and patterns, occurring at various linguistic levels. It is widely observed in different contexts, notably among educated urbanites, and has expanded beyond the Tagalog-speaking region, influenced by mass media and the Filipino diaspora [31].

Transformer-Based Models. A type of machine learning model, specifically deep learning models, which are proficient in understanding and processing natural language through the use of layers of transformers, enabling them to analyze complex linguistic patterns [32].

CHAPTER 2: REVIEW OF RELATED LITERATURE

HATE SPEECH

Hate speech, a topic that has garnered significant attention in recent years, is a matter of grave concern due to its potential impact on the social fabric of communities. In a seminal study the authors embarked on a systematic analysis of hate speech in social media platforms, where they applied machine learning methods to classify and quantify hate speech. Through their meticulous approach, it was revealed that social media platforms were not merely conduits for hate speech but were, in many instances, facilitators and amplifiers [23]. The study highlighted that the algorithms used by these platforms often inadvertently promoted hate speech, which often spiraled into virality, thereby exacerbating the societal issues linked to this matter. These findings are significant as they shed light on the critical role that these platforms play in the dissemination of hate speech, demonstrating a pressing need for the refinement of these algorithms to mitigate the proliferation of such content. The results from this study beckon for more stringent regulations on social media platforms to curtail the spread and impact of hate speech.

Parallelly, a study ventured into analyzing the correlation between hate speech and real-world hate crimes [24]. Through a robust methodology that involved analyzing social media posts and corresponding crime rates, the authors were able to establish a clear link between online hate speech and an uptick in real-world hate crimes. This study is instrumental in understanding the gravity of hate speech's implications. The results are not confined to the digital space, they manifest in the real world, causing tangible harm and fostering an environment of hostility. This research serves as a clarion call to policymakers and stakeholders to enact measures that would effectively curb hate speech, not just to maintain a sense of decorum online but to prevent real-world repercussions that can be devastating.

HATE SPEECH DETECTION

The topic of hate speech detection has gained significant momentum in recent years due to the exponential growth of online platforms and the subsequent increase in online hate speech incidents. Developing efficient and precise hate speech detection systems is crucial in maintaining a safe and inclusive online environment. The first piece of literature worth mentioning is a study which aimed at classifying hate speech through supervised machine learning techniques. The study utilized a dataset comprising tweets labeled as hate speech, offensive language, or neither. The researchers employed various machine learning classifiers such as logistic regression, decision trees, and Naïve Bayes to differentiate between hate speech and other forms of offensive language [33]. A noteworthy result from this study was the elucidation that the classifiers found it relatively difficult to distinguish between hate speech and general offensive language, indicating a significant overlap between the two categories. This study signals the necessity for the development of more nuanced and sophisticated algorithms which can make this distinction more clearly, thereby aiding in the more accurate detection of hate speech on online platforms. Another vital study in this field undertook the task of identifying the linguistic and stylistic features that were most prominent in hate speech from Philippine Election-related tweets. The researchers developed a method which focused on lexical, syntactic, and semantic features, and incorporated these into a machine learning model to effectively identify hate speech [34]. They found that the incorporation of these nuanced features led to an improved performance of the hate speech detection system, particularly in identifying implicit forms of hate speech which often slip through the cracks of simpler models. The research underscored the importance of understanding the deeper linguistic structures and styles that are characteristic of hate speech, thus providing a pathway for future research in this area to focus on the development of algorithms which are capable of identifying such complex patterns.

Another effort is presented by the study that employed the Filipino BERT for hate speech detection using the TikTok Video Transcription dataset [7]. The results from this study showed that the Filipino BERT managed to achieve a 61% micro F1-score. However, when juxtaposed against the Bernoulli Naïve Bayes algorithm, it was observed that the latter was superior, achieving a 74% micro F1-score. This significant difference can be attributed to the fact that Bernoulli Naïve Bayes primarily focuses on word occurrence, which seemed to perform better for this particular dataset than contextual understanding. Furthermore, it is significant to note that the dataset used in this study only consisted of 1,000 records, which might have hindered the performance of more complex models like BERT. Thus, the limited size of the dataset might not have been representative enough for BERT to generalize effectively. For the improvement and evolution of this study, it is suggested to add more data to the dataset, integrate features such as semantic, syntactic, and lexical ones, and explore ensemble learning algorithms.

Most importantly, the results of this study showed that Bernoulli Naive Bayes, LSTM, and BERT performed the best. Bernoulli Naive Bayes was also shown to be a strong performer in the binary classification task of the former study. Another study by Velasco [9] similarly used LSTM on the same hate speech dataset, concluding that it was a viable alternative to transformer-based models especially if time or computing power is lacking. These findings all point to the potential of the three learners in the context of Tagalog hate speech. Moreover, the study by Urbano et al. [7] urged the exploration of a voting ensemble system utilizing the aforementioned learners, which this study aims to fulfill.

MULTILINGUAL HATE SPEECH DETECTION

Karim et al. [35] recognizes the need for multilingual hate speech detection; there is a dearth of material on this topic compared to English. This paper extends beyond NLP to include image processing for detecting hate speech in images, specifically in memes, citing that “only the textual data is not enough to judge [whether it is hate speech or not].” The language chosen in this study, Bengali, is significant for it being one of the most widely spoken languages in the world.

For detecting hate speech through text in this study, the researchers used BanglaBERT, multilingual BERT-cased/uncased, and XLM-RoBERTa; this is in addition to more traditional neural networks, namely: Vanilla CNN, LSTM, Bi-LSTM, and Conv-LSTM.

The study found out that all four BERT variants scored better than the DNNs. Specifically, BanglaBERT achieved 0.80, 0.79, 0.79, and 0.592 in Precision, Recall, F1-score, and MCC, respectively; while Conv-LSTM scored only 0.79, 0.78, 0.78, and 0.694 in the same categories. The best performing BERT variant was XLM-RoBERTa with 0.82, 0.82, 0.82, and 0.808. The results of this study lends credence to the efficiency and suitability of transformer-based models for sentiment analysis in hate speech.

Similarly, Dowlagar and Mamidi [36] used transfer learning with BERT and multilingual BERT models for hate speech detection. As part of the FIRE 2020 event, the researchers used the dataset provided there, taken mostly from X with some from Facebook for English, to fine-tune the model. Alongside English, the multilingual model was tuned in German and Hindi. In doing so for the base BERT model, “very few changes” were applied.

The tasks were twofold: to classify tweets into Non Hate-Offensive and Hate and Offensive, and to perform a more detailed classification into Hate Speech, Offensive, and Profane. The results of this were compared with other machine learning algorithms, namely ELMO and SVM.

On all three languages, in each task, BERT and multilingual BERT performed the best: for English, BERT scored 88.33% in both macro F1 and Accuracy for Hate Speech Detection and 54.44% and 81.57%, respectively, for Offensive Content Identification, which is a marked increase from ELMO and SVM’s 82.43% and 83.78% for Hate Speech Detection and 49.62% and 79.54% for Offensive Content Identification; these results were echoed throughout the other languages. However, it is notable that multilingual

BERT on German only scored 77.91% and 82.51% for Hate Speech Detection and 47.78% and 80.42% for Offensive Content Identification; Hindi scored 63.54% and 74.96% on Hate Speech Detection and 49.71% and 73.15% for Offensive Content Identification. This shows that multilingual BERT performs worse than BERT on English tasks, but it still performs better compared to SVM, and ELMO and SVM.

Deshpande et al. [37] sought to collect a dataset of 11 languages with the goal of classifying them in a binary fashion, whether or not the text is hate speech. The focus was to evaluate models that can recognize multilingual hate speech in three different tasks: Multilingual-Train Monolingual-Test, Monolingual-Train Monolingual-Test, and Language-Family-Train Monolingual-Test. The data was sourced from hatespeechdata.com, which is a listing of datasets used in various recent researches categorized by language, of which English, Arabic, German, Indonesian, Italian, Portuguese, Spanish, French, Turkish, Danish, and Hindi were used. Each dataset was also processed to have the same labels. It noted that the definition of hate speech differed per dataset, further compounded by the cultural nuances present in each language, so the researchers opted not to have a common definition for hate speech. The study found out that the imbalance in the amount of data available affected performance, and due to the models used (including mBERT), they could not explain precisely why the models performed this way.

BERNOULLI NAIVE BAYES

McCallum and Nigam [38] sought to describe two distinct Naive Bayes classifier models that they observed researchers used. One of them “specifies that a document is represented by the set of word occurrences from the document," and is used for statistical language modeling for speech recognition as well as text classification—multinomial Naive Bayes. The other "specifies that a document is represented by a vector of binary attributes indicating which words occur and do not occur in the document," and that the number of occurrences of a word in a given document is not tracked—a distribution based on a multivariate Bernoulli event model, or simply Bernoulli Naive Bayes, which is one of the foci of the study. In this, words are represented by either 0 or 1 based on their presence in the document; this is used in conjunction with the Naive Bayes assumption of each word's occurrence being independent of other words.

