A Study of Hate Speech Detection and Classification

By Jerome HA
Supervised by Makoto OKADA

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]

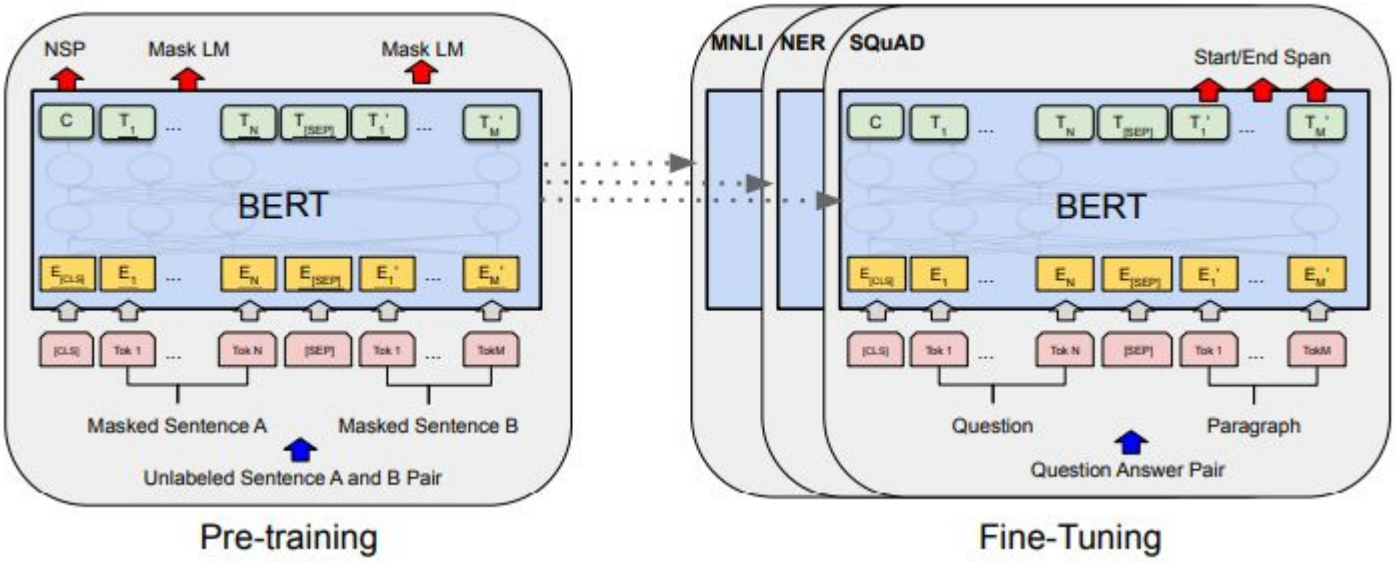
Impact: Effective hate speech detection can contribute to safer online communities.

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

Related research

BERT[1]:

- BERT Architecture : a pre-trained transformer based model for NLP (Natural Language Processing)



pre-training and fine-tuning procedures for BERT

Related research

HateXplain [2]:

Source: Data retrieved from Twitter and Gab, annotated by human raters.

Labels: Hateful (0), Offensive (2), and Normal (1).

