A Study of Hate Speech Detection and Classification

By Jerome HA
Emergent software laboratory

Introduction : Preliminary research

Hate speech is a problem for online etiquette and safe navigation on social media

Objective: Detect and classify hate speech using advanced NLP techniques.

Dataset: HateXplain, a benchmark dataset for explainable hate speech detection, consisting of 20k samples from Twitter and Gab labeled as "hateful," "offensive," and "normal." [2]

Impact: Effective hate speech detection can contribute to safer online communities.

https://twitter.com/
https://gab.com/

Introduction : Preliminary research

First semester :

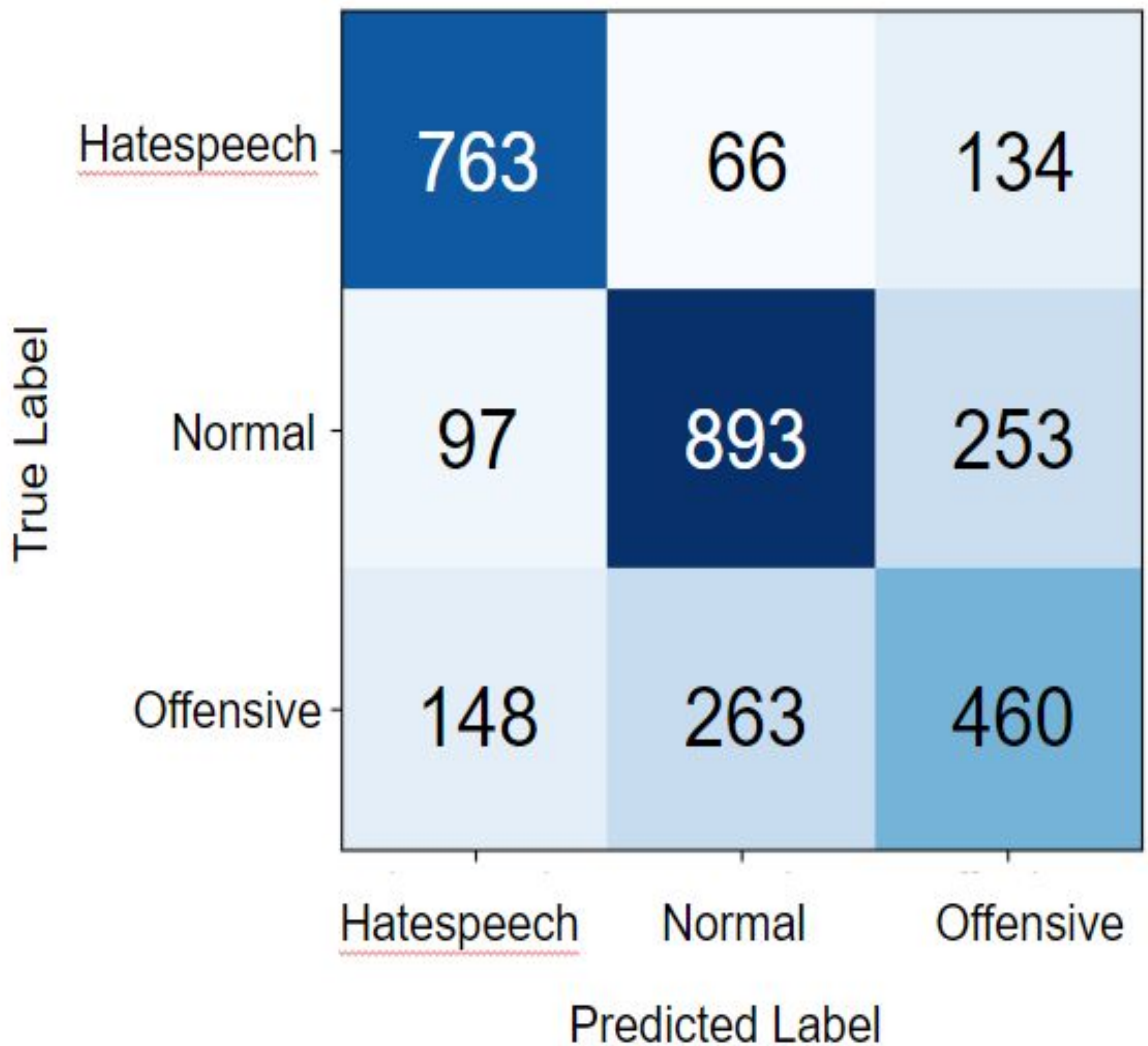
Classifier of 3 classes :

