Implicit and Explicit Hate Speech : An Empirical Analysis

Aarushi Gupta, Abhay Goel, Himanshu Mangla

{agupta857, agoel84, hmangla6} @gatech.edu

GOAL

- As social platforms grow, the communities on them also expand, and so does the amount of online hate
- Offensive and hateful language can incite violence, hurt people's sentiments and cause societal divide. It's important to analyze and control the spread of such hate speech
- Some platforms use human moderators and hate-speech identification techniques to control such hate
- We **identify implicit hate and explicit hate** present in the benchmark binary *ETHOS dataset* [2]
- We differentiate between implicit and explicit hate distribution in data collected from different key **subreddits**
- We perform **Temporal Analysis** of proliferation of hate speech on these platforms around major global events like the 'US Presidential Elections', 'Soccer's Champions League' finals and the ongoing 'Ukraine-Russia conflict.'

PROGRESS

- Until the previous report we had analyzed the number of hateful posts on **three different subreddits**: *r/championsleague*, *r/europe* and *r/politics*
- Performed Classification : **Implicit / Explicit / Not Hate**
- Common platform for different communities. For example, in r/Europe, most posts were about the Russia-Ukraine conflict as seen in the word cloud in Figure 4. People who supported different sides became toxic towards the others. A similar trend was seen in all the other subreddits as well

DATA ANALYSIS

1. SELF COLLECTED DATA

Top 100 Reddit posts for each day, from three different subreddits around respective important world events.

Subreddit Name	No. Of Days	No. of Posts
r/politics	158	15754
r/championsLeague	60	6000
r/europe	18	1796

Table 1. Summary of Data collected from Reddit

2. BENCHMARK DATA

ETHOS data [2] of Reddit and YouTube's hateful comments

Class	Number of Samples
Hate	359
Not Hate	639

Table 2. Summary of ETHOS Dataset

METHODS

1. BINARY CLASSIFIER:

- Benchmark binary classification model's comparison to our model *(Implicit vs Not Hate)*
- BERT based binary classifier (lr = 5e-5, batch size = 64)

Model	Accuracy	F1 Score
Latent Hatred (BERT)	78%	0.68
Our Model	72%	0.71

Table 3. Model Performance

- Also did Cross–Domain inference on Hate vs Not-Hate binary data taken from ETHOS [2]
- **70.3%** accuracy compared to the ground truth labels in [2]

2. THREE WAY CLASSIFIER:

- BERT based 3-way classifier
- Hyper parameters same as above
- Dataset of 5K posts, with a split of 2:2:1 ~ Not-Hate : Implicit : Explicit
- 80:20 Train : Test Split
- **Accuracy - 65%**
- **F1 score - 0.64**

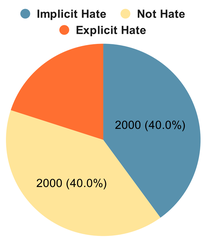

Fig 1. Training Data Summary

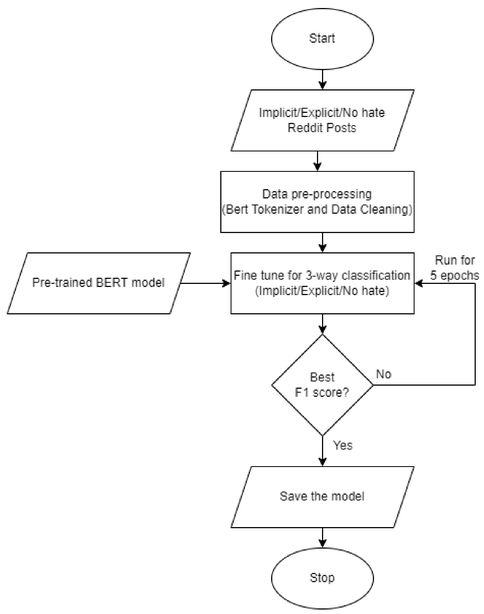

Fig 2. Flowchart explaining the model training process

QUANTITATIVE & QUALITATIVE ANALYSIS

Ground Truth	Implicit Hate	Explicit Hate	Not-Hate
Hate	**173 (48%)**	168 (47%)	18 (5%)
Not Hate	**295 (46%)**	62 (9%)	282 (45%)

