IST 736 Text Mining
Project Proposal

Analyzing Hate Speech Dynamics on Twitter Using Text Mining and NLP Techniques

Team Members

Huzaif Kherani

M K Sowmeya

Gughapriya Elango

Harshita Tanksali

PROJECT GOAL

- The goal of this project is to leverage text mining and natural language processing techniques to analyze hate speech in Twitter data, with the aim of understanding its prevalence, characteristics, and impact.
- By achieving this goal, we aim to contribute to a better understanding of hate speech dynamics on the platform and potentially develop or improve automated tools for hate speech detection and moderation, while considering the ethical implications of such technology.

PROJECT SCOPE

Hate Speech Detection

Train models to classify tweets into categories (hate, offensive and neither). Can be used for content moderation.

Fine- Grained Sentiment Analysis

Perform sentiment analysis with a more fine-grained approach. Instead of just classifying tweets as P, N, Neu, we can differentiate between hate speech, offensive language and non-offensive sentiments.

Language trend analysis

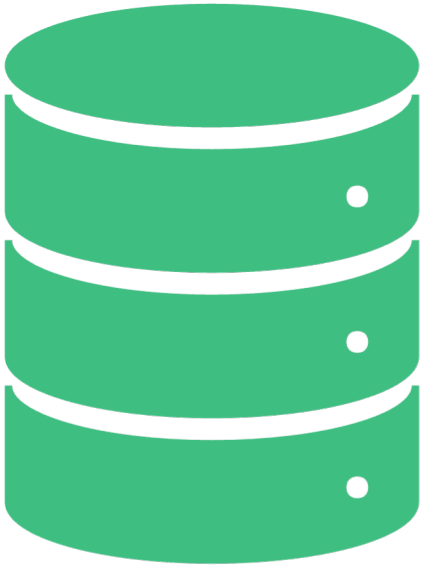
- Analyze to identify trends in the use of hate speech or offensive language over time.

Contextual Analysis

- Capture essential linguistic elements of each tweet which helps distill the significant words and their base forms , creating a representation of essential linguistic components within each tweet.

DATASET DESCRIPTION

hate speech	neither	offensive language	Class	Tweet
•Number of crowdFlower (CF) users who judged the speech to be hate.	•Number of CF users who judged the tweet to be neither offensive nor nor-offensive	•Number of CF users who judged the tweet to be offensive	•Class label for majority of CF users. 0 - hate speech 1 - offensive language 2 - neither	•Contains the text tweet



RESEARCH QUESTIONS

What are the prevailing themes and patterns of hate speech in Twitter data

How accurately can we classify hate speech on Twitter, and what features or techniques contribute to their performance?"

Are there specific linguistic or semantic characteristics that are more commonly associated with hate speech in Twitter data?

What are the most effective strategies for hate speech detection and moderation on Twitter, and how do they compare in terms of accuracy and efficiency?

DATA EXPLORATION

```
missing=df.isnull().sum()
print(missing)
```

```
Unnamed: 0              0
count                   0
hate_speech             0
offensive_language      0
neither                 0
class                   0
tweet                   0
dtype: int64
```

Total: 24783
hate: 1430 (5.77% of total)

Total: 24783
Ofensive: 19190 (77.43% of total)

Total: 24783
Neither: 4163 (16.80% of total)

```
df['hate_speech'].value_counts()
```

```
0    19790
1     3419
2     1251
3      287
4       21
5        7
6        5
7        3
Name: hate_speech, dtype: int64
```

```
df['neither'].value_counts()
```

```
0    18892
3     2790
1     1694
2     1200
6      103
5       54
4       35
9        5
8        5
7        5
Name: neither, dtype: int64
```