-Normal

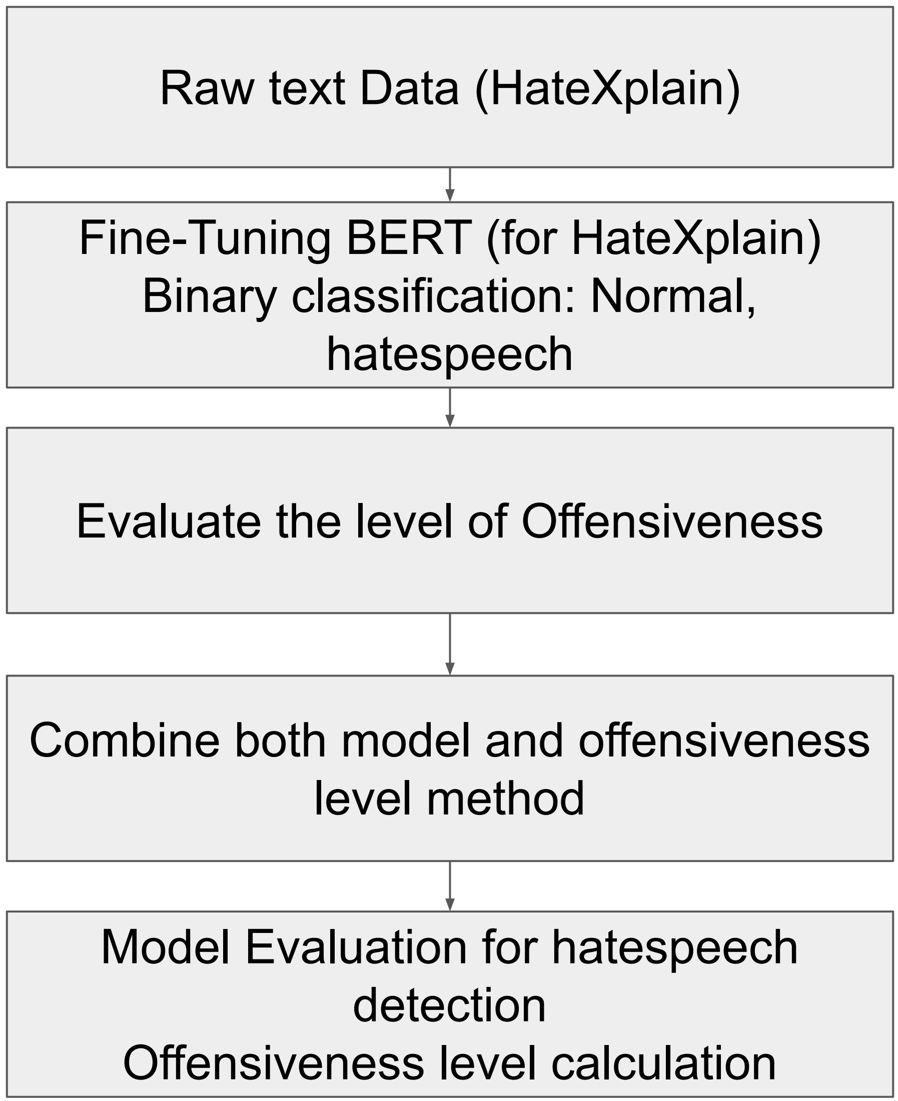
-Offensive

-Hate speech

Accuracy: 69%

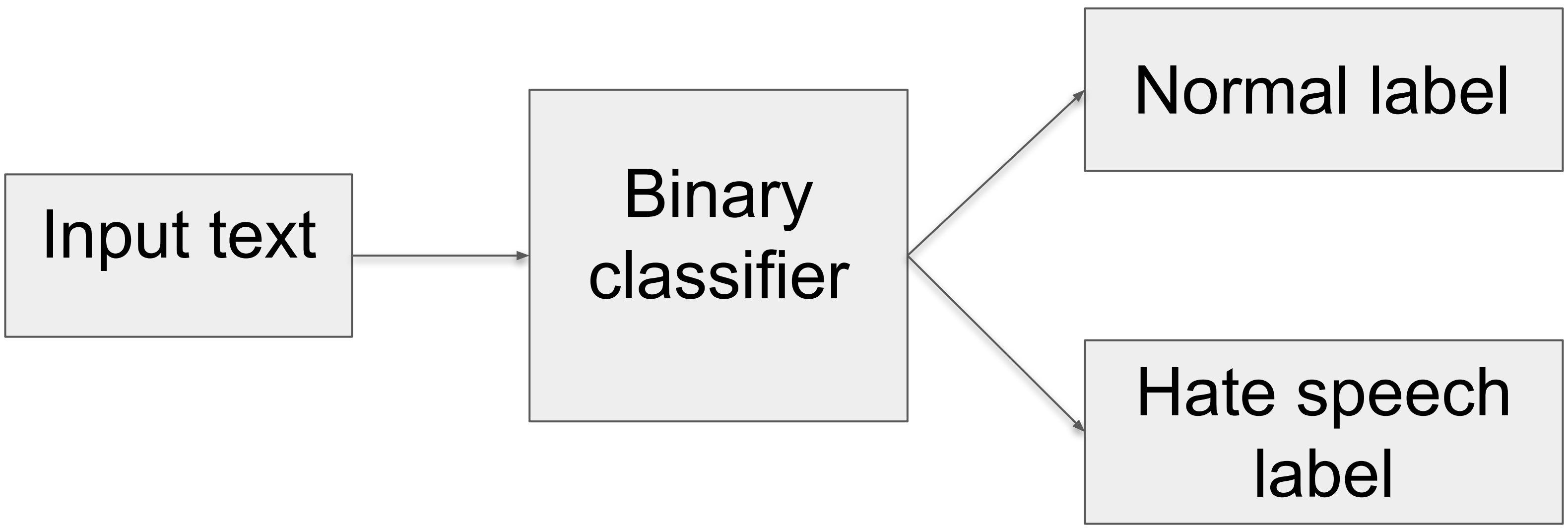


Proposed method



Schematic of the method for our detection model for hate speech and offensiveness

First step: Hate Speech detection



Model Development

- Data Preprocessing:

Tokenize texts using BERT's tokenizer [1].

- Model Architecture:

BERT (Bidirectional Encoder Representations from Transformers):

Pre-trained transformer model for understanding language context.

Fine-tuned on HateXplain for hate speech detection.

Model Development

Training Process:

Use Trainer class from Hugging Face's transformers library to manage training and evaluation.

Total Samples: 20,148

Split:

