PROJECT PROPOSAL

Hate Speech Detection

by San Myint Hlaing & Saranya S.

Table of Contents

- Project Objectives
- Dataset Selection
- Data Preprocessing
- Model Architecture
- Training and Evaluation
- Challenges

Project Objectives

- Content Moderation: Use the AI to automatically detect and remove hate speech and offensive content from online platforms to create a safer and more respectful environment.
- Education and Awareness: Use the AI to flag hate speech and provide educational resources or warnings to users, encouraging responsible online behavior.

Project Objectives

- Content Moderation: Use the AI to automatically detect and remove hate speech and offensive content from online platforms to create a safer and more respectful environment.

- Education and Awareness: Use the AI to flag hate speech and provide educational resources or warnings to users, encouraging responsible online behavior.

Dataset Collection

- Kaggle

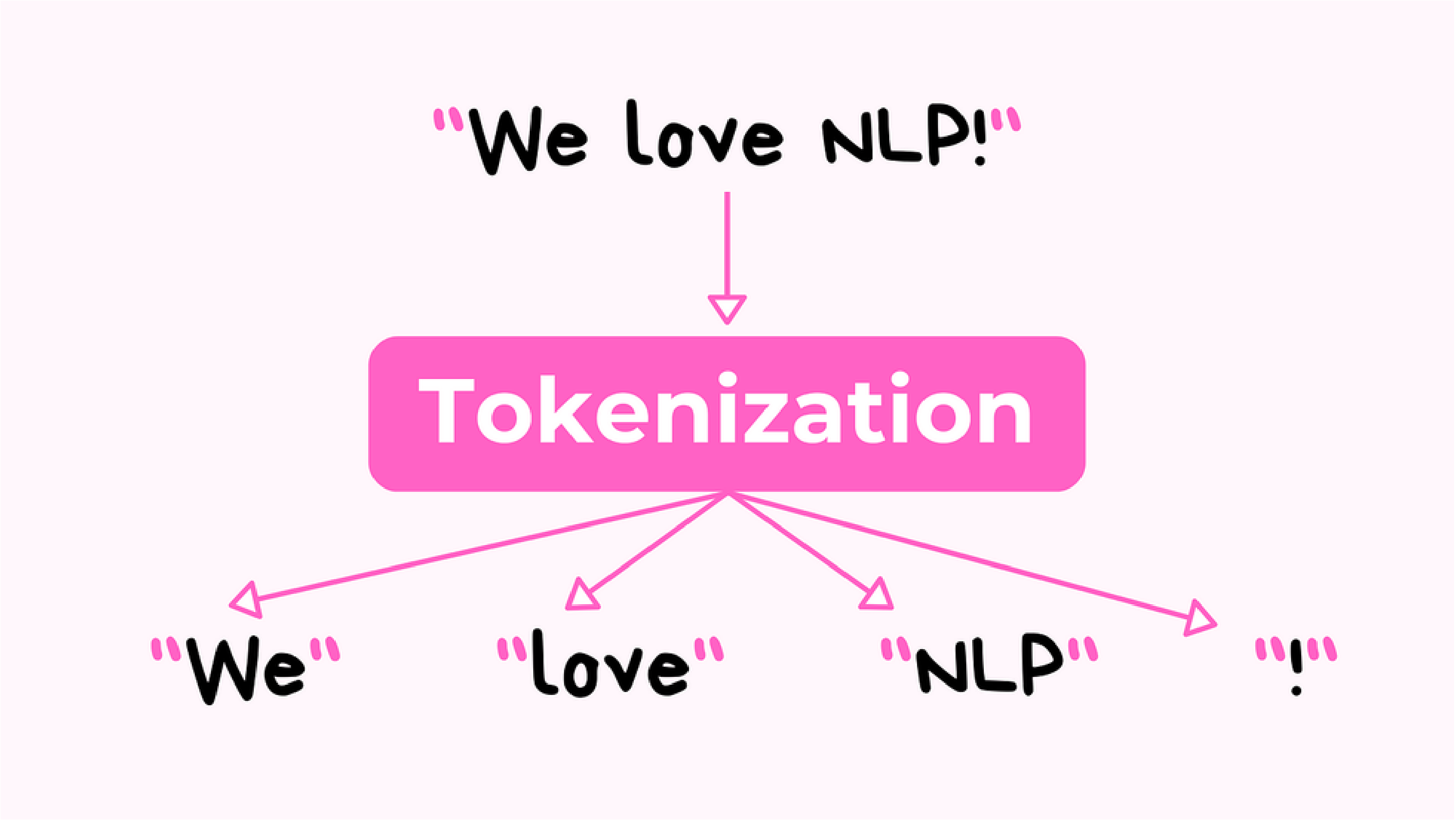
Data Preprocessing

Text Cleaning

Remove special characters, emojis, and irrelevant white spaces from the text data.

