Chapter 3:

Research Design

1. **Research objectives**
2. **Research Methodology**
    - **2.1 Data collection method**
    - **2.2 Data analyzing techniques**
    - **2.3 Model Evaluation**
    - **2.4 Hypothesis**
        - **2.4.1 Contextual Hypothesis**
        - **2.4.2 Model-based Hypothesis**
    - **2.5 Ethical considerations**

1. Research Objectives

1. Create an accurate data collection/data set for hate speech detection in bi-lingual language Singlish (Sinhala + English)

   To make an effective model data is required for any further workings. English and Sinhala as separate languages, have multiple sources to get data sets for hate speech. As per the subject of the study, Singlish is a bilingual language format that is written in English letters to provide the Sinhala meaning terms and as a new-found language it is lacking with the data to work with. To get along with the study it is required to have a sufficient dataset including none-hate and hate speech content to lay the foundation for this study and also for the future researches in the same area.

2. Discover data preprocessing techniques

   Data pre-processing is the major process of data classification and language-based detection systems. In the pre processing identifies the language, text normalization and tokenization of the text input.

   - Language Identification – identifies the language of the text input.
   - Text normalization- removes the noises, punctuations and converts to a uniform lowercase term and standardizes the given text input
   - Tokenization – removes the stop words such as “the”, “a”, “is”, and “am”, etc and creates individual tokens for each and every word in the given statement.
   - Vectorizing – which makes the tokenized words into machine-understandable language for machine learning purposes.

3. Compare and contrast the existing machine learning models to optimize the gap between hate speech detection systems

   Within the previous chapter and conducting the pre-study related to hate speech detection have given the most efficient and effective solutions and findings relevant to the subject. By inheriting and getting inspiration from the existing studies discoveries helps to develop a better solution where the author is able to identify the potential threats and risk along with the opportunities that creates an advantage to make a better, optimized, effective, and efficient bilingual hate speech detection system.

4. Identify limitations and strengths to fine-tune the language model

Using traditional techniques/algorithms to process with novel techniques such as deep learning approaches are able to evaluate one another to find strengths and weaknesses in between.

By inheriting the strengths and omitting the weaknesses to the existing model will help to create a novel solution with better accuracy.

Pre-trained modes can capture more accurate linguistic features on models that are currently available and fine-tuned considering the limitations and the strengths.

5. Evaluating the developed novel model

   Precision, Accuracy, and recall are the three main matrices to evaluate and test against the model to identify the limits of the trained model. Effectiveness of the system and fulfills the initial requirement of identifying hate speech using bilingual language.

2. Research Methodology

2.1. Data Collection Method

In this study, a quantitative approach will be used to implement on detection of hate speech in Singlish. The primary source of data gathering will be extracting hateful and non-hateful text, text-based content which are posted in Singlish from potential social media platforms such as Facebook, x, YouTube, TikTok, etc. By following this approach will allow to get natural expression through extraction and different forms of the terms that are used to interpret the same idea. As the secondary method of collecting data is getting the existing dataset which has been taken into making the previous studies under the same domain area. This approach will bridge the gaps between both primary and secondary approaches to create a robust and more effective machine-learning model to detect accurate hateful content and will create a sufficient and up-to-date data set for further enhancements.

2.2. Data analyzing techniques

Considering the linguistic landscape of the study, to analyze the collected data to generate an accurate outcome number of analyzing steps are needed to be followed within the model.

At first, collected data are needed to be pre-processed which is the process of cleaning the data before analyzing. Therefore eliminating noises,

removing stop words, hashtags, special characters, URLs, symbols emojis, numerical values, and symbols from the data. And making them uniform data by making each and every character to lowercase. After normalizing the uniformed terms are needed to run through a process of tokenization where it splits the text slack into individual word tokens. BERT will be the most relevant technique to use to split and tokenize the text. Tokenized texts are more humanized and to make them machine-understandable, tokenized data are required to vectorize. Which is a process that converts the tokenized text into numerical values which are machine understandable for further operations. Techniques such as mBERT. and BERT for contextual embedding and Word2Vec, and GloVe for word embedding are being used. Vectorized data are now can be analyzed to generate accurate output running through the model.