Training Data: 80% of the total data

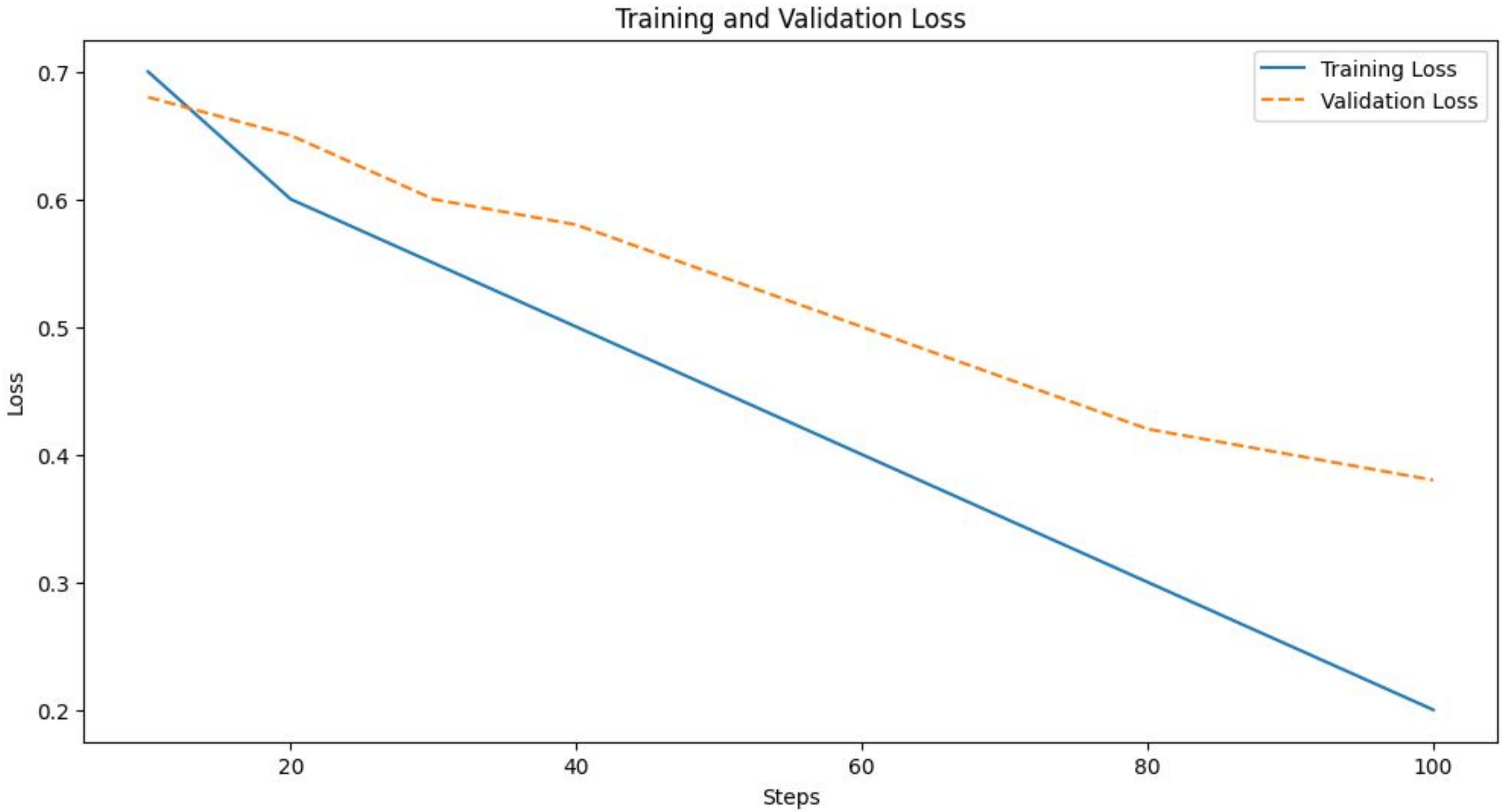
Validation Data: 5% of the total data

Test Data: 15% of the total data

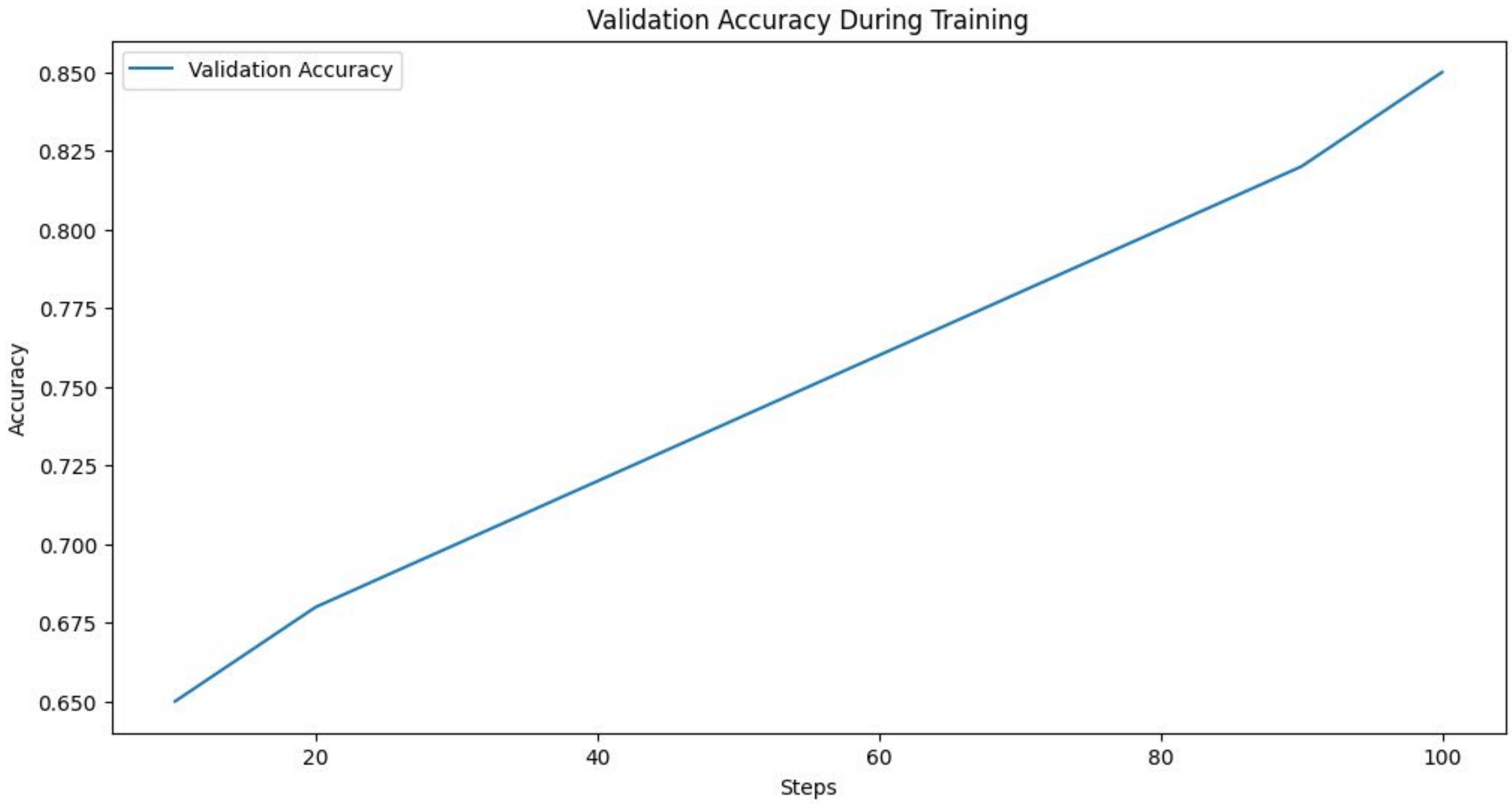
Model Development table