A common use of Naive Bayes classifiers is to perform sentiment analysis. The study by Ressan and Hassan [39] aimed to use them, including Bernoulli Naive Bayes, to categorize a set of 3,057 tweets about COVID-19 into joy, fear, anger, and sadness, as well as another set of 10,000 tweets into positive and negative. Bernoulli Naive Bayes scored 83.4% accuracy for the former and 85.9% accuracy for the latter.

Khezzar et al. [40] used several classification models including a specialized BERT model (AraBERT), Bernoulli Naive Bayes, SVC, and Random Forest, among others, to detect hate speech in Arabic. Notable in this study is the emphasis on detecting dialectical Arabic as well; prior efforts have only done so on the standard vernacular. To achieve this, the researchers have compiled several datasets from prior studies that were mined from X to create the arHateDataset of 34,000 tweets, split between 32% hate and 68% non-hate. After being tested on this dataset, Bernoulli Naive Bayes tied with SVC and Linear SVC for having the strongest results among the machine learning models, with an accuracy of 0.89. Meanwhile, AraBERT performed the best overall with an accuracy of 0.93, a significant improvement.

Kiilu et al. [41] proposed the use of Naive Bayes classifiers to detect hate speech in Kenyan Tweets. Collecting 45,645 tweets for training and 22,820 tweets for testing through the Tweepy library in Python, the researchers preprocessed this data by removing quotes, removing URLs, removing symbols like “@” and “#”, creating n-grams, converting it to lowercase, removing stop words, and tagging the part of speech. The best performing classifier was the Naive Bayes from the NLTK Python library, with 70% accuracy. Meanwhile, Bernoulli Naive Bayes tied with Multinomial Naive Bayes, Logistic Regression, and Linear SVC at 66% accuracy. The researchers also lamented the limitations of X preventing them from collecting more data that could have been used to improve the dataset.

Yati and Pamungkas [42] created a system to detect hate speech in Javanese using variants of Naive Bayes. The researchers used a dataset of 3,477 tweets collected from X; these tweets were then annotated either 0 (non-hate) or 1 (hate). Afterwards, the tweets were preprocessed through eliminating punctuations and digits, changing the text to be lowercase, and removing Javansese stopwords. The data was then transformed using the TF-IDF algorithm. Between Gaussian, Multinomial, and Bernoulli Naive Bayes, Multinomial scored the highest with 1.00 precision, 0.54 recall, 0.70 F!-score, and 0.98 accuracy. However, it is also worth noting Bernoulli's more polarizing spread of scores, with 1.00 precision, 0.09 recall, 0.16 F1-score, and 0.95 accuracy. The researchers suggest this to be due to the relatively uneven amounts of data between the two classifications.

The Bernoulli Naive Bayes classifier presents several compelling advantages in the domain of hate speech detection. Firstly, the model stands out for its computational efficiency. Unlike more resource-intensive deep learning algorithms, Bernoulli Naive Bayes requires less computational power and memory, making it suitable for applications with hardware constraints [43]. This is particularly important for real-time monitoring systems that need to scan vast amounts of data quickly. Secondly, the model has shown robust performance across multiple studies and languages. For instance, Khezzar et al. reported an accuracy rate of 0.89 when applying Bernoulli Naive Bayes to Arabic hate speech detection [40]. This performance suggests that despite its computational simplicity, the model can be a potent tool for identifying hate speech in various linguistic contexts.

However, Bernoulli Naive Bayes is not without its limitations. One significant weakness is its struggle with imbalanced datasets, where the proportion of hate speech to non-hate speech is skewed. In such cases, the model has been observed to perform poorly in terms of recall, as noted by Yati and Pamungkas, who recorded a recall rate of just 0.09. This low recall indicates that the model is prone to false negatives, which could be a critical issue when the aim is to eliminate all instances of hate speech. Another limitation is its relative underperformance when compared to more advanced algorithms [21]. Although Bernoulli Naive Bayes has shown strong results, it still generally lags behind the performance metrics of more sophisticated models. Khezzar et al., for example, reported that AraBERT outperformed Bernoulli Naive Bayes with an accuracy of 0.93. This gap could be consequential in applications where the highest possible accuracy is required.

LSTM

Recognizing the need to improve the storage of information over time via backpropagation, Hochreiter and Schmidhuber [26] proposed the recurrent network architecture with gradient-based algorithm called LSTM. It efficiently enforces a constant error flow, solving this problem. It is also reported to be able to deal with noise and generalization.

Akter et al. [44] created a state-of-the-art model that can detect cyberbullying using a modified LSTM-Autoencoder Network. They experimented with this model and compared it to LSTM, BiLSTM, Word2vec, BERT, and GPT-2.

The dataset used in this study was from Trac-2, containing 25,000 comments from Facebook, Youtube, and X in English, Bengali, and Hindi. These comments were classified based on two sub-tasks: one that divides comments into Non-Aggressive, Overtly Aggressive, and Covertly-Aggressive; and one that indicates whether a comment is misogynistic or not. The study focused on the former sub-task for it was more aligned with the objectives of this study.

To augment their data, the researchers have opted to insert noise due to the imbalance in the dataset, creating a semi-noisy dataset. They also translated the Bangla and Hindi text to English with machine translation, producing a dataset of purely noise.

These two datasets alongside the raw text were tokenized with the BERT tokenizer for the BERT models and TensorFlow for the others.

Each model was then trained with each language separately, with a split of 70% for training and 30% for validation.

For the raw data, of the existing models, GPT-2 was found to have performed the best in English (0.80 accuracy, 0.76 precision, 0.80 recall, and 0.77 F1-score) and Bangla (0.73 accuracy, 0.74 precision, 0.73 recall, 0.73 F1-score). Meanwhile, BERT outperformed in Hindi (0.69 for all four metrics). All of these were surpassed by the proposed model, with figures at 0.90 across the board for all three languages. The proposed model maintained the scores of 0.90 and over for all three languages in the semi-noisy dataset as well as the fully translated noisy English dataset. These results prove the strength of this modified LSTM-Autoencoder, and that when tweaked to fit the problem, can achieve superior results in classification tasks.

Das et al. [45] sought to classify Bengali hate speech collected from Facebook comments into seven categories: hate speech, aggressive comment, religious hatred, ethnical attack, religious comment, political comment, and suicidal comment. The machine learning algorithms used to construct the model were attention, LSTM, and GRU based decoders each in conjunction with CNN; these classification approaches were compared. For feature extraction, they used TF-IDF. Notably, they also included a module to detect the meaning of emojis, which further influence the classification of hate speech. The researchers ultimately found out that the attention-based encoder-decoder performed the best, with 0.77 accuracy, 0.78 precision, 0.75 recall, and

0.78 F1-score; the next highest one being LSTM with 0.74 accuracy, 0.72 precision, 0.71 recall, and 0.72 F1-score.

LSTM networks have exhibited considerable strengths in the field of hate speech detection. One of the most significant advantages is their capability to handle sequential data effectively, which is crucial for text-based tasks. This strength allows LSTMs to capture the context within a text, providing a more nuanced classification [46]. For instance, Das et al. reported that LSTM achieved an accuracy of 0.74, only slightly behind the attention-based encoder-decoder model. Furthermore, Akter et al. found that a modified LSTM-Autoencoder outperformed other notable models like GPT-2 and BERT across multiple languages, showing its adaptability and robustness [47]. This advanced performance implies that LSTM models, especially when fine-tuned, can offer high accuracy, making them a formidable tool for detecting hate speech.

On the flip side, LSTMs also have limitations that can hinder their effectiveness in certain scenarios. One of the main drawbacks is their computational intensity. These models require significant computational resources for training and inference, which may not be viable for all applications. Another potential issue arises from their complexity; LSTMs involve numerous parameters that need to be optimized, making the model prone to overfitting if not properly regulated [47]. Lastly, while LSTMs perform well, they can sometimes be outperformed by other specialized models. For instance, in the study by Das et al., an attention-based encoder-decoder achieved slightly better performance metrics than LSTM. This suggests that while LSTMs are powerful, they may not always be the best choice for every hate speech detection task.

mBERT

In the contemporary landscape of natural language processing (NLP), an array of pre-trained models like BERT and its variants have emerged, reflecting a significant evolution in how text data is processed and analyzed. The following paragraphs undertake a meticulous examination of these two prominent models, with a concentrated emphasis on the results derived from studies and literature. Notably, the information presented here is based on the data available up to September 2021.

mBERT, or Multilingual BERT, represents a significant milestone in the NLP sphere, fostering the integration and analysis of multiple languages within a singular pre-trained model. Developed by researchers at Google, this model is grounded in the BERT architecture, which employs a transformer neural network [48]. A distinctive characteristic of mBERT is its capacity to process and analyze 104 languages, a feature facilitated through the training on a multilingual corpus that incorporates diverse linguistic structures and semantics.

