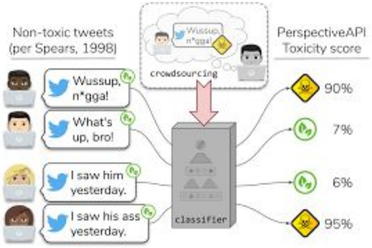
Detecting Hate Speech with BERT model

AGG IES DO

Abigail Yohannes
5/2/24

Hate Speech and Online Extremism

- With the continued rise of online interaction, the risk of hate speech and extremism poses safety risks for users and causes concern for companies and government agencies
- New techniques for masking hate speech by users can make detection difficult
- Lack of understanding for community linguistics can lead to unfair detection



Project Research Problem

- Detecting hate speech from social media platforms using pre-trained Bert model for sequence classification with fine-tuning for text classification of tweets.

index	count	hate_speech	offensive_language	neither	class	tweet
0	3	0	0	3	2	!!! RT @mayasolovely: As a woman you shouldn't...
1	3	0	3	0	1	!!!!! RT @mleew17: boy dats cold...tyga dwn ba...
2	3	0	3	0	1	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby...
3	3	0	2	1	1	!!!!!!!!! RT @C_G_Anderson: @viva_based she lo...
4	6	0	6	0	1	!!!!!!!!!!!!! RT @ShenikaRoberts: The shit you...

- **Dataset: "Hate Speech and Offensive Language Dataset" (Sourced from Kaggle)**
- **Text classification:** **Hate-speech = 0 , Offensive language =1, and Neither 2**
- **Data includes 24,783 labeled tweets**
- **Dataset also includes columns for hate speech/offensive language/neither ratings (classified best ⅔ by various teams)**