Table 4. Results of Inference over Ethos Data

3-way classification over benchmark ETHOS Dataset:

- Only 5% (18) of Hate posts were misclassified as Not-Hate and 9% (62) of Not-Hate posts got misclassified as Explicit Hate
- From the 639 Not-Hate posts, our model identified an **extra 46% (295) Implicit Hate** posts
- Manual Inspection of the Implicit Hate Posts pending

TEMPORAL ANALYSIS

1. r/politics:

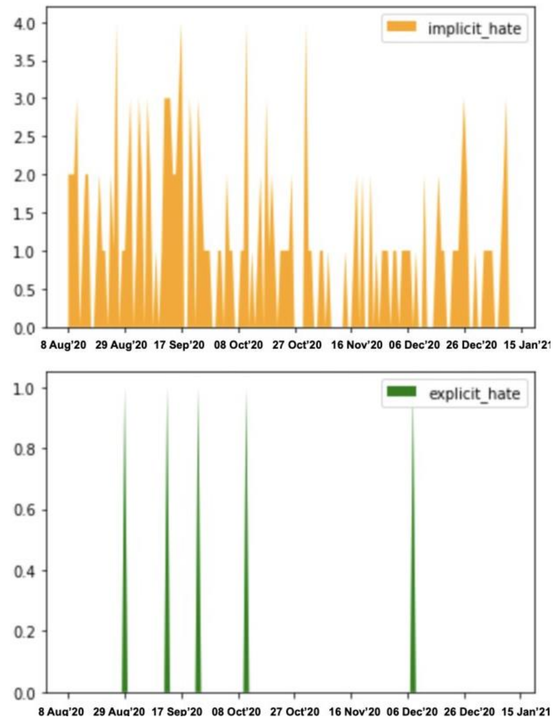

Fig 3. Number of Hatful posts as a function of time

A. Both Implicit and Explicit Hate decreased after 8th November 2020

B. Hateful content was propagated before elections to negatively influence the choice of the voters

TEMPORAL ANALYSIS

2. r/europe

Fig 4. Word Cloud for *r/europe*

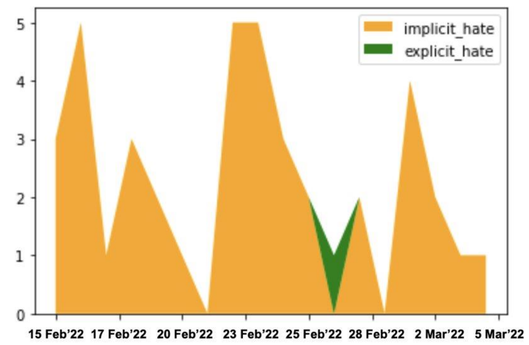

Fig 5. Number of Hatful posts as a function of time

A. **Implicit Hate** peaked around **21st February 2022, the day of conflict** beginning.

B. **Explicit Hate** also observed near the same time

3. r/championsLeague

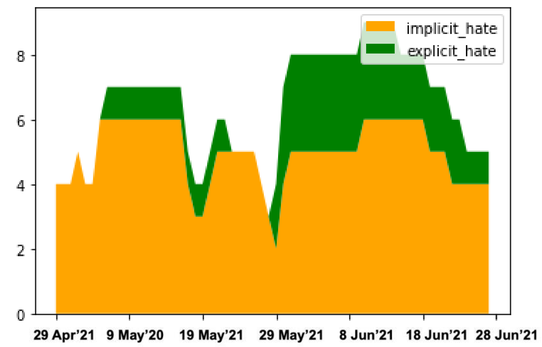

Fig 6. Number of Hatful posts as a function of time

A. **Constant Implicit Hate** : English Clubs with prior rivalry, might cause such hate

B. **Explicit Hate** after 29 May 2021**, the Finals,** as expected

REFERENCES

1. ElSherief, M., Ziems, C., Muchlinski, D., Anupindi, V., Seybolt, J., De Choudhury, M., & Yang, D. (2021). Latent Hatred: A Benchmark for Understanding Implicit Hate Speech.
2. Mollas, I., Chrysopoulou, Z., Karlos, S., & Tsoumakas, G. (2020). Ethos: an online hate speech detection dataset. *arXiv preprint arXiv:2006.08328*.