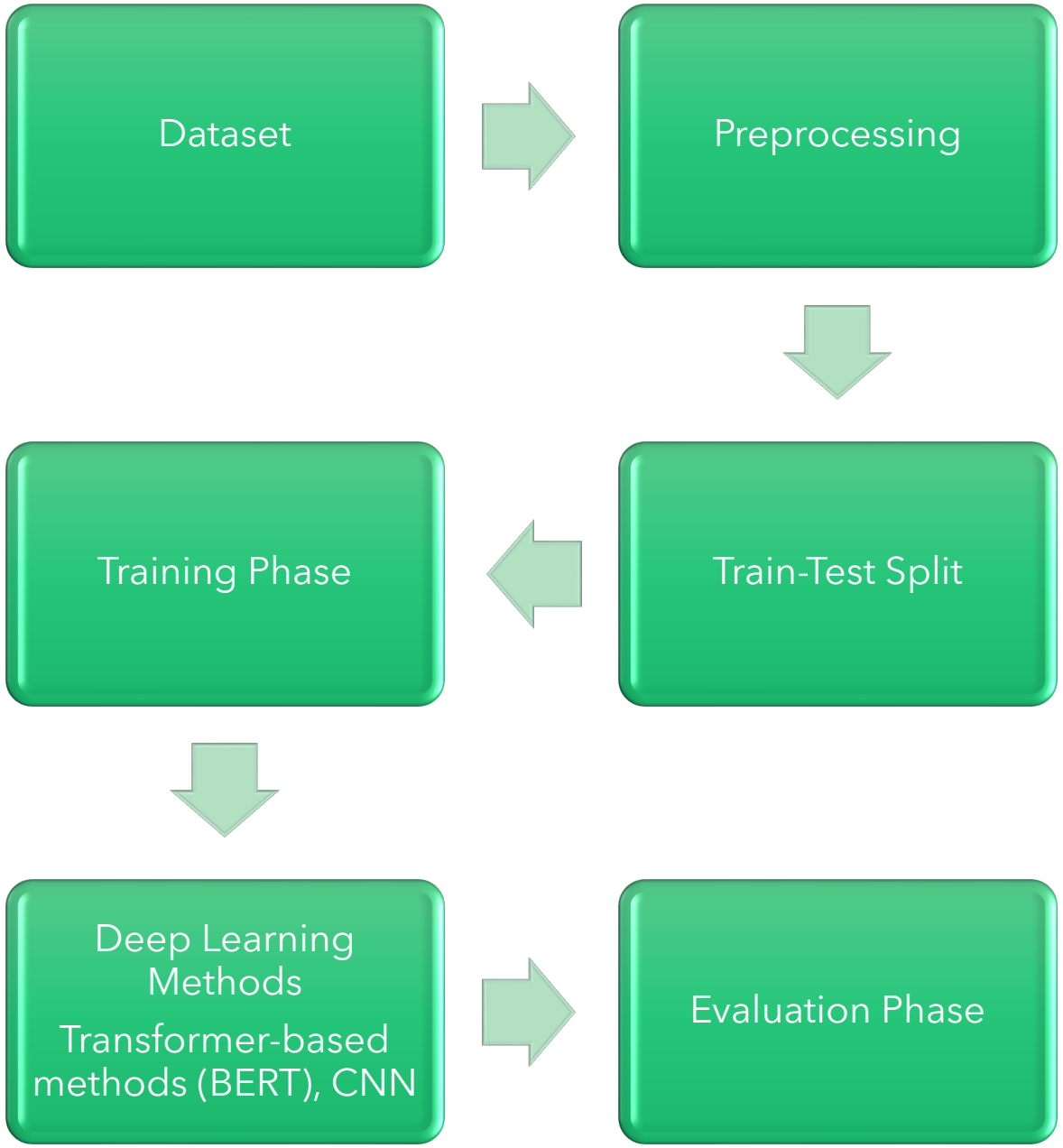
```
df['offensive_language'].value_counts()
```

```
3    13383
2     4246
0     3475
1     2066
6      857
5      369
4      251
9       66
8       37
7       33
Name: offensive_language, dtype: int64
```

```
df['class'].value_counts()
```

```
1    19190
2     4163
0     1430
Name: class, dtype: int64
```