Hyperparameter	value
num_train_epochs	3
batch_size	8
Learning rate	Dynamic adjustment with warmup steps (initial: 5e-5)
warmup steps	500
BERT model	Bert-base-uncased pre-trained on BooksCorpus (11000 books), English Wikipedia (2.5 billion words)

Results

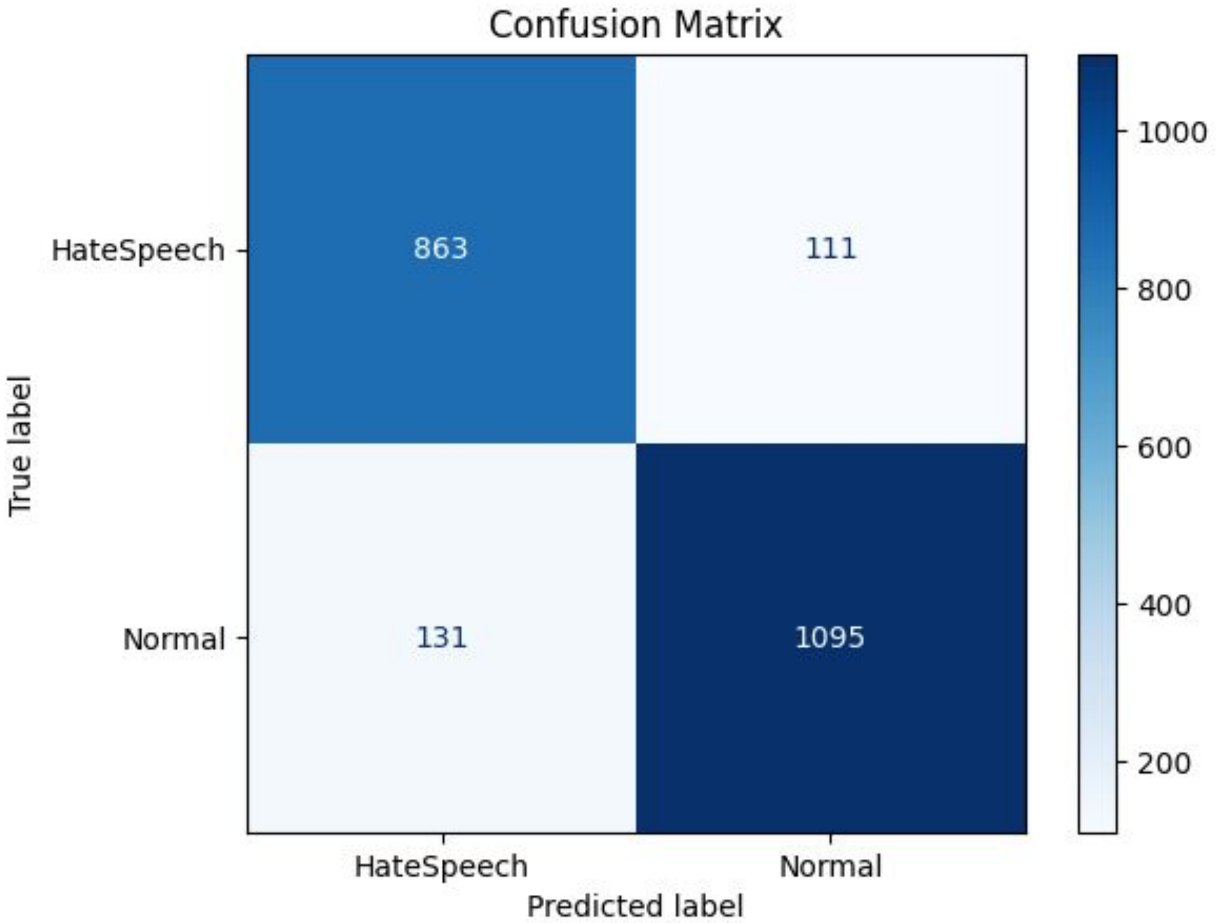

Training and Validation Loss
Training Loss
Validation Loss
0.7
0.6
0.5
0.4
0.3
0.2
Loss
20
40
60
80
100
Steps