In a research study, mBERT displayed notable proficiency in zero-shot cross-lingual transfer, a process where a model trained in one language is able to successfully perform tasks in another language without additional training [49]. This study meticulously analyzed the performance of mBERT across a variety of linguistic tasks, demonstrating that the model could achieve competitive results even in scenarios where language-specific training data was sparse or non-existent. Furthermore, the research highlighted that mBERT exhibited remarkable performance in sentence classification tasks, outperforming several monolingual baselines. However, it was also noted that the performance could potentially be influenced by the typological proximity between the languages involved.

Nozza [50] also sought to research the extent of the effectiveness of zero-shot multilingual hate speech detection. This study recognized several issues that arise with respect to this domain. One is that hate speech is a broad concept, covering misogyny, racism, and others; they must be handled differently depending on the specific kind of hate speech a text is. Second is that the available data from prior research on this topic are inconsistent; the exact definitions and limits to what counts as hate speech can differ. Finally, it was noticed that the vast majority of ground covered is in English alone, when hate speech is not exclusive to any one language; there are specific nuances in each language that require specific consideration as well. Thus, to fill in these gaps, the researcher investigated the detection of hate speech targeting women and immigrants exclusively and experimenting with zero-shot, cross-lingual hate speech detection using mBERT.

The datasets used in the study were HatEval for English and Spanish, alongside the automatic misogyny identification challenge and hate speech detection shared task for Italian. Surprisingly, for the datasets on hate speech towards immigrants as well as the one on women, English consistently had the lowest F1-scores at 0.368 and 0.559, respectively for monolingual, with cross-lingual results being close. So for the purposes of cross-lingual zero-shot transfer learning for hate speech detection, it does not solve the lack of models and data on the problem, positing that hate speech is highly intertwined with a specific language and must be studied on those merits.

The task of using BERT for hate speech detection is so strong that Caselli et al. [51] have sought to retrain BERT for this exact purpose. Dubbed HateBERT, it was trained on RAL-E, or the Reddit Abusive Language dataset; this means it is equipped to handle more than just hate speech, including "microaggression, stereotyping, offense, abuse, … threats, and doxxing." Experiments using this new pre-trained model were ran on the following datasets: OffensEval 2019, AbusEval, and HatEval. Compared to BERT, HateBERT fared better with 80.9% over 80.3% in the macro F1 Pos. class for OffensEval2019, 76.5% over 72.7% for AbusEval, and 51.6% over 48% in HatEval. The results of this study show how effective it is to further train models like BERT towards a specific problem domain to achieve better results.

Saleh et al. [52] recognized the challenges of identifying hate speech, especially with models that have previously taken a long time to extract results from. To this end, they sought to evaluate the effectiveness of a BiLSTM and BERT model, compared to previous results from Word2Vec and GloVe, which are models built on the word embedding technique. The datasets tested on were from Davidson et al. (hate speech, offensive, and neither), Waseem (racism, sexism, both, and neither), and Waseem and Hovy (sexist, racist, and neither); the classes were collapsed to be uniform among datasets. The results showed that BERT, between its base and large models, performed better than the other models and methods. That is, BERT outperformed even models that are trained to be domain-specific. To further understand the results of BERT in this study, the researchers used the LIME strategy which essentially tinkers with parameters in the model and observes how it affects output.

Another innovative study that warrants discussion sought to develop an audio-based hate speech classifier using traditional machine learning algorithms, particularly focusing on TikTok videos [53]. This study stands apart as it leveraged audio-based features, such as MFCCs, Spectral Centroid, Rolloff, Bandwidth, Zero-Crossing Rate, and Chroma values, from a considerably larger dataset of 4,746 videos. The outcomes of this study were promising, with an optimized Random Forest model achieving an accuracy of 78.5%. An important takeaway from this research is the revelation that Spectral Rolloff and MFCCs stand out as top predictors when it comes to identifying hate speech in the Filipino language. This indicates that while textual content is essential, audio-based features can be equally pivotal and provide nuanced insights into the task of hate speech detection.

Thus, while advancements in the field of hate speech detection using models like BERT are commendable, traditional algorithms and diverse features can sometimes provide comparable or even superior results. This underscores the importance of understanding the dataset at hand and choosing the appropriate model and features accordingly.

The BERT model has demonstrated remarkable strengths in the domain of hate speech detection. One significant advantage is its adaptability to domain-specific tasks. Caselli et al. retrained BERT into HateBERT and achieved superior performance over the original BERT model across multiple datasets, showing that fine-tuning BERT can yield better results. Another strength is its ability to handle complex linguistic constructs, such as context and polysemy, which are often found in hate speech [51]. This capability makes BERT exceptionally useful in identifying more subtle forms of hate speech, including microaggressions and stereotyping. Furthermore, Saleh et al. found that BERT outperformed other models, including domain-specific ones, indicating its robustness and generalizability across different types of hate speech and languages.

Despite its strengths, BERT also has limitations when applied to hate speech detection. One major drawback is its computational intensity. BERT models are resource-heavy, requiring significant computational power for both training and inference, which may not be feasible for all applications [51]. Additionally, while BERT's complexity allows it to capture nuances in language, it can also make the model difficult to interpret. Saleh et al. had to employ the LIME strategy to better understand BERT's output, indicating that the model's decisions are not easily explainable. Lastly, although BERT performs exceptionally well, it does not always guarantee the best results for every type of hate speech. Traditional algorithms, as seen in the study focusing on audio-based hate speech detection, can sometimes yield comparable or even superior results.

ENSEMBLE APPROACHES

Aljero and Dimililer [54], in their study, created a novel stacked ensemble approach using SVM, LR, and XGBoost for detecting hate speech in English tweets. The framework of their ensemble is composed of first-level base-level classifiers (SVM, LR, XGBoost) and a second-level meta-level classifier (LR). After employing feature extraction with Word2Vec and USE, the outputs of the former were fed to the latter to obtain the final prediction. This study made use of four public datasets: HatEval, Davidson, COVID-HATE, and ZeerakW; all four of these datasets were recategorized to fit the binary classification scheme of non-hate (0) and hate (1). and subsequently oversampled using SMOTE to balance the datasets. This ensemble approach scored 0.6551, 0.9713, 0.7301, and 0.7392, in each dataset while the best scoring classifier per dataset was LR, SVM, XGBoost, and LR, respectively. The ensemble's results on the first three datasets surpass that of the state of the art, proving the effectiveness of ensemble systems.

Nurriski et al. [55] aimed to further sentiment analysis of hate speech with a stacking ensemble learning technique, focusing on improving accuracy and F1-score. The researchers noted the limitations of sentiment analysis in this problem, as hate speech "has its own characteristics and is not always related to the opinion or sentiment expressed," but nevertheless expressed its importance. The dataset utilized in this study was Tweeter Hate Speech Analysis available on kaggle, containing 31,692 tweets and their corresponding label of whether or not the text is hate speech (0 for non-hate, 1 for hate). In sentiment analysis, each tweet had their subjectivity and polarity extracted, which fed into the analysis of whether the tweet was positive, negative, or neutral. The stacked ensemble learning approach used involved the combination of SVM, Decision Tree, and Random Forest; the predictions from each model is fed to a Meta-Learner to produce the output. This approach scored 96.11% in accuracy, 96.07% in precision, 96.10% in recall, and 96.07 in F1-score, a marked improvement from other models like CNN (79% accuracy) and BERT (75% accuracy), fulfilling the objective of the study and showcasing the potential of ensemble learning systems for this task.

Markov et al. [56] used a combination of BERT models to create an ensemble approach for detecting cross-domain hate speech in Dutch. This study's emphasis on cross-domain hate speech is due to the "substantial drop in performance when evaluating hate speech detection approaches on out-of-domain datasets." In doing so, the researchers made use of two existing Dutch datasets: LiLaH from Facebook comments and DALC from X tweets. They also made use of the models BERTje, a monolingual Dutch variant of BERT; RobBERT, a Dutch variant of the RoBERTa model; and SVM to create their ensemble. These models were trained with stratified five-fold cross-validation. In-domain, the ensemble performed the best for both datasets over any single model, scoring 78.8 precision, 78.2 recall, and 78.4 F1-score for LiLaH; and 84.9 precision, 75.0 recall, and 77.2 F!-score for DALC. For cross-domain, the ensemble performed worse but still superior in all metrics except one: BERTje scored 74.4 precision over 74.1 in the "train DALC, test LiLaH (entire)" experiment. This result notwithstanding, the use of an ensemble proved to be more fruitful than any singular model.