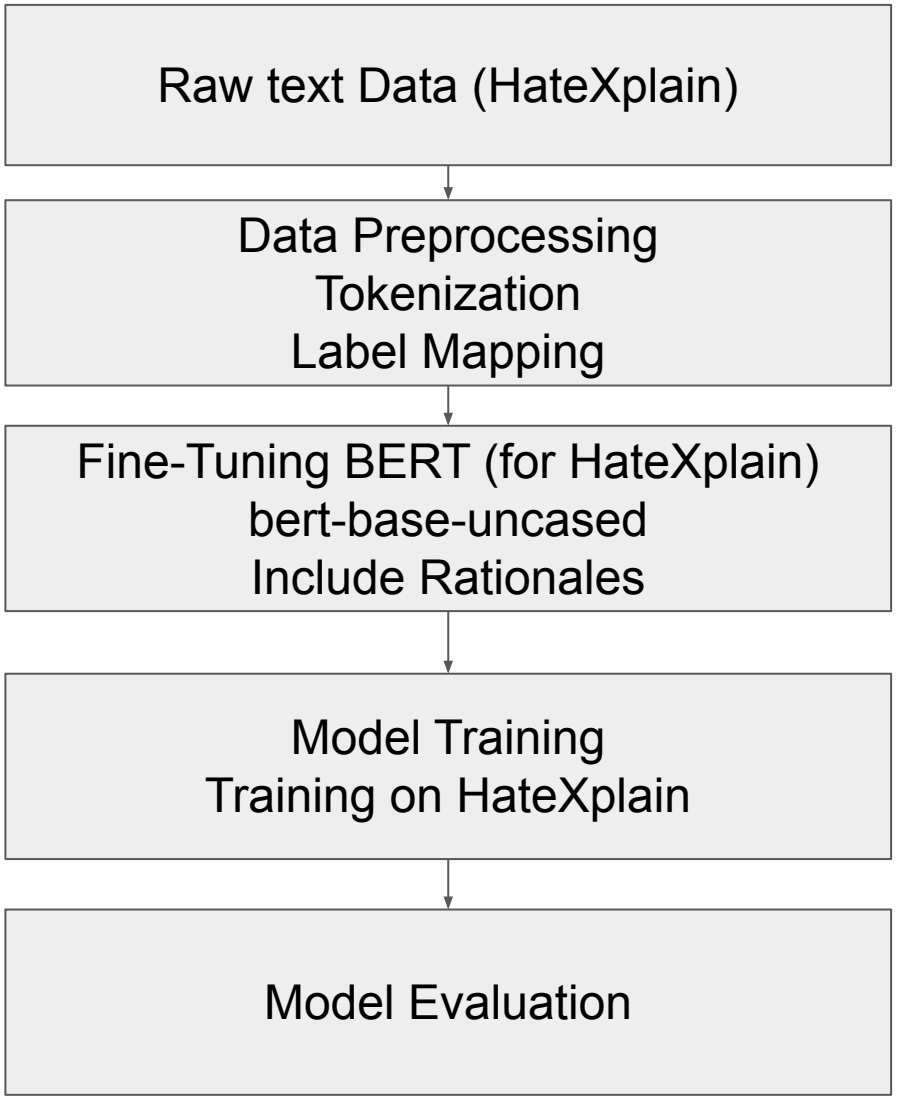
Additional Information: Each sample includes target groups and rationales for the annotations.

Initial Analysis

Class Distribution:

- Hateful: 4,748 posts
  Hateful speech targeting individuals or groups based on attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender.
- Offensive: 4,384 posts
  Content that includes rude or vulgar language but may not be specifically targeted towards a group or individual based on their identity.
- Normal: 6,251 posts
  Content that does not contain any hate speech or offensive language. This includes neutral, informative, or benign comments.

Proposed method (Base point of the research)



Schematic of the method for our classification model for hatespeech

Model Development

- Data Preprocessing:

Tokenize texts using BERT's tokenizer.