Dataset outlook

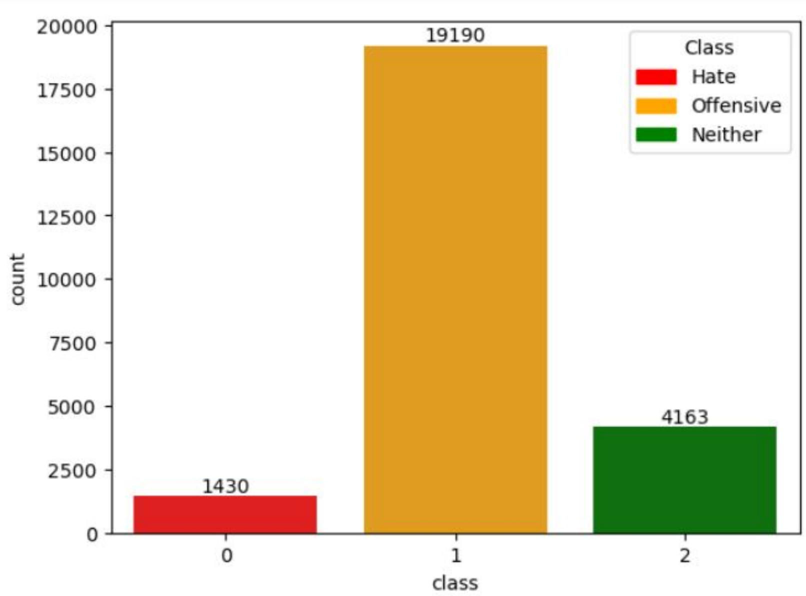


Fig. 1 Dataset

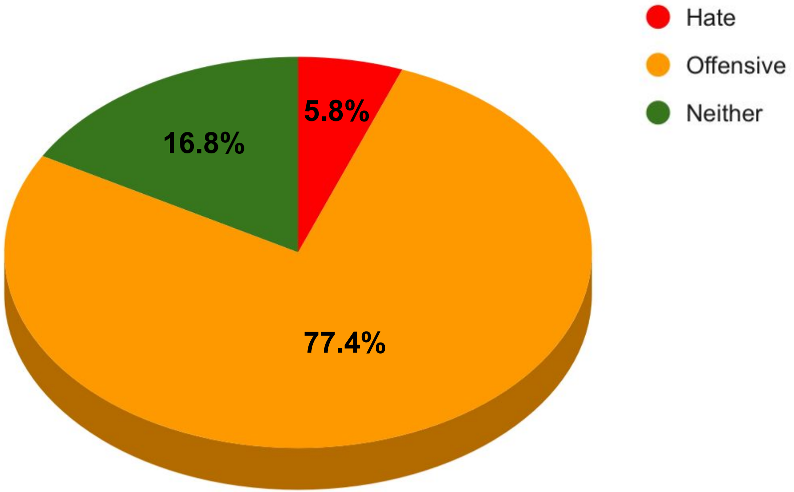


Fig. 2 Text Classification Distribution by Percentage (%)

Research Article

Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model

Hind Saleh, Areej Alhothali & Kawthar Moria

Article: 2166719 | Received 26 Jun 2022, Accepted 04 Jan 2023, Published online: 02 Feb 2023

Cite this article https://doi.org/10.1080/08839514.2023.2166719 Check for updates

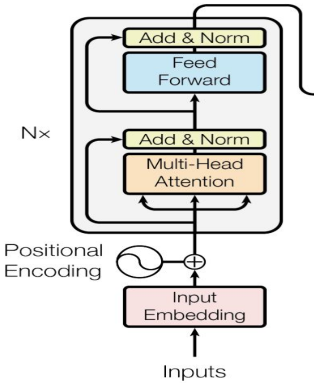
Full Article Figures & data References Citations Metrics Licensing Reprints & Permissions

View PDF View EPUB

Relatec

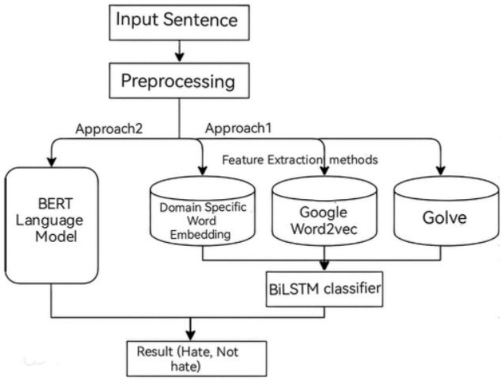
Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model

- **Objective: With increased demand for detecting hate speech on social media platforms, this study encourages the use of NLP to reduce the impact hate speech and hate content has on social media users**
- **The study proposes the use of two models to detect hate speech**



Transformer encoder architecture

Figure 1. Block diagram of the experiments.



- **Approach 1: Domain-Specific Embedding Features with BiLstm Based Deep Model Classifier**
- **Approach 2: Pre-Trained Bert Model**

Paper Dataset

Table 1. Datasets description.

Dataset	Original labels	Number of hate	Number of non-hate	Total
Davidson-ICWSM (Davidson et al. 2017)	Hate speech, offensive, and neither	20620	4163	24783
Waseem-EMNLP (Waseem 2016)	Racism, sexism, both, and neither	1059	5850	6909
Waseem-NAACL (Waseem and Hovy 2016)	Sexist, racist, and neither	5406	11501	16910
Balanced Combined	Racism, sexism, and neither, both	16260	16260	32520

- Combined hate and offensive languages classes in their dataset to be hate class (Waseem & Davidson)
- Racism and sexism as a hate class, Neither as a non-hate class. (Waseem and Davidson)
- Offensive language and hate as a hate class and neither as a non-hate class (Davidson)
- Created a large combine dataset randomly selecting a similar number of examples for each class and the number of classes specified according to the lowest class