Mutanga et al. [15] applied an ensemble of models to identify hate speech in both English and Spanish, demonstrating the versatility and adaptability of their approach across different languages. The paper used decision trees, logistic regression, SVM, CNN, and LSTM.

The CNN and LSTM models were built using the Keras library. The CNN model used pooling techniques, which help reduce the spatial size of the convoluted feature, reducing the computational complexity and helping with overfitting issues. The LSTM model tackled the vanishing gradient problem, a difficulty often encountered in traditional neural networks where information gets lost over time or through layers.

Mnassri et al. [16] presented a groundbreaking study on the detection of hate speech using an ensemble approach. The research utilized three distinct models, each with its own unique configuration.

The first model, BERT+LSTM, incorporated LSTM layers into BERT’s sequence output. This model consisted of two LSTM layers with 512 units each, followed by a dropout and a dense layer.

The researchers employed four ensembling methods to combine BERT+MLP, BERT+CNN, and BERT+LSTM. These methods included soft voting or averaging, maximum voting, hard voting, and stacked generalization ensemble or stacking.

Soft voting took the average of predicted class probabilities of each individual classifier and used argmax to obtain the final class. Maximum voting considered the maximum prediction probability from the models, resulting in the class with the highest probability among the classifiers. Hard voting used the principle of majority voting (of an odd number of classifiers). It took the predictions of each model and outputted the most frequent class.

Stacked generalization ensemble combined heterogeneously trained base learners by training a meta-model to output a prediction based on the predictions of the base models. The researchers implemented the Stratified k-fold cross-validation technique to partition the training set between the models, maintaining a ratio of 10% for the validation set. Each base learner was trained on a subset of the training set, and their predictions were fed as training sets into the meta learner (a Linear Regression classifier in this case), which output the final predictions.

CHAPTER 3: THEORY AND METHODS

THEORETICAL BACKGROUND

The theoretical background plays a critical role in this study by providing the necessary context for the research on ensemble learning models for hate speech detection in Tagalog and Taglish. This background comprises an overview of the ensemble learning models considered in this study (Bernoulli Naive Bayes, LSTM, mBERT), along with their different classification algorithms. This background information is essential for establishing the current state of knowledge in the field of hate speech detection, identifying gaps or limitations in the existing literature, and providing a theoretical framework for the proposed research. The use of ensemble learning models for hate speech detection in less commonly spoken languages is a relatively new area of research, and this study aims to contribute to this field by developing and evaluating a novel ensemble learning framework for Tagalog and Taglish.

VOTING

Voting is a heterogeneous ensemble method that combines the predictions of different machine learning algorithms to make a prediction, used in this study due to the heterogeneous set of learners used to build the ensemble. It is typically used for classification problems, and there are two main types of voting schemes: hard voting and soft voting. In hard voting, the class that was predicted by the majority of base models is used as the final prediction. In soft voting, the base models have varying degrees of influence when it comes to the final prediction, which is weighted by the classifier's importance and summed up [57].

$$\widehat{y} = softmax\left\{\sum_{i=1}^{n}\sum_{j=1}^{m} p_{i,j}\right\} \qquad (1)$$

Equation (1) shows soft voting [16] [57], which specifically averages the predicted class j and probabilities p of each classifier, where n is the amount of classifiers, then utilizes softmax to get the final output. That is, for every classifier, each class iis weighted the same.

$$\widehat{y} = mode\{C_1, C_2, \dots C_m\} \qquad (2)$$

Meanwhile, hard voting, as shown in (2) [16], simply gets the class $\widehat{y}$ from the majority vote of each individual classifier C, where m is the amount of classifiers. This majority is represented by the $mode$ of all results.

Both methods will be used, wherein a model will be built with each method. This is due to the appreciable differences in their algorithms, thereby potentially leading to different results that are worth exploring.

STACKING

Stacking or Stacked Generalization is a type of ensemble learning that combines the predictions of multiple machine learning models (the base models) to produce a more accurate and robust prediction. Stacking does this by training a second-level model (the meta-learner) on the predictions of the base models. The meta-learner learns to combine the predictions of the base models in a way that minimizes the overall error [54]. Like voting, it is also ideal for heterogeneous learners.

$$z = stack\left\{ \sum_{i=1}^{n} h_i(D) \right\} \tag{3}$$

Equation (3) [16] illustrates how the stacking ensemble works. Each base learner h is trained on a partition of the dataset D, where n is the total number of learners. These comprise the $stack$ and their predictions z are used to train the meta learner—logistic regression, in this study. To partition the training dataset for each learner, the stratified k-fold cross-validation technique will be used, where the k is equal to the number of learners which is three. The use of stratified cross-validation specifically mitigates potential imbalances between partitions, which is especially helpful for classification problems such as this.

BERNOULLI NAIVE BAYES

Bernoulli Naive Bayes (BNB) is a simple yet effective machine learning algorithm for hate speech detection. It is a probabilistic model that assumes that the features of a text document are independent of each other. BNB works by calculating the probability of a document belonging to each class based on the presence or absence of each feature in the document.

$$P(tk|c) = \frac{1+|Trtk,c|}{2+Trc} \tag{4}$$

Equation (4) [39] illustrates the probability of a binary event, where the event can only have two outcomes, such as success or failure, heads or tails, or true or false.

The equation states that given a term t, the occurrence of the term in a document tk, and a class c, the probability of tk in a document of c is equal to the number of documents where the term appeared $Trtk, c$ over the total number of documents for the class Trc. Additional numbers—1 in the numerator and 2 in the denominator—were added to perform Laplace smoothing, preventing the probability from being totally zero and therefore unusable.

LSTM

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) that is well-suited for sequence-to-sequence learning tasks, such as hate speech detection. LSTM networks are able to learn long-term dependencies in sequences of data, which is important for hate speech detection, as hate speech can be expressed in a variety of ways, including through the use of sarcasm, irony, and other indirect forms of language [58].

$$\begin{aligned}
f_t &= \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \\
i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \\
o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \\
c_t &= f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \\
h_t &= o_t \circ \sigma_h(c_t)
\end{aligned} \tag{5}$$

Equation (5) [58] above illustrates how LSTM works. Forget gate (f_t) controls what information to retain, while the input gate (i_t) and candidate cell state (c_t) determine new information to incorporate. The cell state (c_t) is updated based on forget, input, and candidate cell state results. The output gate (o_t) determines the current hidden state (h_t) from the updated cell state. LSTMs capture long-term dependencies through these coordinated components.

Other legends include:

- x_t: Input to the current timestamp.
- U: Weight associated with the input.
- h_{t-1}: The hidden state of the previous timestamp.
- W:The weight matrix associated with the hidden state.
- b: Represents the bias term associated with each gate in the LSTM cell.
- σ: Sigmoid activation function. It takes an input value (x) and maps it to a value between 0 and 1.

mBERT

Multilingual Bidirectional Encoder Representations from Transformers (mBERT) is a pre-trained language model that has been shown to be effective for a variety of natural language processing (NLP) tasks, including hate speech detection. mBERT is a bidirectional encoder model, which means that it can learn the context of a word from both the words that come before it and the words that come after it. This makes mBERT well-suited for identifying complex patterns in text, such as those that are indicative of hate speech [59].

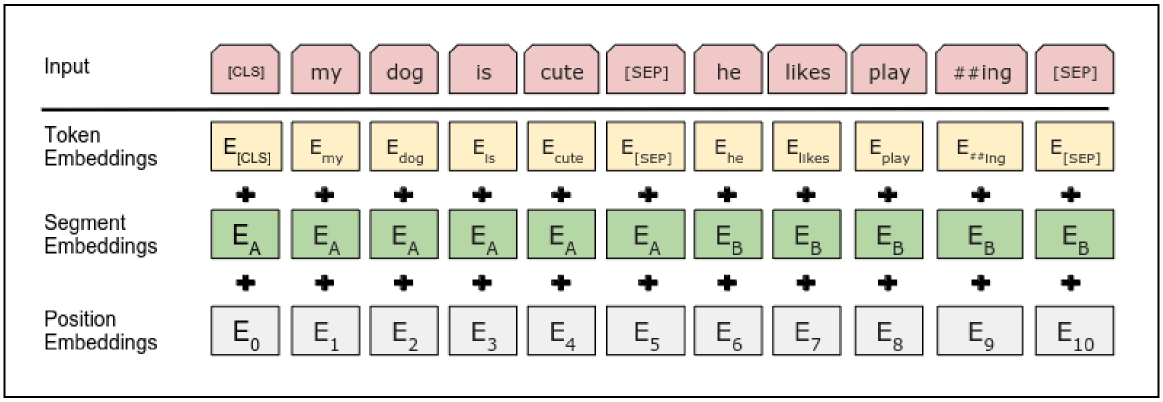

Input
[CLS]
my
dog
is
cute
[SEP]
he
likes
play
##ing
[SEP]
Token Embeddings
$E_{[CLS]}$ E_{my} E_{dog} E_{is} E_{cute} $E_{[SEP]}$ E_{he} E_{likes} E_{play} $E_{\#\#ing}$ $E_{[SEP]}$
Segment Embeddings
E_A E_A E_A E_A E_A E_A E_B E_B E_B E_B E_B
Position Embeddings
E_0 E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10}

Figure 1. BERT Input Representation

Figure 1 [32] visualizes how BERT handles each token sequence. Every sequence begins with a special classification *[CLS]* token, and a special separator *[SEP]* token denotes another sentence. Each token is represented by the sum +of its token, segment, and position embedding.

