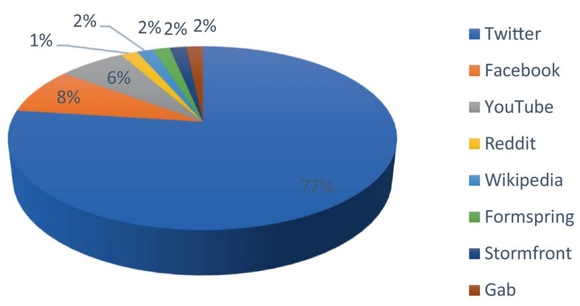
Hate Speech Detection Pipeline: Comparing Traditional and Transformer Methods

Devanshu Dixit, Saeid Ghayour, Vrinda Valaboju
Virginia Polytechnic Institute and State University

Motivation

As social media usage increases, so is the prevalence of hate-speech. The speech targets marginalized communities promoting violent and hurtful behavior. Manual moderation is slow, expensive, and inconsistent and automated detection can moderate and create safe online spaces.



Prevalence of Hate-Speech on social media 2024

Research Questions

1. Can we accurately classify tweets into *Normal*, *Offensive*, or *Hate Speech*?
2. Can we correctly identify the *target group* of hate speech (e.g., religion, race, gender)?
3. How do traditional NLP approaches such as TF-IDF+SVM and GloVe compare with modern approaches such as BiLSTM and DeBERTa?
4. What trade-offs exist between accuracy, computational cost, interpretability, and speed?

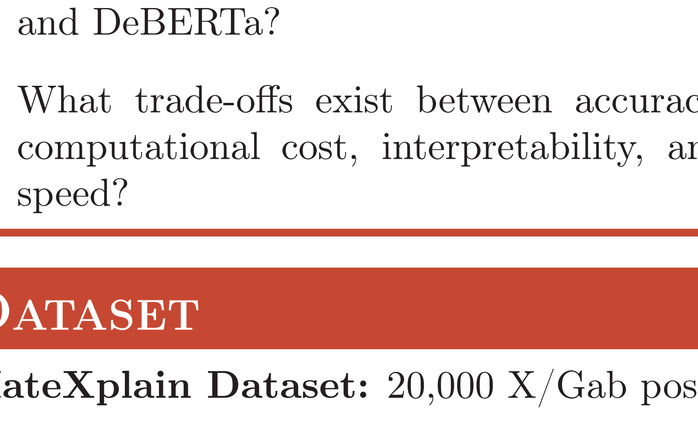
Dataset

HateXplain Dataset: 20,000 X/Gab posts

- Three labels: **Normal**, **Offensive**, **Hate Speech**
- Stratified **80/10/10** train/validation/test split

7	14567516_gab	nogs jews and dykes how enriching	offensive	2	Jewish	train
8	19679648_gab	guilty until proven innocent unless you are a ...	hatespeech	0	African	train
9	15858421_gab	tired i can not support abortion from a moral ...	hatespeech	0	African	val
10	23844035_gab	<number> <number> <percent> of brits think mul...	hatespeech	0	Islam	train
11	1089312238102609921_twitter	tariq nasheed beating the fuck out of that dru...	normal	1	None	train
12	1159278048069464065_twitter	<percent> cotton thigh highs to avoid contact ...	normal	1	None	train

HateXplain Data Sample



Task Setup / Approaches

The project builds a complete pipeline for analyzing hate speech in social media posts and identifying harmful language.

- **Task 1:** Classification of posts into categories: hate-speech, offensive, normal.
- **Task 2:** Multi-class classification to detect which groups are targeted (e.g., race, religion, gender, sexuality).
- **Task 3:** Use a Large Language Model (LLM) to generate a explanation hate-speech posts.