Results


Validation Accuracy During Training
Validation Accuracy
Accuracy
0.850
0.825
0.800
0.775
0.750
0.725
0.700
0.675
0.650
20
40
60
80
100
Steps

Results

'eval_accuracy': 0.89



Analysis on falsely classified data

True Label: hatespeech, Predicted Label: Normal

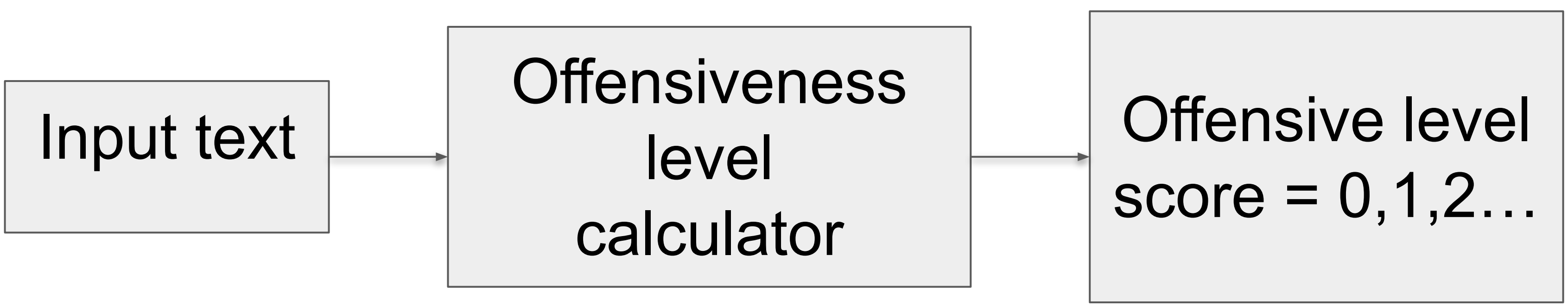
Example: me too but i still dis like jews and rather not have them follow me

----

True Label: normal, Predicted Label: Hatespeech

Example: it was most certainly high even for secular jews during the ellis island days but i think the percentage of jews relative to their population that use welfare is higher than say asians or white catholics not sure though