TF-IDF

Term Frequency - Inverse Document Frequency, known as TF-IDF, is a widely utilized technique for extracting features in the field of natural language processing. Plain text cannot be directly processed by machine learning algorithms. It is necessary to convert text into numerical values before using it as input for a machine learning model. The TF-IDF score of a word is determined by multiplying two components: term frequency, which indicates the frequency of a term within a document or sentence, and inverse document frequency, which reflects how common or uncommon a word is within the entire set of documents [60].

Term-Frequency

$$TF(t) = \frac{(number\ of\ times\ term\ t \text{ appears in a document})}{(total\ number\ of\ terms\ in\ the\ document)}$$

Inverse Document Frequency

$$IDF(t) = \frac{log\ (total\ number\ of\ documents)}{(number\ of\ documents\ with\ term\ t\ in\ it)} \tag{6}$$

$$TF\text{-}IDF = TF \times IDF$$

The above equation, (6) [8], illustrates how TF-IDF works. Essentially, to calculate the term frequency, it divides the occurrences of a given term t by the amount of all terms in the document. Its result is multiplied to the inverse document frequency, which takes the logarithm of the number of documents over the amount of documents with the given term.

LOGISTIC REGRESSION

A statistical method for estimating the probability of binary outcomes based on input features, is highly instrumental in NLP, finding extensive utility in text classification tasks like spam filtering, sentiment analysis, and topic modeling. In NLP, it leverages feature-text relationships to predict the likelihood of specific outcomes, such as categorizing emails as spam or not, determining sentiment in text (positive or negative), or identifying topics within a corpus of text. Its simplicity, interpretability, and capacity to offer insights into feature importance make logistic regression a valuable and efficient solution in NLP, aiding in addressing various text classification challenges while maintaining a probabilistic perspective for nuanced language data [61].

$$z = \left(\sum_{i=1}^{n} w_i l_i\right) + b \qquad (7)$$

The process of logistic regression is shown in (7) [62]. Each learner l up to n times its weight w is summed together; the result of this is added to the bias b to create the final predicted output. In this study however, each learner is weighted the same, thereby eschewing the need for w in the equation.

STRATIFIED K-FOLD CROSS-VALIDATION

Stratified k-fold cross-validation is a vital resampling technique in the field of NLP for evaluating machine learning models, especially when dealing with imbalanced datasets. It extends the traditional k-fold cross-validation method by ensuring that the distribution of different classes within each fold closely mirrors the distribution in the original dataset. This approach is critical in NLP tasks where certain categories or labels may be rare, as it allows for a more accurate assessment of a model's performance by ensuring it is tested on representative examples from all classes, including the minority ones. This ensures robust evaluation and reliable model development, making it particularly valuable in NLP applications where maintaining class balance is often challenging [63].

$$\begin{gathered} D = \{D_1, D_2, \dots D_k\} \\ train = D_i \\ test = D/D_i \end{gathered} \qquad (8)$$

The partitioning of the dataset is covered in (8) [16]. The dataset D is comprised of several partitions i up to k. The training dataset in each instance is D_i while its respective test datasets are the rest of the partitions represented by D/D_i. For this study, as part of the stacking ensemble, there are three folds—representative of the three learners in the ensemble—that the training dataset will be further partitioned into.

SOLUTION TO THE PROBLEM

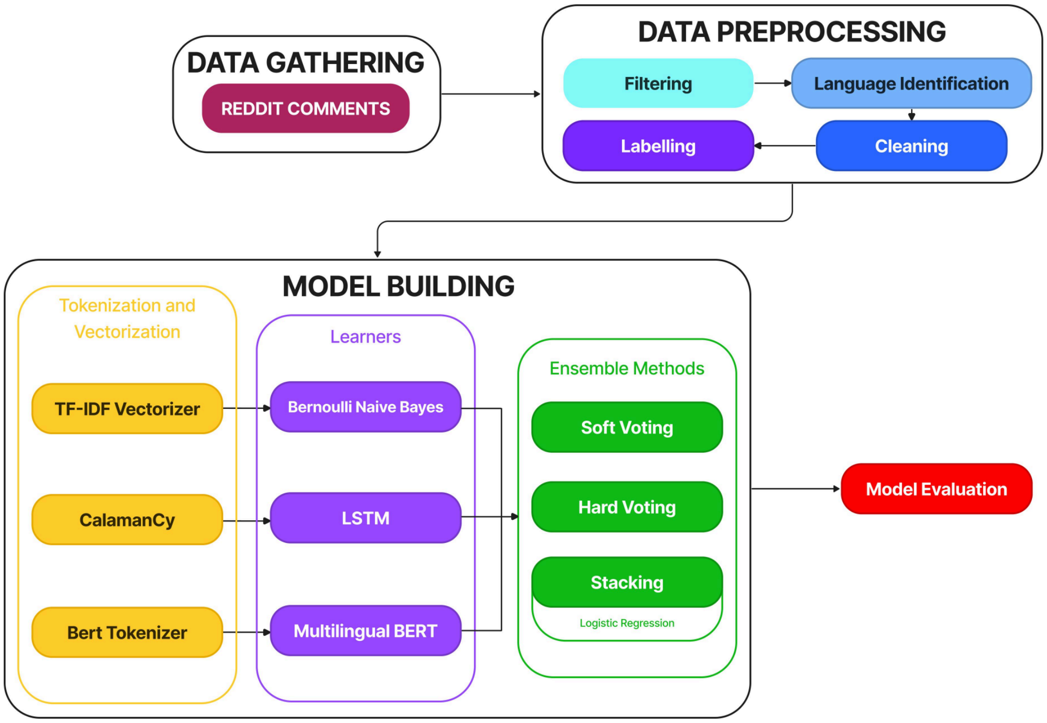


Figure 2. Conceptual Framework

Figure 2 illustrates the process of the study. First, data is gathered from Reddit comments, as outlined in the *Data Gathering* section. The gathered data is then preprocessed through a series of procedures: filtering, language identification, cleaning, and finally labeling, as explained in the *Data Annotation* portion. Afterwards, the model proper is built—this first involves the *Tokenization and Vectorization* of the data, converting it into a format suitable for processing by the learners Bernoulli Naive Bayes, LSTM, and mBERT. These three learners compose the ensemble formed by two different methods: voting and stacking, as shown in the *Model Building* section, and its training and testing are detailed in the *Training and Testing* section. Finally, *Model Evaluation* talks about the metrics used to show the performance of the models.

DATA GATHERING

The additional data to be used will be collected from certain Philippine and Filipino-related subreddits throughout October 2023; these subreddits are listed in Appendix B [64]. The researchers will use the *Python Reddit API Wrapper (PRAW)* library to obtain this data.

Specifically, the researchers will filter each subreddit by "controversial" and "top." In each filter, the comments of each post will be gathered. The attributes to be collected are: the comment ID, the subreddit name, the post's name, the post's body text, the author of the comment, the text of the comment, the score of the comment, and the timestamp.

DATA PREPROCESSING

The collected data will be saved into a CSV file. All of the aforementioned attributes are to be dropped, save for the post's name, post's body, and text of the comment. This will be done to preserve the original context of the comment as much as possible, which can aid in the data annotation process.

First, the collected data will be filtered as such:

- Exclusion of comments made by "*AutoModerator*" users
- Exclusion of empty, deleted, and removed comments

Additionally, since some of the comments gathered may be written fully in English, language detection will be performed using the *Lingua* Python library. The researchers set it to compute the language confidence of a comment in both Tagalog and English. Comments that did not reach a minimum threshold of 0.50 in Tagalog will be excluded.

Afterwards, cleaning the data will involve the following processes:

- Removal of URLs
- Removal of username mentions
- Removal of Markdown formatting characters
- Removal of characters that caused issues in the scripts when the data was loaded

DATA ANNOTATION

After initially filtering and cleaning the data, it will be annotated by an expert—Mrs. Jessica Kitane, RGC, a guidance counselor at the Silliman University Medical School. The expert will be able to read the post name, the post body, and the comment itself; they are to label based on the comment as the details of the post are there to provide further context. The two possible labels to be given are *0* for non-hate speech, and *1* for hate speech.