Wordcloud Representations

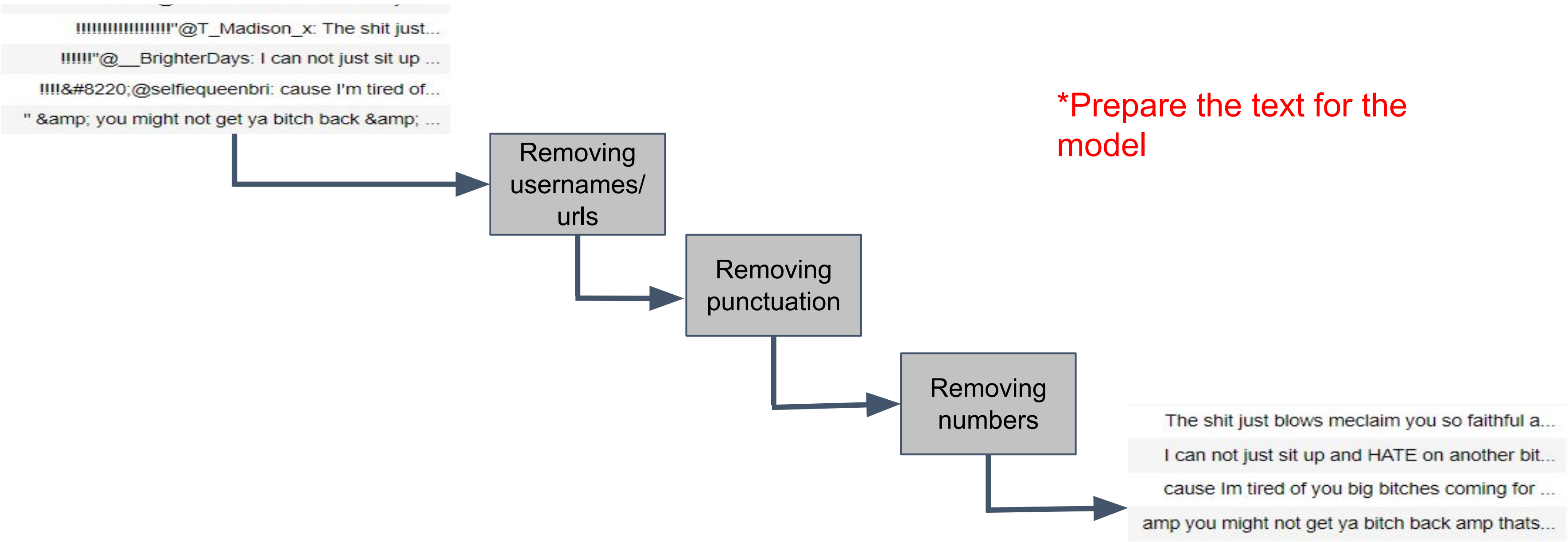
Word Cloud for Common Words in Hate Labeled Tweets

Word Cloud for Common Words in Offensive Labeled Tweets

Word Cloud for Common Words in Neither Labeled Tweets

Text Pre-processing

Removing noise from the model (unnecessary text & characters)



*Prepare the text for the model

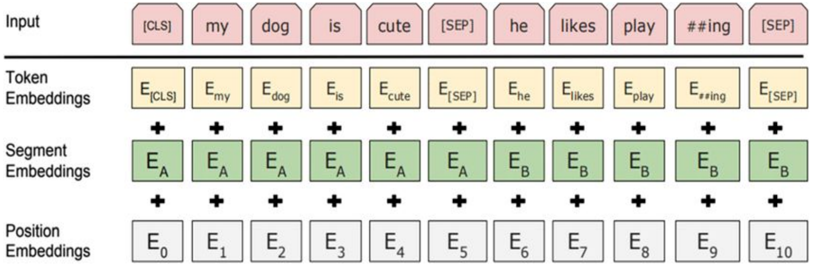
Text Pre-processing

Tokenization using BertTokenizer.from_pretrained('bert-base-uncased')

```
Text:   If Richnow doesnt show up with hella tinder hoes Im not his friend anymore chill I brought like like  prople
Token identification: [ 101  2065  4138 19779  2987  2102  2265  2039  2007  3109  2050  9543
  4063  7570  2229 10047  2025  2010  2767  4902 10720  1045  2716  2066
  2066 17678  2571  102     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
```