- 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

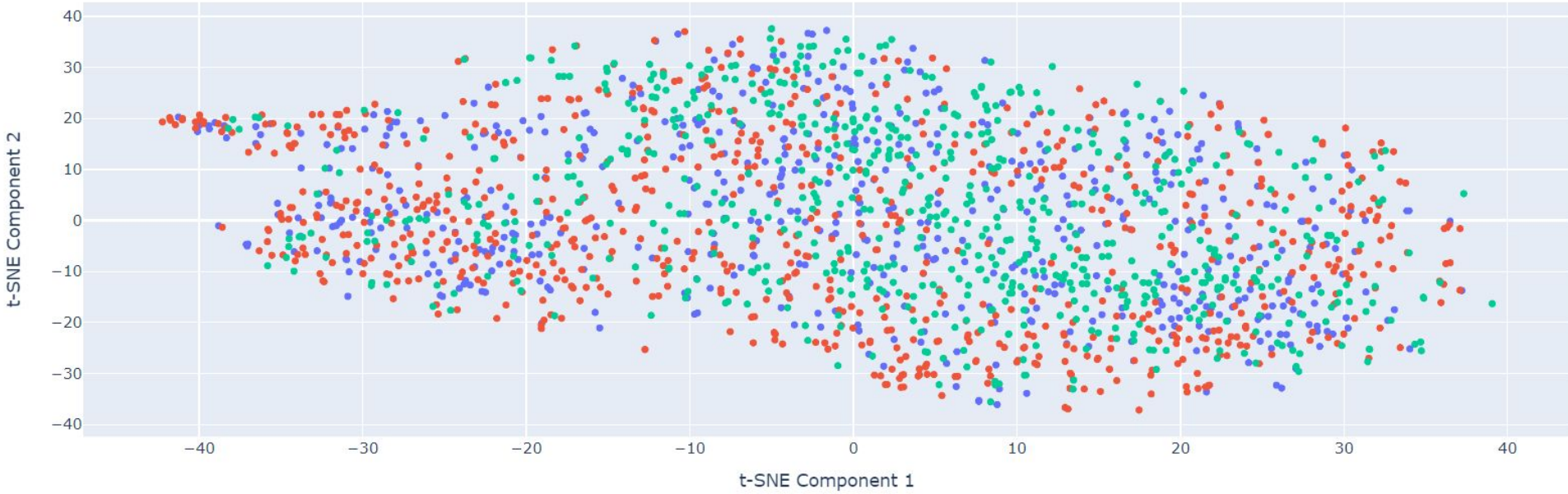
Hyperparameter	value
num_train_epochs	3
batch_size	8
Learning rate	Dynamic adjustment with warmup steps (initial: 5e-5)
warmup steps	500
BERT model	Bert-base-uncased pre-trained on BooksCorpus (11000 books), English Wikipedia (2.5 billion words)

Results

	precision	recall	f1-score	support
hatespeech	0.76	0.79	0.77	963
normal	0.73	0.72	0.72	1243
offensive	0.54	0.53	0.54	871
accuracy			0.69	3077

This is the baseline result which is same as the creator of the dataset which also used BERT [1].