TOKENIZATION AND VECTORIZATION

In order to feed the data to the models, the text must be transformed first into numerical values representing the words themselves—tokenization and vectorization. Both of these will be achieved through the TF-IDF vectorizer in the *scikit-learn* Python library, as well as *CalamanCy* [65], which contains word vectors trained on the TLUnified dataset.

For mBERT, it will require the use of its own *BERT Tokenizer*, which encodes the data, retrieving input IDs and their respective attention masks; this functionality necessitates the use of another tool to perform this task.

MODEL BUILDING

Three ensemble learning models are to be built: a stacking model, using logistic regression as the meta-learner; and two voting models, utilizing a soft voting strategy for one and hard voting for the other. Each model will make use of the same learners, namely: Bernoulli Naive Bayes, LSTM, and mBERT.

TRAINING AND TESTING

The models will be trained and tested on a dataset formed from the combination of the 2016 and 2022 Philippine Presidential Elections Hate Speech Dataset [8] [22] and the gathered data from the comments on Philippine-related subreddits. This dataset will be split by a ratio of 80:20 for training and testing, respectively. Additionally, the stacking ensemble will further partition the training data for its individual learners using the stratified k-fold cross-validation technique.

MODEL EVALUATION

After training and testing, each model will be evaluated using four metrics: accuracy, precision, recall, and F1-score.

$$\begin{aligned}
&Precision = \frac{TP}{TP+FP} \\
&Recall = \frac{TP}{TP+FN} \\
&Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \\
&F1\ Score = 2\frac{Precision * Recall}{Precision + Recall}
\end{aligned} \tag{9}$$

Equation (9) [40] illustrates each of these measures. Precision refers to the amount of correctly predicted positive samples TP (i.e. hate speech) over all the possible positive samples, which includes the false positives FP. Recall details the quotient of the true positives alongside the true positives and false negatives FN. Accuracy measures the amount of correct predictions over the total amount of predictions, inclusive of the true negative TN. Finally, F1-score refers to the mean of the precision and recall. The results among the three models will be compared to see which performs the best in each metric.

PROJECTED TIMELINE

The data will be collected from October 2023 to November 2023. The months of December 2023 to January 2024 will be for data annotation. The months of January

2024 to February 2024 will be for creating the model and evaluation. Interface building will happen in the month of February 2024.

ETHICAL CONSIDERATIONS

This research paper makes extensive use of professional and ethical standards of research, including the requirement that all data presentations and interpretations be truthful, free of dishonesty and fraud, and properly credit the contributions of others. As a result, the document conveys honesty, integrity, and truthfulness. Additionally, the researchers ensure that the study is conducted with the following ethical considerations in mind:

Data Gathering: The data collected for this study are obtained from Reddit, specifically the subreddits listed in Appendix A. The research team ensures that all data collection and usage adhere to Reddit's terms and conditions. As the data used are publicly available and part of the public discourse on the platform, informed consent is not required. The ethical use of public data from the subreddit is strictly followed.

Data Preparation: To respect user privacy, only non-identifying data are collected and utilized in the study. Specifically, the content from the subreddit is used for analysis, while data that could potentially identify users or raise privacy concerns, such as specific usernames, are not gathered. These measures are put in place to protect the privacy and anonymity of individuals participating in the subreddit.

Data Sharing: To uphold user privacy and data integrity, the dataset used in the research will not be made publicly available. The results presented in the paper are conclusive and reflect the ethical considerations of data usage and user privacy.

Data Disposal: After the research is concluded, data disposal procedures are rigorously followed to ensure that all data collected is securely and irreversibly deleted, in compliance with data privacy regulations and ethical standards. Data disposal is carried out to prevent any unintended or unauthorized access to the data used in the study.

BIBLIOGRAPHY

[1] D. Antypas and J. Camacho-Collados, “Robust Hate Speech Detection in Social Media: A Cross-Dataset Empirical Evaluation.” arXiv, Jul. 04, 2023. Accessed: Sep. 14, 2023. [Online]. Available: http://arxiv.org/abs/2307.01680

[2] “Hard Questions: Who Should Decide What Is Hate Speech in an Online Global Community?,” Meta. Accessed: Sep. 14, 2023. [Online]. Available: https://about.fb.com/news/2017/06/hard-questions-hate-speech/

[3] A. I. | Dec 7 and 2020, “Filipinos Lead the World in... Social Media ‘Addiction,’” Esquiremag.ph. Accessed: Oct. 02, 2023. [Online]. Available: https://www.esquiremag.ph/money/industry/filipinos-social-media-addiction

[4] N. Manarpiis, K. Cortez, M. Cortez, B. Nicole, and L. Mendoza, “Online Hate Speech and the Personal Experiences of Young Adult Filipinos,” presented at the Parañaque National Research Conference 2021, Parañaque City College, Dec. 2021.

[5] P. Lammerts, P. Lippmann, Y.-C. Hsu, F. Casati, and J. Yang, “How do you feel? Measuring User-Perceived Value for Rejecting Machine Decisions in Hate Speech Detection,” in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, Aug. 2023, pp. 834–844. doi: 10.1145/3600211.3604655.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” arXiv, May 24, 2019. Accessed: Aug. 31, 2023. [Online]. Available: http://arxiv.org/abs/1810.04805

[7] R. Hernandez Urbano Jr., J. Uy Ajero, A. Legaspi Angeles, M. N. Hacar Quintos, J. M. Regalado Imperial, and R. Llabanes Rodriguez, “A BERT-based Hate Speech Classifier from Transcribed Online Short-Form Videos,” in *2021 5th International Conference on E-Society, E-Education and E-Technology*, Taipei Taiwan: ACM, Aug. 2021, pp. 186–192. doi: 10.1145/3485768.3485806.

[8] N. V. P. Cabasag, V. R. C. Chan, S. C. Y. Lim, M. E. M. Gonzales, and C. K. Cheng, “Hate Speech in Philippine Election-Related Tweets: Automatic Detection and Classification Using Natural Language Processing”.

[9] D. J. Velasco, “Pagsusuri ng RNN-based Transfer Learning Technique sa Low-Resource Language.” arXiv, Oct. 14, 2020. Accessed: Sep. 15, 2023. [Online]. Available: http://arxiv.org/abs/2010.06447

[10] D. Osborne, “Chapter 10. Me, myself, and ako: Locating the self in Taglish tweets,” in *Pragmatics & Beyond New Series*, vol. 304, P. Bouissac, Ed., Amsterdam: John Benjamins Publishing Company, 2019, pp. 235–252. doi: 10.1075/pbns.304.10osb.

[11] Z.-H. Zhou, *Machine learning*. Singapore: Springer, 2021.

[12] A. F. M. de Paula, R. F. da Silva, and I. B. Schlicht, “Sexism Prediction in Spanish and English Tweets Using Monolingual and Multilingual BERT and Ensemble Models.” arXiv, Nov. 08, 2021. Accessed: Oct. 02, 2023. [Online]. Available: http://arxiv.org/abs/2111.04551

[13] S. Agarwal and C. R. Chowdary, “Combating hate speech using an adaptive ensemble learning model with a case study on COVID-19,” *Expert Syst. Appl.*, vol.

185, p. 115632, Dec. 2021, doi: 10.1016/j.eswa.2021.115632.

[14] A. Haque and M. N.-U.-R. Chowdhury, “Hate Speech Detection in Social Media Using the Ensemble Learning Technique,” *Int. J. Adv. Netw. Appl.*, vol. 15, no. 01, pp. 5815–5821, 2023, doi: 10.35444/IJANA.2023.15111.

[15] R. T. Mutanga, N. Naicker, and O. O. Olugbara, “Detecting Hate Speech on Twitter Network using Ensemble Machine Learning,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 3, 2022, doi: 10.14569/IJACSA.2022.0130341.

[16] K. Mnassri, P. Rajapaksha, R. Farahbakhsh, and N. Crespi, “BERT-based Ensemble Approaches for Hate Speech Detection.” arXiv, Sep. 15, 2022. Accessed: Sep. 26, 2023. [Online]. Available: http://arxiv.org/abs/2209.06505

[17] “bert-base-multilingual-cased · Hugging Face.” Accessed: Sep. 27, 2023. [Online]. Available: https://huggingface.co/bert-base-multilingual-cased

[18] “Long Short-Term Memory (LSTM) Networks.” Accessed: Sep. 27, 2023. [Online]. Available: https://www.mathworks.com/discovery/lstm.html

[19] A. Kharwal, “Bernoulli Naive Bayes in Machine Learning | Aman Kharwal,” thecleverprogrammer. Accessed: Sep. 27, 2023. [Online]. Available: https://thecleverprogrammer.com/2021/07/27/bernoulli-naive-bayes-in-machine-learning/

[20] G. Singh, B. Kumar, L. Gaur, and A. Tyagi, “Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification,” in *2019 International Conference on Automation, Computational and Technology Management (ICACTM)*, London, United Kingdom: IEEE, Apr. 2019, pp. 593–596. doi: 10.1109/ICACTM.2019.8776800.