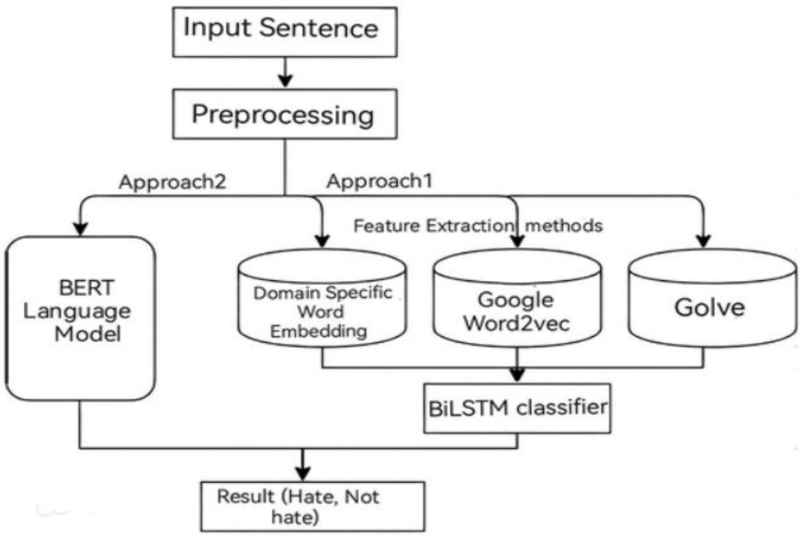
[CLS] = 101
[SEP] = 102

MAX LENGTH TWEET (TEXT) = 150

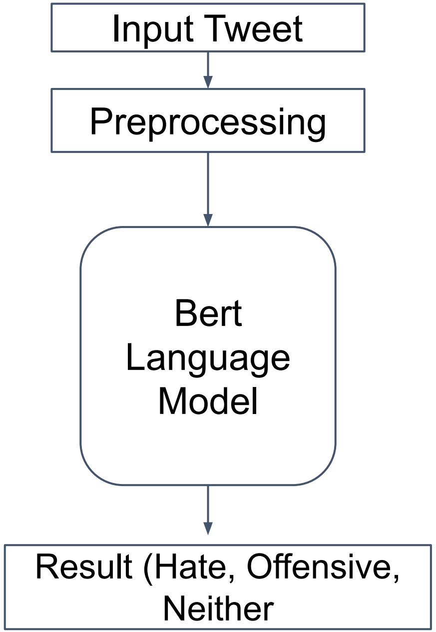


Model Overview

Figure 1. Block diagram of the experiments.

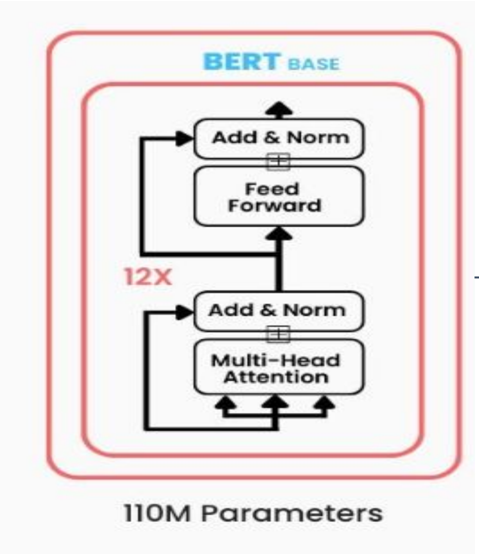


Saleh, H., Alhothali, A., & Moria, K. (2023).