2.3. Model evaluation

The model itself contains three different data models such as Logistic regression which performs well in binary classification contexts, and support vector machines model which is more effective in classifications and is multi-dimensional. Both models are known to be traditional machine learning models. mBERT which is a deep-learning model that is capable of understanding natural language and has the ability to handle linguistic features. All three models can be examined and evaluated as a whole. Key performance matrices will be accuracy, precision, sensitivity and result validation matrix.

Accuracy – correctness of the given outcome from the model,

Precision – the ratio of the number of correct outcomes out of all the positives,

Sensitivity – the ratio of the number of correct outcomes out of all the observations,

Result validation matrix – matrix that depicts true positive, true negatives, false positives, and false negative.

Model evaluation of a bilingual model is challenging due to the advance features of the text classification parameters itself.

2.4. Hypothesis

2.4.1. Contextual hypothesis

Hypothesis 1 – Implementing the hate speech detection tool will ensure the safety of Sri Lankan Facebook users.

Hypothesis 2 – The Hate speech detection tool will increase the number of users on Facebook due to the safe environment of the platform.

Hypothesis 3 – The Hate speech detection tool might fail to detect both languages and will not be able to achieve the study objectives.

2.4.2. Model-based hypothesis

Hypothesis 4 – Preprocessing texts consumes more time when generating an outcome

Hypothesis 5 – Preprocessing doesn't affect the result accuracy.

Hypothesis 6 – language-specific pre-processing enhances the model performances

2.5. Ethical considerations

Data Privacy

Privacy of the collected data will be upheld with the author and no user, user name, or person identification mechanism wasn't used or is not being used. Only the statements that were posted to the social media platforms were taken to examine the model.

Bias and Fairness

Three models identified contexts in three different perspectives are used to get multiple perspectives and get a better view of the posted content. This will increase the fairness of the data classification and will be unbiased when making the decisions. This will affect to accuracy and precision as well.

Impact on Freedom of Expression

Hate speech is bounded by the freedom of expressions. Balance is required to identify the contextual aspects of criticism and hateful contexts. Classification may affect to the freedom of expression by identifying a valid criticism as a hateful context. By using the necessary parameters model can be recalibrated to balance the ratio of freedom of speech and as well as the hate speech

```python
from transformers import BertTokenizer, BertForSequenceClassification, Trainer, TrainingArguments
import torch

# Load tokenizer and model
tokenizer = BertTokenizer.from_pretrained('bert-base-multilingual-cased')
model = BertForSequenceClassification.from_pretrained('bert-base-multilingual-cased', num_labels=2)

# Tokenize data
train_texts = ["your train texts here"]
train_labels = [0, 1]  # 0 for non-hate speech, 1 for hate speech
val_texts = ["your validation texts here"]
val_labels = [0, 1]

train_encodings = tokenizer(train_texts, truncation=True, padding=True, max_length=128)
val_encodings = tokenizer(val_texts, truncation=True, padding=True, max_length=128)

# Create dataset class
class HateSpeechDataset(torch.utils.data.Dataset):
    def __init__(self, encodings, labels):
        self.encodings = encodings
        self.labels = labels

    def __getitem__(self, idx):
        item = {key: torch.tensor(val[idx]) for key, val in self.encodings.items()}
        item['labels'] = torch.tensor(self.labels[idx])
        return item

    def __len__(self):
        return len(self.labels)

train_dataset = HateSpeechDataset(train_encodings, train_labels)
val_dataset = HateSpeechDataset(val_encodings, val_labels)

# Training arguments
training_args = TrainingArguments(
    output_dir='./results',
    num_train_epochs=3,
    per_device_train_batch_size=8,
    per_device_eval_batch_size=8,
    warmup_steps=500,
    weight_decay=0.01,
    logging_dir='./logs',
    logging_steps=10,
    evaluation_strategy="epoch",
)

# Trainer
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=train_dataset,
    eval_dataset=val_dataset,
)

# Train model
```