[21] J. S. Malik, G. Pang, and A. van den Hengel, “Deep Learning for Hate Speech Detection: A Comparative Study.” arXiv, Feb. 18, 2022. Accessed: Oct. 03, 2023. [Online]. Available: http://arxiv.org/abs/2202.09517

[22] “mapsoriano/2016_2022_hate_speech_filipino · Datasets at Hugging Face.” Accessed: Oct. 08, 2023. [Online]. Available: https://huggingface.co/datasets/mapsoriano/2016_2022_hate_speech_filipino

[23] E. W. McLaughlin, “How to Regulate Online Platforms: Why Common Carrier Doctrine is Inappropriate to Regulate Social Networks and Alternate Approaches to Protect Rights,” vol. 90.

[24] Oup. Editor, “The connection between online hate speech and real-world hate crime,” OUPblog. Accessed: Sep. 27, 2023. [Online]. Available: https://blog.oup.com/2019/10/connection-between-online-hate-speech-real-world-hate-crime/

[25] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, “A survey on ensemble learning,” *Front. Comput. Sci.*, vol. 14, no. 2, pp. 241–258, Apr. 2020, doi: 10.1007/s11704-019-8208-z.

[26] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[27] P. Domingos, “A few useful things to know about machine learning,” *Commun. ACM*, vol. 55, no. 10, pp. 78–87, Oct. 2012, doi: 10.1145/2347736.2347755.

[28] N. Proferes, N. Jones, S. Gilbert, C. Fiesler, and M. Zimmer, “Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics,” *Soc. Media*

Soc., vol. 7, no. 2, p. 205630512110190, Apr. 2021, doi: 10.1177/20563051211019004.

[29] C. Buntain and J. Golbeck, “Identifying social roles in reddit using network structure,” in *Proceedings of the 23rd International Conference on World Wide Web*, Seoul Korea: ACM, Apr. 2014, pp. 615–620. doi: 10.1145/2567948.2579231.

[30] “Tagalog language | Philippines, Austronesian, Dialects | Britannica.” Accessed: Sep. 14, 2023. [Online]. Available: https://www.britannica.com/topic/Tagalog-language

[31] J. D. Lesada, “TAGLISH IN METRO MANILA: AN ANALYSIS OF TAGALOG-ENGLISH CODE-SWITCHING”.

[32] A. Vaswani *et al.*, “Attention Is All You Need.” arXiv, Aug. 01, 2023. Accessed: Aug. 31, 2023. [Online]. Available: http://arxiv.org/abs/1706.03762

[33] J. C. Pereira-Kohatsu, L. Quijano-Sánchez, F. Liberatore, and M. Camacho-Collados, “Detecting and Monitoring Hate Speech in Twitter,” *Sensors*, vol. 19, no. 21, p. 4654, Oct. 2019, doi: 10.3390/s19214654.

[34] R. C. K. Enriquez and M. R. J. E. Estuar, “Determining Linguistic Features of Hate Speech from 2016 Philippine Election-Related Tweets,” in *2023 International Conference on IT Innovation and Knowledge Discovery (ITIKD)*, Manama, Bahrain: IEEE, Mar. 2023, pp. 1–6. doi: 10.1109/ITIKD56332.2023.10100008.

[35] M. R. Karim, S. K. Dey, T. Islam, M. Shajalal, and B. R. Chakravarthi, “Multimodal Hate Speech Detection from Bengali Memes and Texts.” arXiv, Dec. 21, 2022. Accessed: Sep. 03, 2023. [Online]. Available: http://arxiv.org/abs/2204.10196

[36] S. Dowlagar and R. Mamidi, “HASOCOne@FIRE-HASOC2020: Using BERT and Multilingual BERT models for Hate Speech Detection.” arXiv, Jan. 22, 2021. Accessed: Sep. 03, 2023. [Online]. Available: http://arxiv.org/abs/2101.09007

[37] N. Deshpande, N. Farris, and V. Kumar, “Highly Generalizable Models for Multilingual Hate Speech Detection.” arXiv, Jan. 26, 2022. Accessed: Sep. 04, 2023. [Online]. Available: http://arxiv.org/abs/2201.11294

[38] A. McCallum and K. Nigam, “A Comparison of Event Models for Naive Bayes Text Classification”.

[39] M. B. Ressan and R. F. Hassan, “Naïve-Bayes family for sentiment analysis during COVID-19 pandemic and classification tweets,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 28, no. 1, p. 375, Oct. 2022, doi: 10.11591/ijeecs.v28.i1.pp375-383.

[40] R. Khezzar, A. Moursi, and Z. Al Aghbari, “arHateDetector: detection of hate speech from standard and dialectal Arabic Tweets,” *Discov. Internet Things*, vol. 3, no. 1, p. 1, Mar. 2023, doi: 10.1007/s43926-023-00030-9.

[41] K. K. Kiilu, G. Okeyo, R. Rimiru, and K. Ogada, “Using Naïve Bayes Algorithm in detection of Hate Tweets,” *Int. J. Sci. Res. Publ. IJSRP*, vol. 8, no. 3, Mar. 2018, doi: 10.29322/IJSRP.8.3.2018.p7517.

[42] J. D. Yati and E. W. Pamungkas, “HATE SPEECH DETECTION ON SOCIAL MEDIA CONTENT IN JAVANESE LANGUAGE WITH NAÏVE BAYES ALGORITHM”.

[43] N. L. Lavenia and R. Permatasari, “Sentiment Analysis on Twitter Social Media Regarding Depression Disorder Using the Naive Bayes Method,” *CoreID J.*, vol. 1, no. 2, pp. 66–74, Jul. 2023, doi: 10.60005/coreid.v1i2.14.

[44] M. S. Akter, H. Shahriar, and A. Cuzzocrea, “A Trustable LSTM-Autoencoder Network for Cyberbullying Detection on Social Media Using Synthetic Data”.

[45] A. K. Das, A. Al Asif, A. Paul, and Md. N. Hossain, “Bangla hate speech detection on social media using attention-based recurrent neural network,” *J. Intell. Syst.*, vol. 30, no. 1, pp. 578–591, Apr. 2021, doi: 10.1515/jisys-2020-0060.

[46] M. S. Jahan and M. Oussalah, “A systematic review of hate speech automatic detection using natural language processing,” *Neurocomputing*, vol. 546, p. 126232, Aug. 2023, doi: 10.1016/j.neucom.2023.126232.

[47] G. L. De la Peña Sarracén, R. G. Pons, C. E. Muñiz Cuza, and P. Rosso, “Hate Speech Detection using Attention-based LSTM,” in *EVALITA Evaluation of NLP and Speech Tools for Italian*, T. Caselli, N. Novielli, V. Patti, and P. Rosso, Eds., Accademia University Press, 2018, pp. 235–238. doi: 10.4000/books.aaccademia.4784.

[48] G. Manias, A. Mavrogiorgou, A. Kiourtis, C. Symvoulidis, and D. Kyriazis, “Multilingual text categorization and sentiment analysis: a comparative analysis of the utilization of multilingual approaches for classifying twitter data,” *Neural Comput. Appl.*, vol. 35, no. 29, pp. 21415–21431, Oct. 2023, doi: 10.1007/s00521-023-08629-3.

[49] T. Li and K. Murray, “Why Does Zero-Shot Cross-Lingual Generation Fail? An Explanation and a Solution,” in *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada: Association for Computational Linguistics, 2023, pp. 12461–12476. doi: 10.18653/v1/2023.findings-acl.789.

[50] D. Nozza, “Exposing the limits of Zero-shot Cross-lingual Hate Speech Detection,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Online: Association for Computational Linguistics, 2021, pp. 907–914. doi: 10.18653/v1/2021.acl-short.114.

[51] T. Caselli, V. Basile, J. Mitrović, and M. Granitzer, “HateBERT: Retraining BERT for Abusive Language Detection in English,” in *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, Online: Association for Computational Linguistics, 2021, pp. 17–25. doi: 10.18653/v1/2021.woah-1.3.

[52] H. Saleh, A. Alhothali, and K. Moria, “Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model,” *Appl. Artif. Intell.*, vol. 37, no. 1, p. 2166719, Dec. 2023, doi: 10.1080/08839514.2023.2166719.

[53] M. Ibanez, R. Sapinit, L. A. Reyes, M. Hussien, J. M. Imperial, and R. Rodriguez, “Audio-Based Hate Speech Classification from Online Short-Form Videos,” in *2021 International Conference on Asian Language Processing (IALP)*, Singapore, Singapore: IEEE, Dec. 2021, pp. 72–77. doi: 10.1109/IALP54817.2021.9675250.

[54] M. K. A. Aljero and N. Dimililer, “A Novel Stacked Ensemble for Hate Speech Recognition,” 2021.