Data Preprocessing

Tokenization

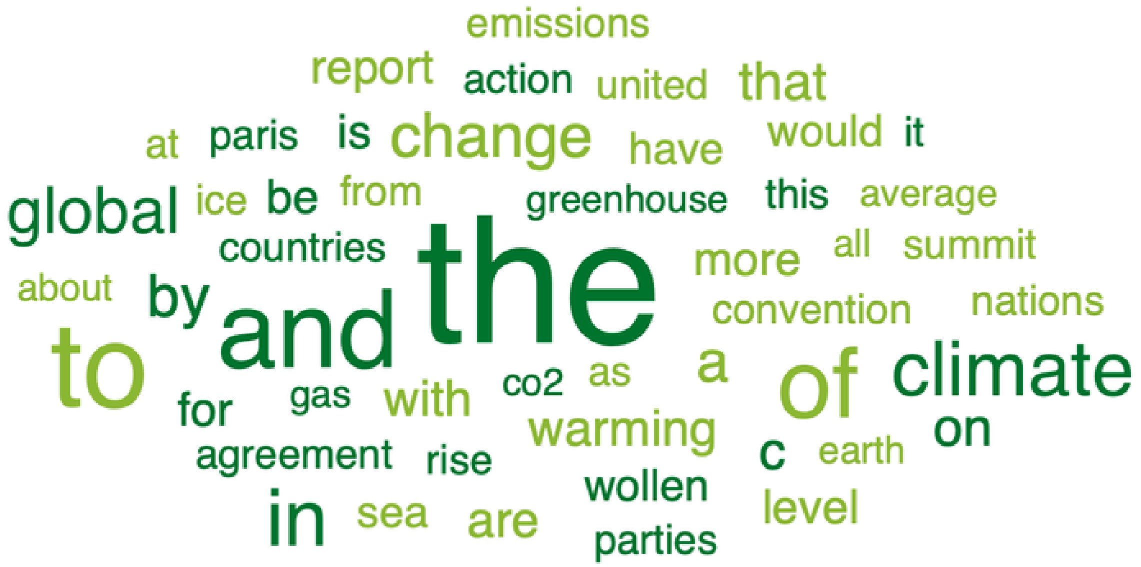


techniques: word tokenization or Byte-Pair Encoding (BPE)

Data Preprocessing

Stopword Removal

Eliminate common stopwords and low-information words from the text to reduce noise.

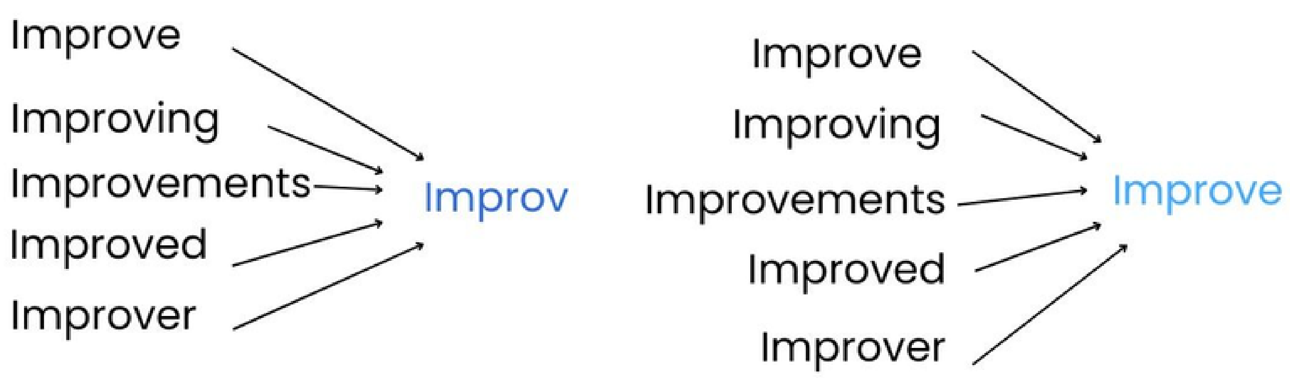


Data Preprocessing

Text Normalization

Apply text normalization techniques such as lemmatization or stemming to standardize text forms.

Stemming vs Lemmatization



Python
for Unix and Linux System Administration
Noah Gift & Jeremy M. Jones
O'REILLY

Model Architecture

BERT

A pre-trained language model that is fine-tuned for the specific task of hate speech detection. BERT is used to convert the text data into numerical representations that can be fed into a deep learning model for classification.

```
tokenizer = BertTokenizer.from_pretrained('bert-base-uncased')
```

Training and Evaluation

Training

- Data Collection

- Data Labeling

- Data Preprocessing

Validation

Utilize cross-validation to assess the model's generalization performance effectively.

Evaluation Metrics

- accuracy
- precision
- recall
- F1-score

Challenges

Data Quality and Bias

Ensuring unbiased and representative training data.

False Positives and Negatives

Balancing accuracy while minimizing both types of errors.

User Acceptance

Gaining user trust and acceptance of AI-based content moderation.

PROJECT PROPOSAL

THANK YOU

by San Myint Hlaing & Saranya S.