Model Architecture and Parameters

BERT MODEL



BertForSequenceClassification

```
**Embedding Layer**
**bert.embeddings.word_embeddings.weight                    (30522, 768)**
**bert.embeddings.position_embeddings.weight                  (512, 768)**
**bert.embeddings.token_type_embeddings.weight                  (2, 768)**
**bert.embeddings.LayerNorm.weight                                (768,)**
**bert.embeddings.LayerNorm.bias                                  (768,)**

**First Transformer**
**bert.encoder.layer.0.attention.self.query.weight            (768, 768)**
**bert.encoder.layer.0.attention.self.query.bias                  (768,)**
**bert.encoder.layer.0.attention.self.key.weight              (768, 768)**
**bert.encoder.layer.0.attention.self.key.bias                    (768,)**
**bert.encoder.layer.0.attention.self.value.weight            (768, 768)**
**bert.encoder.layer.0.attention.self.value.bias                  (768,)**
**bert.encoder.layer.0.attention.output.dense.weight           (768, 768)**
**bert.encoder.layer.0.attention.output.dense.bias                (768,)**
**bert.encoder.layer.0.attention.output.LayerNorm.weight          (768,)**
**bert.encoder.layer.0.attention.output.LayerNorm.bias            (768,)**
**bert.encoder.layer.0.intermediate.dense.weight             (3072, 768)**
**bert.encoder.layer.0.intermediate.dense.bias                   (3072,)**
**bert.encoder.layer.0.output.dense.weight                   (768, 3072)**
**bert.encoder.layer.0.output.dense.bias                          (768,)**
**bert.encoder.layer.0.output.LayerNorm.weight                    (768,)**
**bert.encoder.layer.0.output.LayerNorm.bias                      (768,)**

**Output Layer**
**classifier.weight                                             (3, 768)**
**classifier.bias                                                   (3,)**
```

Train: 80% Test:10% Val:10%

AdamW

Learning Rate = 3e-5

Eps = 1e-8

Epochs = 4

Batch Size = 32

LR Scheduler = 3e-5

Paper parameters:

LEARNING RATE = 2e−5

NUM TRAIN EPOCHS = 3.0

BATCH SIZE = 16,8

Approach 1 Results: BiLSTM with various word embedding features

Source	Machine Learning Approach	Methods	Dataset	P	R	f1-score (hate)	f1-score (non-hate)	f1-score	AUC
Gupta and Waseem (2017)	LR	Hate W2V(300)	Davidson ICWSM	0.91	0.91	-	-	0.9120	0.8400
			Waseem EMNLP	0.84	0.86	-	-	0.8440	0.6380
			Waseem NAACL	0.76	0.77	-	-	0.7500	0.6790
Our proposed deep model	BiLSTM Deep model	GoogleNews- vectors-negative300	Davidson ICWSM	0.94	0.94	0.9679	0.8457	0.9473	0.9132
			Waseem EMNLP	0.91	0.91	0.6737	0.9484	0.9033	0.7697
			Waseem NAACL	0.80	0.80	0.6805	0.8542	0.7990	0.7654
			Combined balanced	0.94	0.94	0.9365	0.9376	0.9371	0.9370
		GloVe.6B.300d	Davidson ICWSM	0.94	0.94	0.9646	0.8310	0.9421	0.9073
			Waseem EMNLP	0.90	0.91	0.6631	0.9463	0.9028	0.7792
			Waseem NAACL	0.80	0.80	0.6794	0.8569	0.8002	0.7634
			Combined balanced	0.94	0.94	0.9414	0.9413	0.9414	0.9414
		GloVe.Twitter. 27B.200d	Davidson ICWSM	0.95	0.95	0.9676	0.8347	0.9453	0.8927
			Waseem EMNLP	0.91	0.90	0.6917	0.9399	0.9018	0.8324
			Waseem NAACL	0.80	0.80	0.6911	0.8503	0.7994	0.7738
			Combined balanced	0.94	0.94	0.9381	0.9394	0.9388	0.9388
		HSW2V(300)	Davidson ICWSM	0.95	0.95	0.9703	0.8551	**0.9509**	0.9162
			Waseem EMNLP	0.91	0.91	0.6928	0.9469	**0.9079**	0.8094
			Waseem NAACL	0.81	0.81	0.7055	0.8634	**0.8129**	0.7831
			Combined balanced	0.94	0.94	0.9428	0.9439	**0.9434**	0.9434

Approach 2 Paper Results: BERT (Base vs Large)

Table 4 of 6

Table 4. BERT for sequence classification hate speech experiment results (base-large).