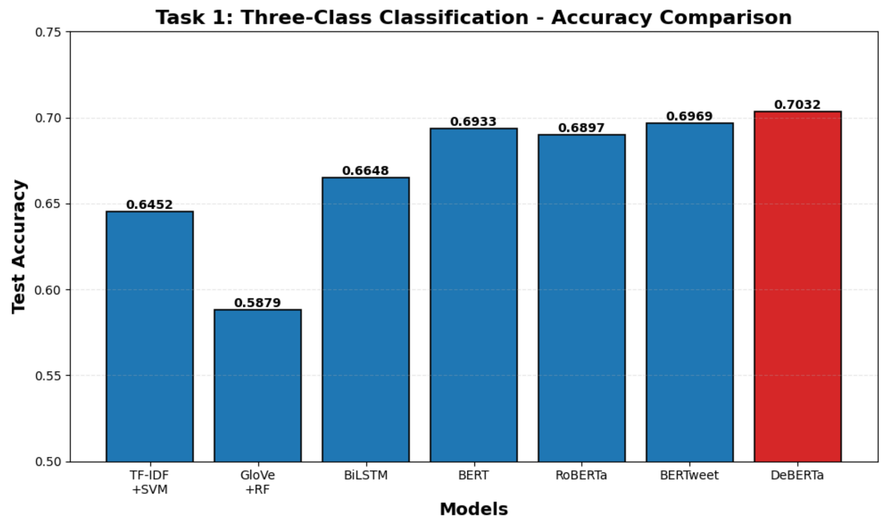
Model Approaches

Category	Model	Notes
Traditional ML	TF-IDF+SVM	Baseline
	GloVe+SVM/RF	Embeddings
Deep Learning	BiLSTM	Sequential
Transformers	BERT	Baseline
	RoBERTa	Improved
	DeBERTa	Robust
	BERTweet	Twitter
LLM	Gemini/GPT	Explain

Experiments & Results

Task 1: Three-Class Classification

Model	Macro F1
TF-IDF + SVM	0.6451
GloVe + RF	0.5771
BiLSTM	0.6569
BERT	0.6827
RoBERTa	0.6766
BERTweet	0.6855
DeBERTa	**0.6925**



Accuracy comparison across models

Task 2: Target Group Classification

Model	Macro F1
TF-IDF + SVM	0.6329
GloVe + RF	0.4608
BiLSTM	0.5830
DeBERTa	**0.6720**

	text	pred_targets_str	task3_gemini_explanation
0	i dont think im getting my baby them white 9 h...	no specific target group	no target group – model predicted normal/other
1	laura loomer raped me while screaming at me in...	Jewish	This tweet contains antisemitic hate speech an...
2	<user> what did the old lady do was she a nazi...	no specific target group	no target group – model predicted normal/other
3	as much as i appreciate a format of open discu...	Hispanic	This tweet contains offensive language express...
4	sex be so good a bitch be slow stroking and cr...	Women	This tweet contains derogatory and sexually su...
5	<user> <user> <user> with a room full of white...	no specific target group	no target group – model predicted normal/other
6	i do not blame orlando pie rats for loosing li...	no specific target group	no target group – model predicted normal/other
7	they are going to figure out a way to kike bit...	Jewish	This tweet contains hate speech targeting Jewi...
8	<user> <user> a camel jockey midde east wateri...	Islam	This tweet contains hate speech using derogato...

Task 3: LLM Generated Explanations

Key Results:

- DeBERTa achieves best performance on both Task 1 (F1: 0.6925) and Task 2 (F1: 0.6720)
- Transformers outperform traditional ML approaches
- Task 2 (target group classification) proves more challenging than Task 1
- LLM-based explanations align with ground truth target groups in most of cases

Key Takeaways

1. Traditional methods remain competitive. TF-IDF has 90% of transformer performance with 5% of training time.
2. Hate-speech detection relies on explicit lexical patters -> baselines are effective
3. Transformers: Explicit slurs overtake dataset, making contextual understanding less helpful
4. Domain adaption important! BERTweet > RoBERTa

Misclassification Insights

- Sarcastic or indirect hate speech is often misclassified as "Offensive".
- Tweets with overlapping stereotypes (e.g., nationality + religion) confuse the model.
- Some offensive tweets without group targeting are incorrectly labeled as hate speech.
- Hybrid approach reduces bias by providing retrieved example anchors.

Acknowledgments

We thank the Virginia Tech Department of Computer Science and Professor Vaibhav Garg for guidance, and the HateXplain team for providing annotated open data.

Reference

Rawat, A., Kumar, S., Samant, S. S. (2024). Hate speech detection in social media: Techniques, recent trends, and future challenges. WIREs Computational Statistics, 16(2). https://doi.org/10.1002/wics.1648