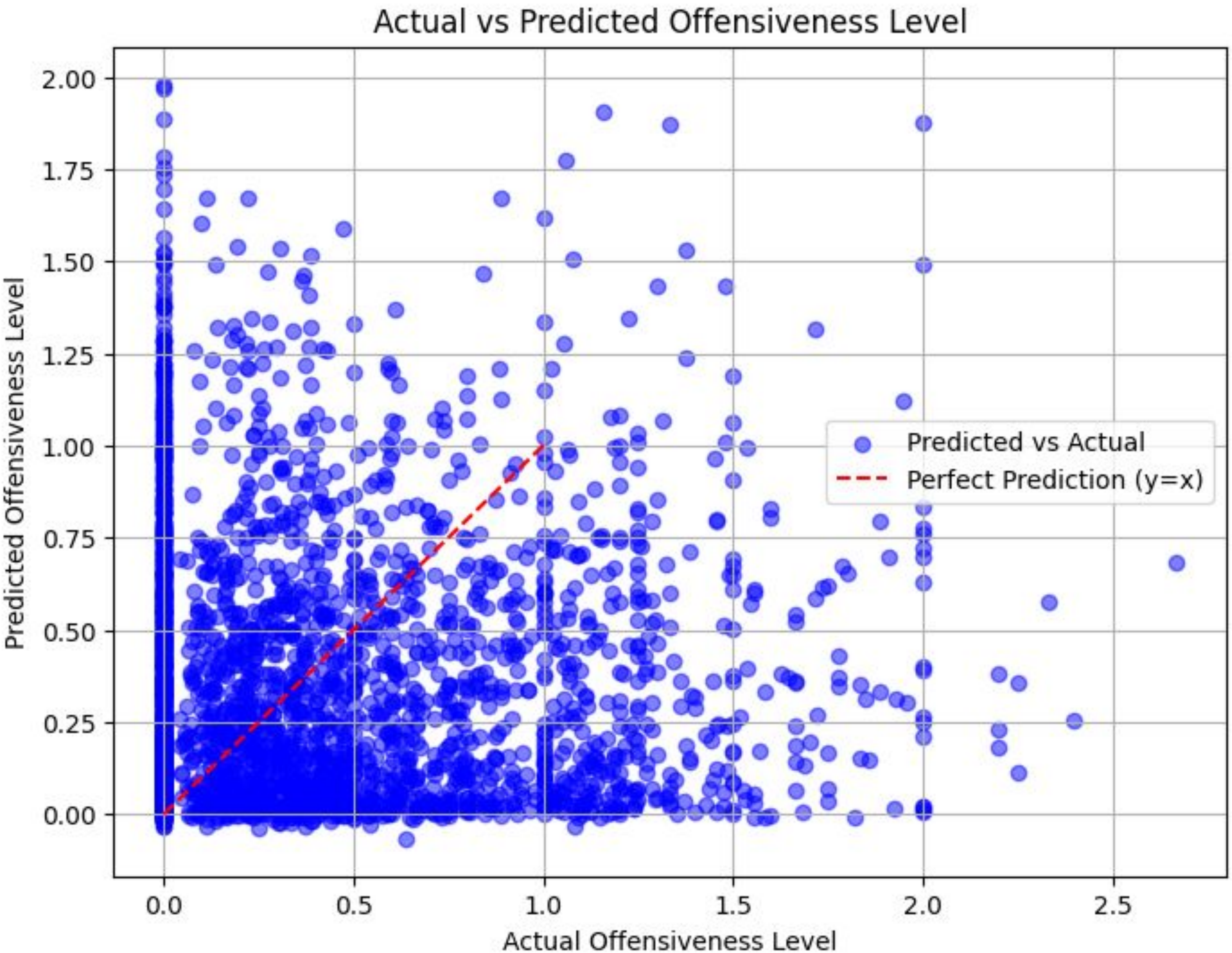
Second step: Offensiveness level calculation