Methods	Datasets	P	R	f1-score (Hate)	f1-score (non- Hate)	f1-score	AUC
BERT Base	Davidson- ICWSM	0.96	0.96	0.98	0.89	0.962	0.9309
	Waseem- EMNLP	0.92	0.92	0.7654	0.9541	**0.9216**	0.8455
	Waseem- NAACL	0.85	0.85	0.7612	0.8881	0.8472	0.8227
	Combined Balanced	0.95	0.95	0.9543	0.9552	0.9547	0.9547
BERT Large	Davidson- ICWSM	0.96	0.96	0.9788	0.8924	**0.9646**	0.9345
	Waseem- EMNLP	0.91	0.91	0.6939	0.9458	0.9103	0.8371
	Waseem -NAACL	0.85	0.85	0.7643	0.8937	**0.8521**	0.823
	Combined Balanced	0.96	0.96	0.962	0.9625	**0.9623**	0.9623

Results

Confusion Matrix

Test set prediction

	precision	recall	f1-score	support
0	0.50	0.37	0.43	134
1	0.93	0.96	0.95	1904
2	0.90	0.87	0.89	440
accuracy			0.91	2478
macro avg	0.78	0.73	0.75	2478
weighted avg	0.91	0.91	0.91	2478

```
======== Epoch 1 / 4 ========
Training...

  Average training loss: 0.30

Running Validation...
  Accuracy: 0.91

======== Epoch 2 / 4 ========
Training...

  Average training loss: 0.21

Running Validation...
  Accuracy: 0.91

======== Epoch 3 / 4 ========
Training...

  Average training loss: 0.16

Running Validation...
  Accuracy: 0.91

======== Epoch 4 / 4 ========
Training...

  Average training loss: 0.11

Running Validation...
  Accuracy: 0.91

Training finished!
```