[55] Y. J. Nurriski and T. F. Agustina, “Stacking Ensemble Learning Technique for Sentiment Analysis of Hate Speech on Twitter,” vol. 1, no. 1.

[56] I. Markov, I. Gevers, and W. Daelemans, “An Ensemble Approach for Dutch Cross-Domain Hate Speech Detection,” in *Natural Language Processing and Information Systems*, vol. 13286, P. Rosso, V. Basile, R. Martínez, E. Métais, and F.

Meziane, Eds., in Lecture Notes in Computer Science, vol. 13286. , Cham: Springer International Publishing, 2022, pp. 3–15. doi: 10.1007/978-3-031-08473-7_1.

[57] S. Cui, Y. Han, Y. Duan, Y. Li, S. Zhu, and C. Song, “A Two-Stage Voting-Boosting Technique for Ensemble Learning in Social Network Sentiment Classification,” *Entropy*, vol. 25, no. 4, p. 555, Mar. 2023, doi: 10.3390/e25040555.

[58] C. N. Arbaatun, D. Nurjanah, and H. Nurrahmi, “Hate Speech Detection on Twitter through Natural Language Processing using LSTM Model,” *Build. Inform. Technol. Sci. BITS*, vol. 4, no. 3, Dec. 2022, doi: 10.47065/bits.v4i3.2718.

[59] K. Mnassri, P. Rajapaksha, R. Farahbakhsh, and N. Crespi, “Hate Speech and Offensive Language Detection using an Emotion-aware Shared Encoder.” arXiv, Feb. 17, 2023. Accessed: Oct. 21, 2023. [Online]. Available: http://arxiv.org/abs/2302.08777

[60] W. Scott, “TF-IDF for Document Ranking from scratch in python on real world dataset.,” Medium. Accessed: Oct. 21, 2023. [Online]. Available: https://towardsdatascience.com/tf-idf-for-document-ranking-from-scratch-in-python-on-real-world-dataset-796d339a4089

[61]Bhartendoo Vimal and RV College of Engineering, “Application of Logistic Regression in Natural Language Processing,” *IJERT*, vol. V9, no. 06, p. IJERTV9IS060095, Jun. 2020, doi: 10.17577/IJERTV9IS060095.

[62] A. Mehmood et al., “Threatening URDU Language Detection from Tweets Using Machine Learning,” Applied Sciences, vol. 12, no. 20, p. 10342, Oct. 2022, doi: 10.3390/app122010342.

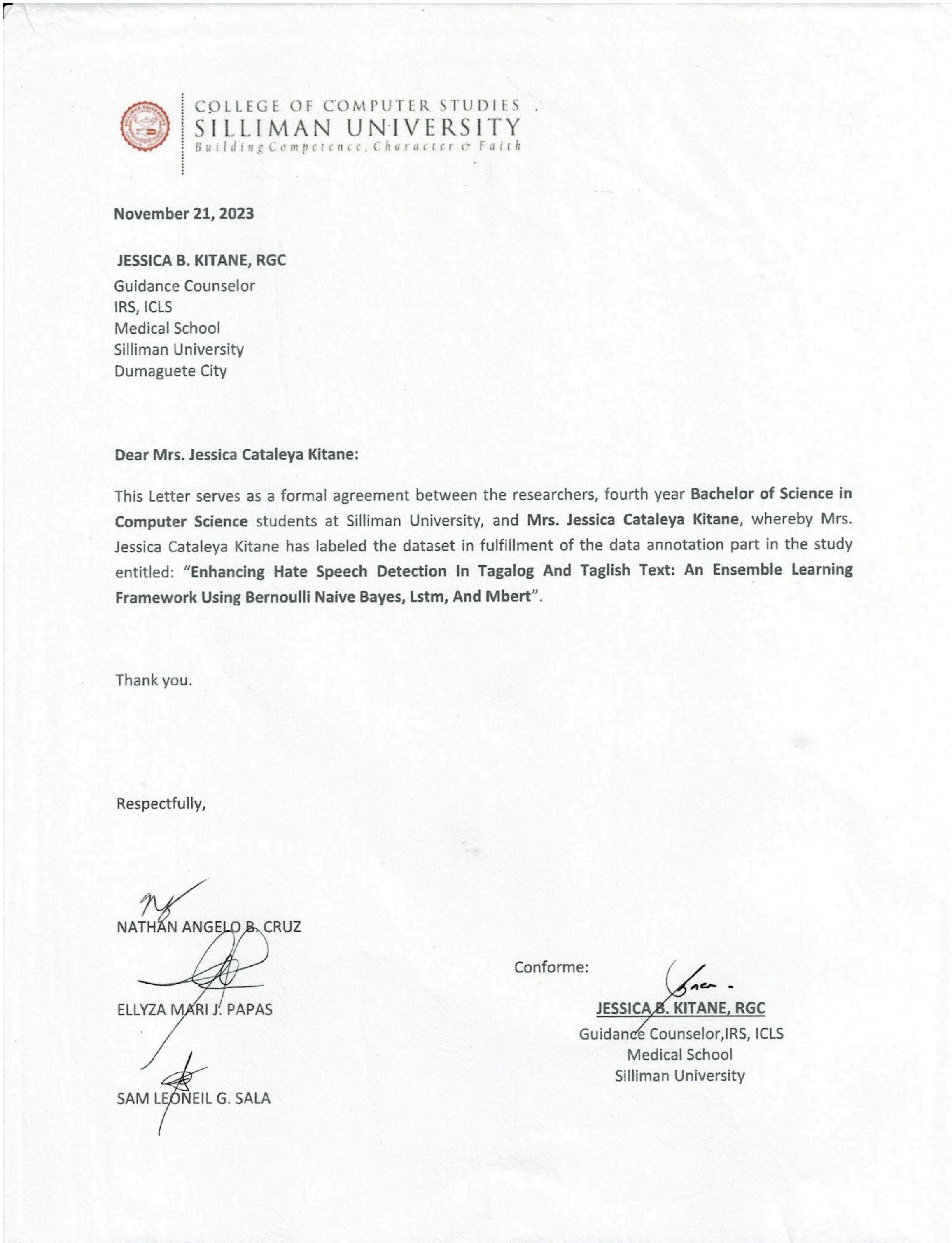
[63] S. Prusty, S. Patnaik, and S. K. Dash, “SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer,” *Front. Nanotechnol.*, vol. 4, p. 972421, Aug. 2022, doi: 10.3389/fnano.2022.972421.

[64]“pinoy subreddits,” Reddit. [Online]. Available: https://www.reddit.com/user/the_yaya/m/pinoy/

[65] L. J. V. Miranda, “calamanCy: A Tagalog Natural Language Processing Toolkit.” arXiv, Nov. 13, 2023. [Online]. Available: http://arxiv.org/abs/2311.07171

APPENDICES

APPENDIX A: LETTER OF AGREEMENT



COLLEGE OF COMPUTER STUDIES
SILLIMAN UNIVERSITY
Building Competence, Character & Faith

November 21, 2023

JESSICA B. KITANE, RGC
Guidance Counselor
IRS, ICLS
Medical School
Silliman University
Dumaguete City

Dear Mrs. Jessica Cataleya Kitane:

This Letter serves as a formal agreement between the researchers, fourth year **Bachelor of Science in Computer Science** students at Silliman University, and **Mrs. Jessica Cataleya Kitane**, whereby Mrs. Jessica Cataleya Kitane has labeled the dataset in fulfillment of the data annotation part in the study entitled: “**Enhancing Hate Speech Detection In Tagalog And Taglish Text: An Ensemble Learning Framework Using Bernoulli Naive Bayes, Lstm, And Mbert**”.

Thank you.

Respectfully,

NATHAN ANGELO B. CRUZ

ELLYZA MARI J. PAPAS

SAM LEONEIL G. SALA

Conforme:

JESSICA B. KITANE, RGC
Guidance Counselor,IRS, ICLS
Medical School
Silliman University

APPENDIX B: LIST OF FILIPINO SUBREDDITS FOR DATA COLLECTION

Subreddit	Name	Subscribers as of October 2023
/r/ADMU	Ateneo de Manila University	46,221
/r/CasualPH	Casual PH	193,206
/r/dlsu	De La Salle University	36,288
/r/LawPH	Law PH	24,952
/r/LeopardsAteMyFacePH	Leopards Ate My Face PH	10,387
/r/MentalHealthPH	Mental Health PH	35,519
/r/OffMyChestPH	Off My Chest PH	297,322
/r/peyups	University of the Philippines	77,637
/r/phlgbt	PH LGBT	13,928
/r/Philippines	Philippines	1,241,653
/r/relationship_advicePH	Relationship Advice PH	82,565
/r/studentsph	Students PH	134,747
/r/Tomasino	University of Santo Tomas	53,767