METHODOLOGY



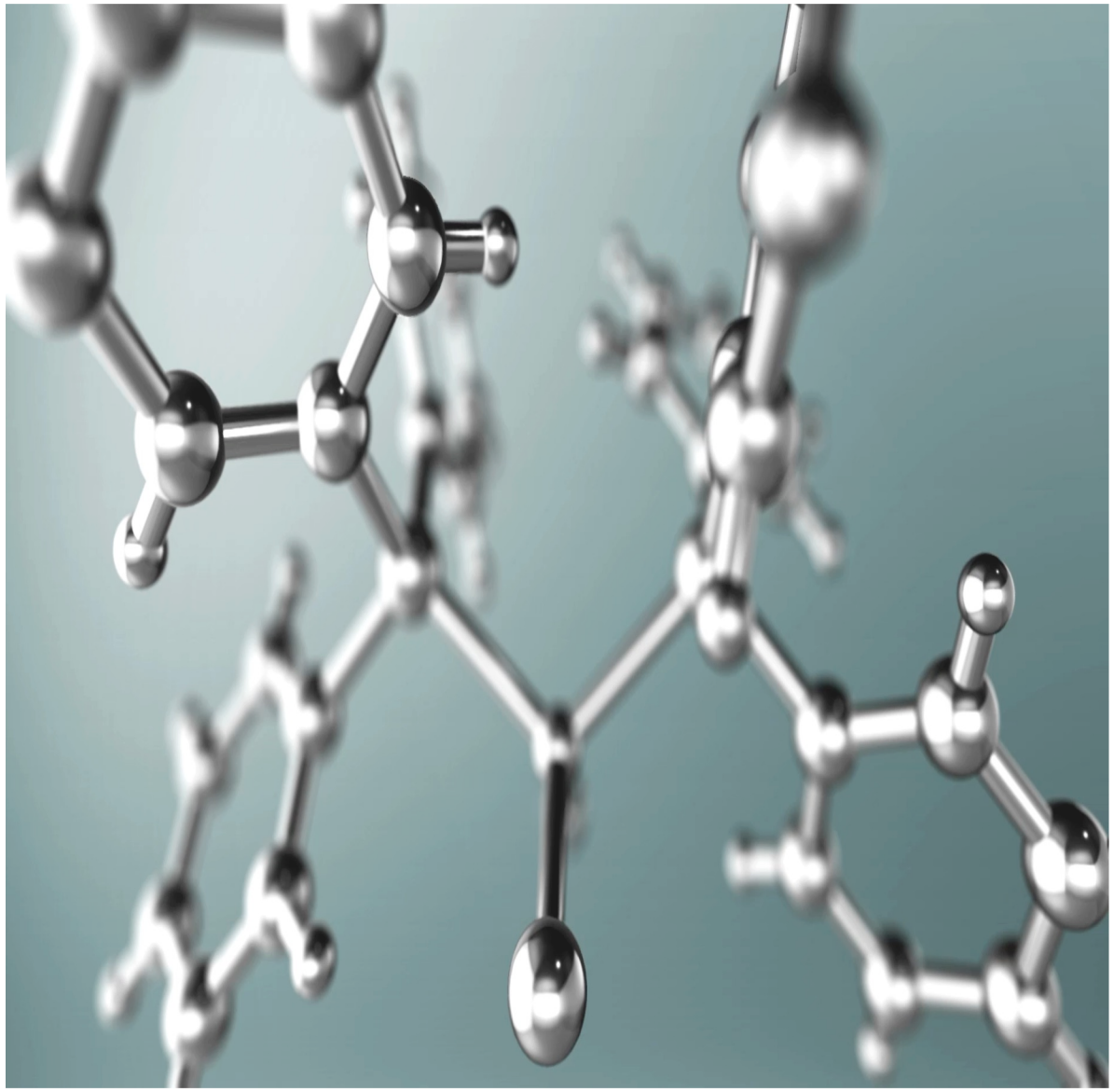
ALGORITHMS

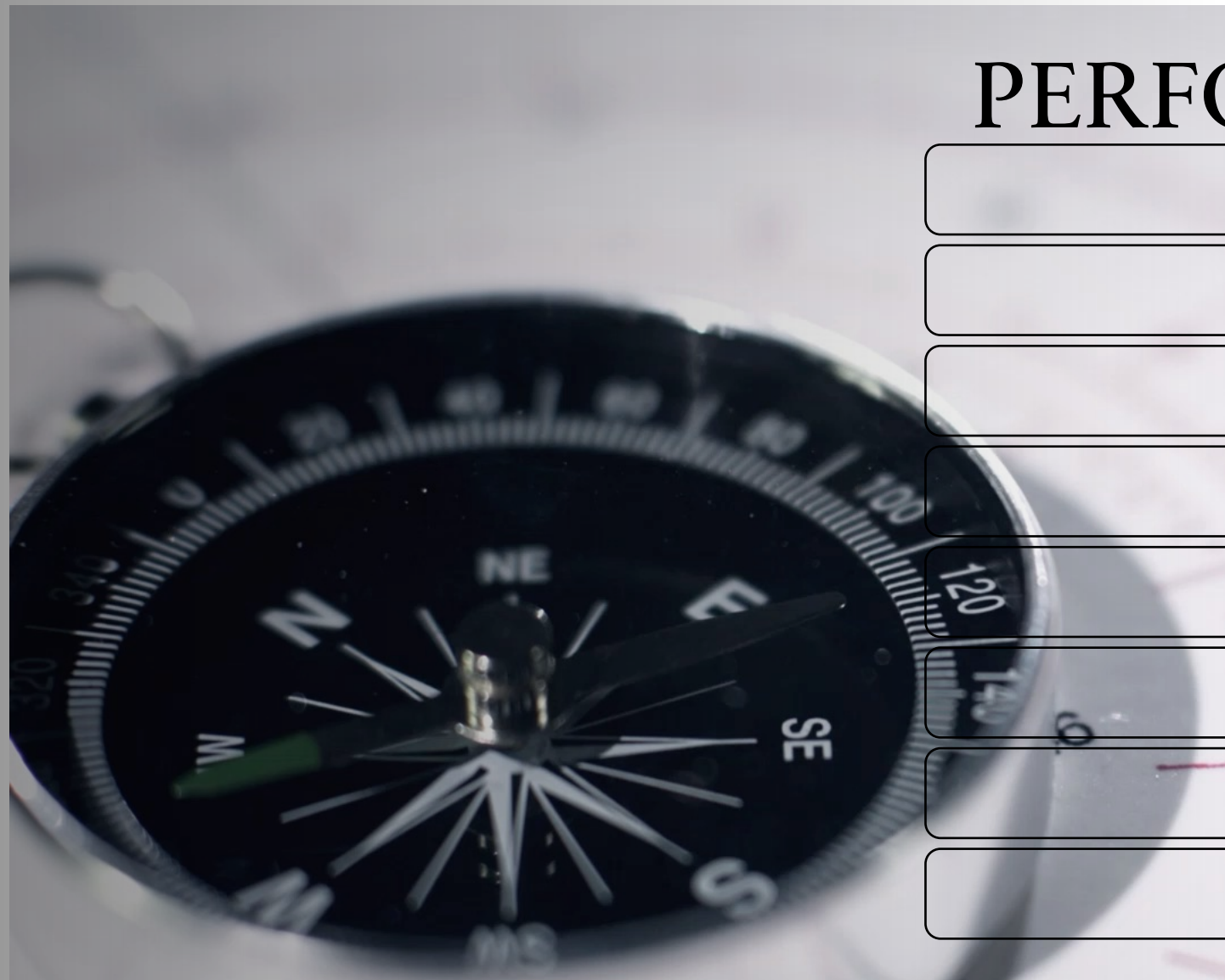
For Classification

- PyTorch Tranformer
- BERT
- Multi channel CNN + Glove
- LSTM

For Hate speech recognition (Prediction)

- Decision Tree
- Naive Bayes
- logistic regression
- KNN
- SVM
- LGBM





PERFORMANCE METRICS

- Accuracy
- Precision
- Recall
- F1 Score
- ROC-AUC
- Confusion Matrix
- False Positive Rate
- Cross Validation

Potential Technical Challeneges

DATA IMBALANCE - CAN IMPACT MODELS ABILITY TO GENERALIZE WELL, PARTICULARLY IF ONE CLASS IS UNDERREPRESENTED

NOISY DATA - PRESENCE OF SPECIAL CHARACTERS, MENTIONS AND NON-STANDARD LANGUAGES IN TWEETS COULD BE A POTENTIAL CHALLENGE TO CLEAN WHILE MAINTAINING ESSENTIAL INFORMATION

MISCLASSIFICATION AND FALSE POSITIVES: MODELS MAY GENERATE FALSE POSITIVES, INCORRECTLY IDENTIFYING NON-OFFENSIVE CONTENT AS HATE SPEECH. THIS CAN LEAD TO UNJUST CONSEQUENCES FOR USERS WHO ARE WRONGLY FLAGGED.

THANK YOU