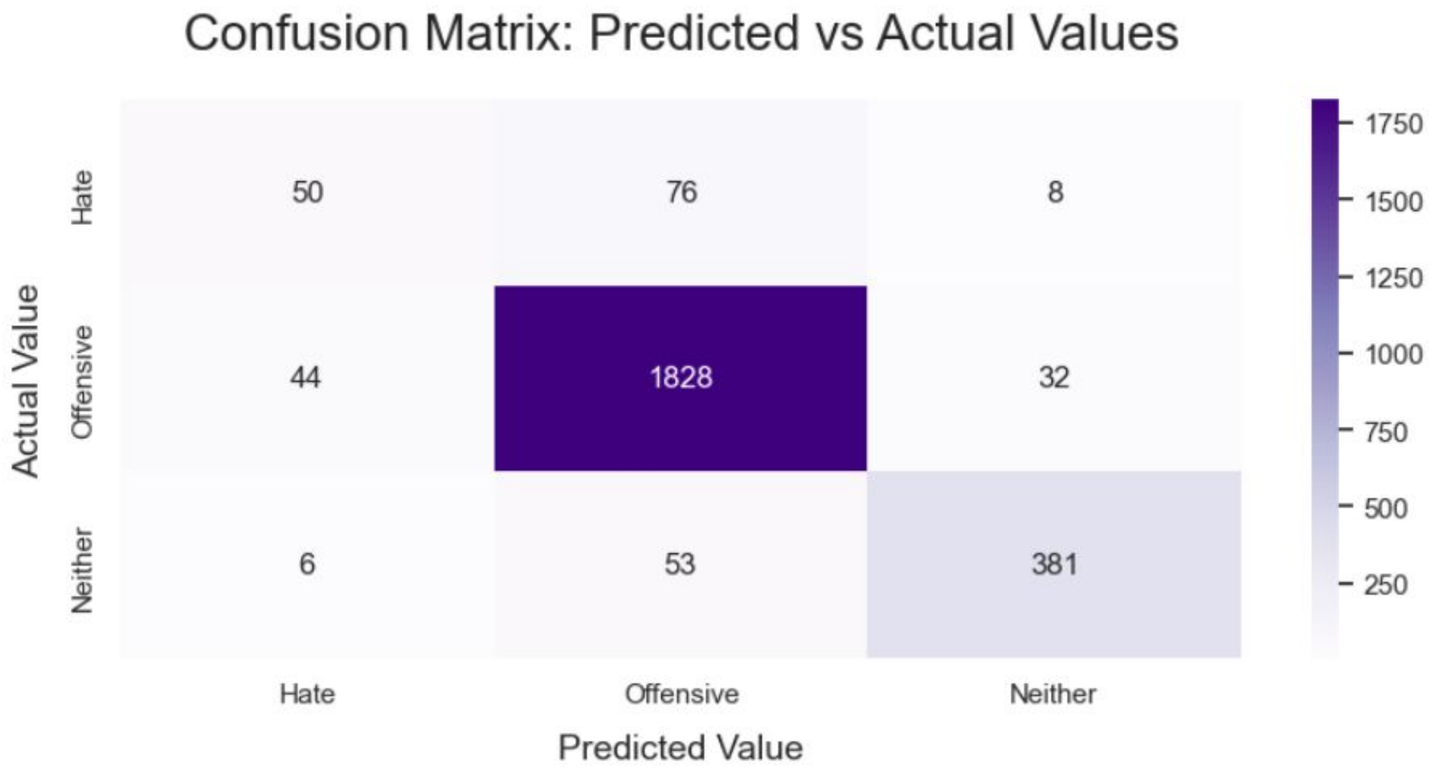
Results

Paper Results: Hate vs Non Hate

Methods	Datasets	P	R	f1-score (Hate)	f1-score (non- Hate)	f1-score
BERT Base	Davidson- ICWSM	0.96	0.96	0.98	0.89	0.962

Results: Hate vs Offensive vs Non Hate

Methods	Datasets (Test Data)	Class	P	R	f-1 Score
Bert Base	Hate Speech & Offensive Language Dataset	Hate (0)	0.50	0.37	0.43
Bert Base	Hate Speech & Offensive Language Dataset	Offensive (1)	0.93	0.96	0.95
Bert Base	Hate Speech & Offensive Language Dataset	Neither (2)	0.90	0.87	0.89


Confusion Matrix: Predicted vs Actual Values
Actual Value
Hate
Offensive
Neither
50
76
8
44
1828
32
6
53
381
Hate
Offensive
Neither
Predicted Value
1750
1500
1250
1000
750
500
250

Conclusion

- Helpful to build large pre-trained models from rich domains specific content in current social media platforms. BERT combines the benefits of domain-agnostic, domain-specific word embedding and domain-specific data (fine-tuning) due to the large amount of data it was trained (Hind Saleh, Areej Alhothali & Kawthar Moria (2023)
- Signifies the importance of differentiating hate speech vs offensive data
- Requires further investigation on how text in the dataset is classified
- Overfitting due to small size of test data for hate speech classification
- **Future exploration: Racial Bias in existing Hate Speech Detection Models. Building LLM's specifically with hate speech signifiers/ classified words.**
  - > Issues: context, sarcasm, bias against cultural and linguistic speech patterns
  - > Trust and Safety policies and regulations

Thank You!

AGGIES DO

References

- Hind Saleh, Areej Alhothali & Kawthar Moria (2023) Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model, Applied Artificial Intelligence, 37:1, DOI: 10.1080/08839514.2023.2166719
- Coldewey, D. (2019, August 15). *Racial bias observed in hate speech detection algorithm from google*. TechCrunch. https://techcrunch.com/2019/08/14/racial-bias-observed-in-hate-speech-detection-algorithm-from-google/
- Waseem, Z., and Hovy, D. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In SRW@HLT-NAACL, 88–93.
- Waseem, Z. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In Proceedings of the First Workshop on NLP and CSS, 138–142.
- Davidson, T., D. Warmsley, M. Macy, and I. Weber. 2017. "Automated hate speech detection and the problem of offensive language." In Eleventh international aaai conference on web and social media, Canada.