t-SNE analysis of hateXplain dataset

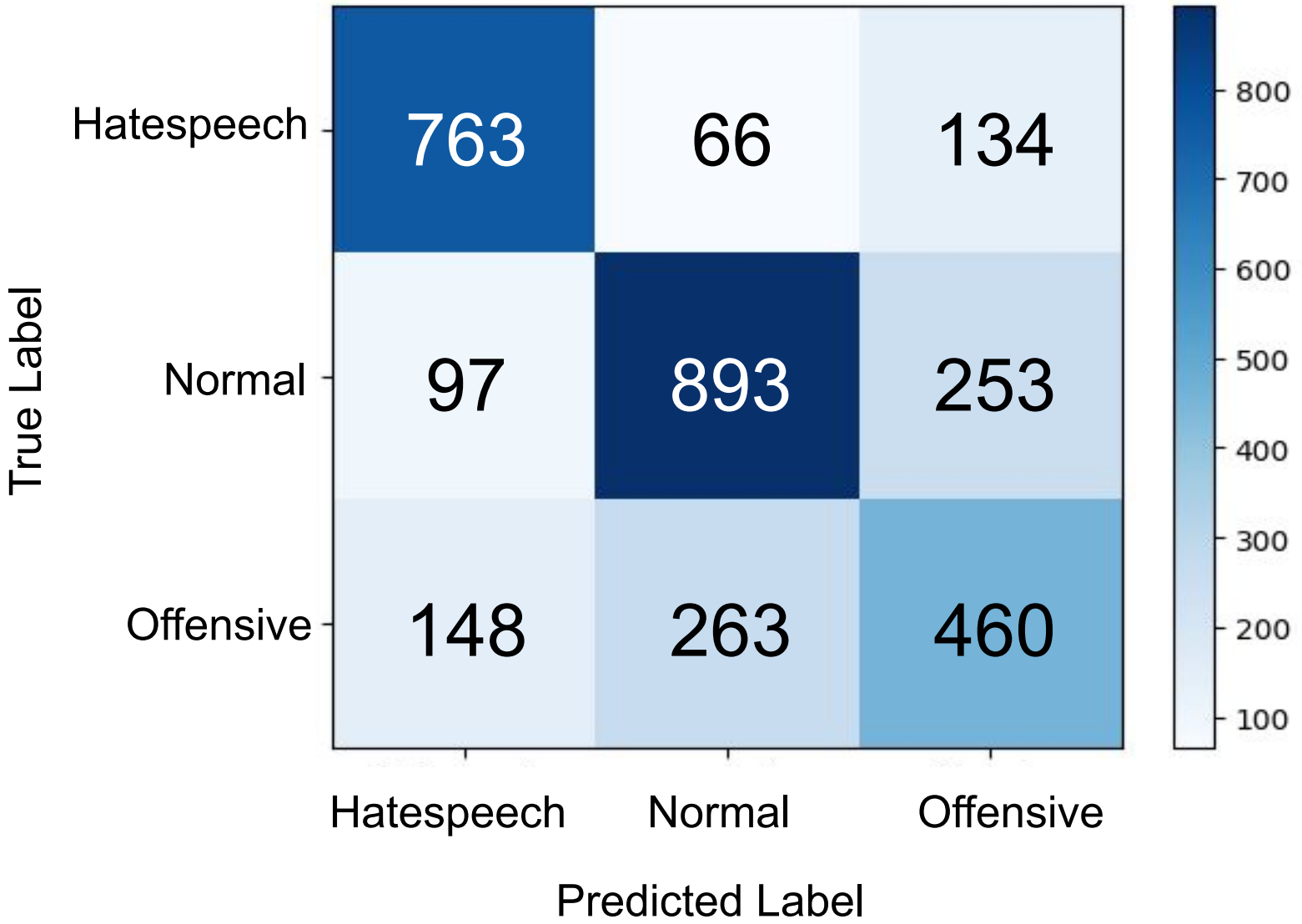


Labels

- Offensive
- Normal
- Hatespeech

True Label

Confusion matrix



Hatespeech examples better visibility

True Label: hatespeech, Predicted Label: hatespeech

Example: curryn*gger and known race traitor <user> has now hit the force unfollow button on me <number> times in the last <number> hours this little freeway shitstain can not pull off even pretending to be pro speech it just wants to cut things into swiss cheese safespace hugboxes for it weakminded little self gab fam

Hatespeech examples

True Label: hatespeech, Predicted Label: Normal
Example: i do not moozlim shitty hands on my food fuck tyson and starbucks

---

True Label: hatespeech, Predicted Label: offensive
Example: warning women men boy girls and toddlers all are fair game for rapes and other forms violence from these obama friendly terrorist immigrants

---

True Label: normal, Predicted Label: hatespeech
Example: can not find ne good content on youtube help u n*ggers

Challenges and Observations

- Misclassification Issues:

Significant misclassifications between "normal" and "offensive" categories.

Difficulty in handling subtle offensive language and context-dependent phrases.

- Unique and Shared Words Analysis:

Identified unique words in each category and words shared across multiple categories.

Examples of context-dependent words that change meaning based on usage.

- Also :

Ambiguity in language, contextual dependence, high inter-annotator variability, evolving language and subtle offensive content.

Potential Improvements

Data Augmentation: Increase the diversity of training data.

Class Weight Adjustment: Balance the impact of underrepresented classes.

Hyperparameter Tuning: Optimize model parameters for better performance.

Advanced Models: Explore more sophisticated models like RoBERTa or BERT-large.

Ensemble Methods: Combine multiple models to improve accuracy.

Future Work

Cross-Domain Transfer Learning: Adapt the model to different datasets and domains.

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

Explainability: Improve the model's ability to provide transparent and understandable explanations for its predictions.

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.