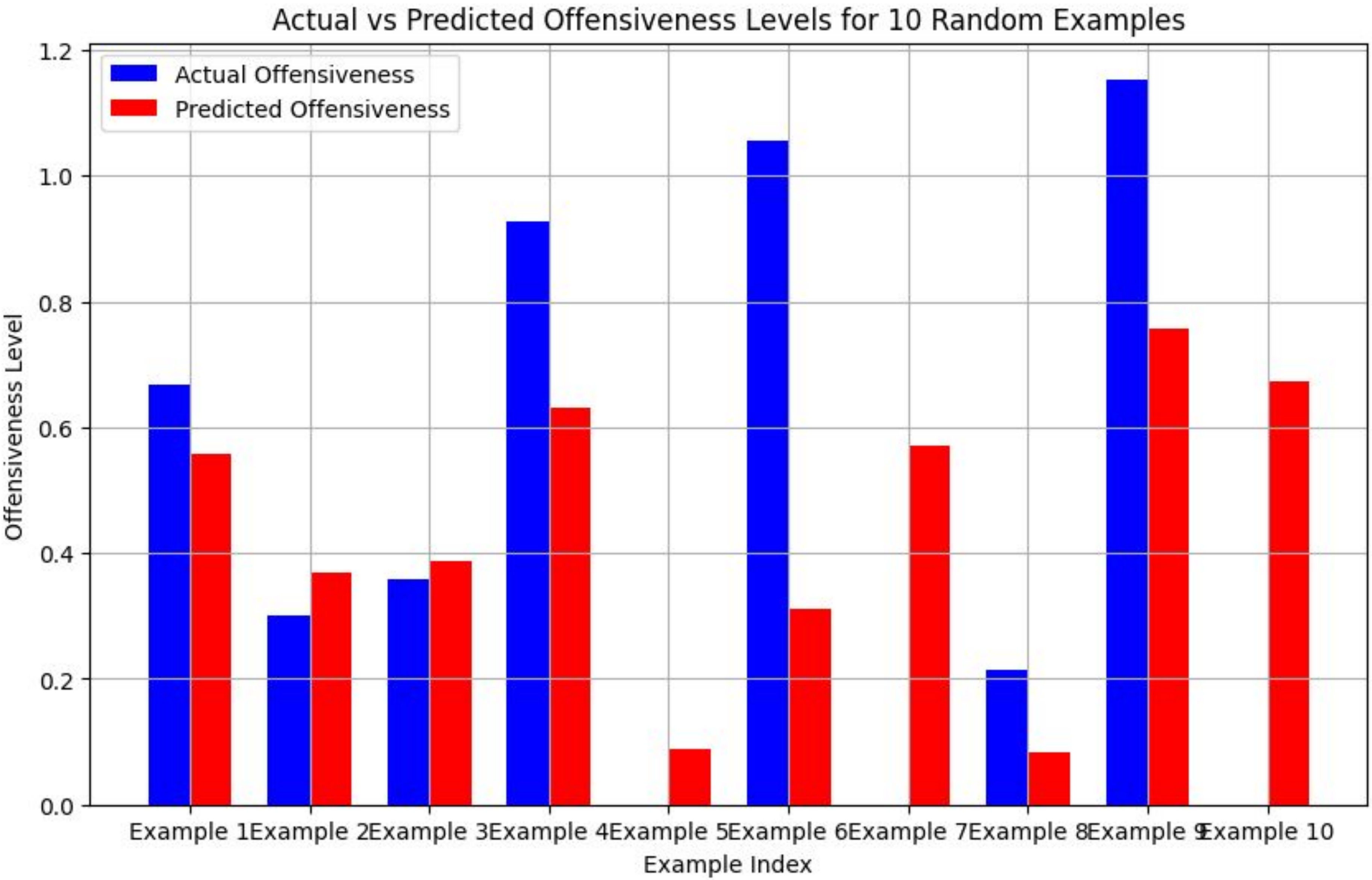
First method

A regression model using the Offensive labels by the annotator data

Each sentence of the dataset has 3 annotators that labeled Rationales or specific parts of the text with normal, offensive or hate speech.



Actual vs Predicted Offensiveness Level
Predicted Offensiveness Level
Actual Offensiveness Level
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.5
1.0
1.5
2.0
2.5
Predicted vs Actual
Perfect Prediction (y=x)


Actual vs Predicted Offensiveness Levels for 10 Random Examples
Actual Offensiveness
Predicted Offensiveness
Offensiveness Level
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Example 1
Example 2
Example 3
Example 4
Example 5
Example 6
Example 7
Example 8
Example 9
Example 10
Example Index

Second method

A predefined dictionary (profanity.txt) Using Wiktionary list of profanity for English (3171 words)

Accounting for slang, abbreviations, and misspellings.

Total offensive words = Offensiveness Level

Pages in category "English vulgarities"

The following 200 pages are in this category, out of 3,171 total.

(previous page) (next page)

4

- 4uck

A

- A2M
- abso-fucking-lutely
- absofuckinglutely
- abso-fuckin'-lutely
- abso-fuckin-lutely
- absofuckinlutely
- a damn sight
- AF
- arse over tit
- arse-up
- arsewash
- arseways
- arsewipe
- arsy varsy
- arsy versy
- artfag
- artfuck
- as all fuck
- as balls
- asf
- as fuck

https://en.wiktionary.org/wiki/Category:English_vulgarities

Result examples

True Label: normal, Predicted Label: normal

Example: the whole drink water cure all shit is fake i drink at least a gallon a day and have shit skin hair brain

Offensive score: 2

True Label: hatespeech, Predicted Label: hatespeech

Example: <user> she was screaming nigger and that she wish she could kill all the niggers in public at a cvs bitch

Offensive score: 3

True Label: normal, Predicted Label: normal

Example: done wit yo nigga i just want his wallet

Offensive score: 1

Potential bypass

Obfuscation with Special Characters

Example: f@ck, f*ck, fu&k, f#ck

Intentional Misspellings

Example: fuuuuuck, fucc, fuk, fasshole

Leet Speak (1337)

Example:h8, n1gger, b1tch

Potential Improvements and future works

Some Offensive words are more offensive than others, find a way to include that information

Improve the accuracy of Hate speech detection.

Multilingual: Enhance the model's ability to handle multiple languages and mixed-language content.

Code-Switching: Try to include deliberate misspelling on offensive words

Think of other programs, how to improve the idea of the research

Use **generative IA** future works and compare it with my method

Conclusion

Ongoing Research: Continuous improvement in model accuracy, interpretability, and adaptability.

Impact: Effective hate speech detection can contribute to safer online communities.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Long and Short Papers. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[2] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "Hatexplain: A benchmark dataset for explainable hate speech detection," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020, pp. 3451–3463.

https://en.wiktionary.org/wiki/Category:English_vulgarities