Implicit and Explicit Speech: An Empirical Analysis

Aarushi Gupta[1], **Himanshu Mangla**[1], **Abhay Goel**[1]
College of Computing, Georgia Institute of Technology
[1]{agupta857, hmangla6, agoel84}@gatech.edu

Abstract

Most of the prior research in Hate Detection deals with Explicit Hate. While some prior work also discusses identifying Implicit Hate and ways of classifying it, most of this work has been done by identifying hateful posts over Twitter. In this work, we study the proliferation of Implicit vs Explicit Hate across the Reddit domain. We study different subreddits to see the difference in the type of hate and the amount of hate across them. Also, we argue that proliferation of hate increases around major global events. To verify this, we perform Temporal Analysis over the Reddit data from r/politics around the US Presidential Elections, from r/championsleague around the UEFA Champions League finals and from r/europe around the start of the Ukraine-Russia conflict. We trained our model using data from (1). Our fine-tuned version of a pre-trained BERT model classifies the data across implicit, explicit and not hateful with an accuracy of 65% over test set. Performing inference over unlabeled Reddit data, we verified their quality through manual labelling of a subset of data and obtained Fleiss Kappa annotator conformance of 0.61. On analyzing the results for r/politics, we see that overall hate decreases after the presidential elections. We see an increase in Explicit Hate in r/championsleague after the finals and we also see that explicit hate appears in r/europe as soon as the conflict in Ukraine-Russia happens. Such findings can help shape activity monitoring and content moderation policies in social media platforms. The code for our project is hosted at our **Github Repository**

Introduction

Social media platforms provide several use cases. Not only are they a primary source of direct communication between people, but they also enable the spreading of news to masses and acting as a source of entertainment for people through memes, tweets and videos. As their usage increases, there are cases of both good and bad usage of online platforms. Some people engage in Explicit Hateful comments about other people, groups or institutions. Others use means such as sarcasm and metaphor to depict their hate implicitly. It becomes imperative that we identify both such types of hate.

A lot of work has been done on identifying hate in Twitter Data. In this work, we collect Reddit data ourselves for different subreddits across varying timelines. We develop a model using data from (1) and verify its performance on a binary classification task in the same domain and across domain using data from both (1) and (2) sources.

Then we use our model and apply its learning to develop a three-class implicit vs explicit vs non-hate classifier that identifies hate across a different domain of Reddit data. We verify the quality of inference by manual inspection of a subset of results and evaluating annotator conformance and then also try to analyze the reasons why some of the not-hate labelled data previously, are now identified by our model as cases of implicit hate.

Building on the progress from our midterm report, we improved our model's accuracy. We also performed an evaluation of our model over a new standard (2) data. Overall, we study the amount of implicit and explicit hate present in different subreddits. We also study how hate patterns change around major global events such as the US Presidential Elections, Soccer's Champions League Finals and the ongoing Russia-Ukraine conflict.

Related Work

In the past, we have seen Hate speech being studied and measures being taken against them (3). The paper from Warner and Hirschberg et al. (4) has been one of the first to study this. They detected stereotypical words which might be inappropriate towards certain communities. This and much other similar research had a common problem that they did not consider implicit hate, which could be propagated through sarcasm, implying inferiority, etc. Our study was on both Implicit and Explicit hate, and we conducted an Empirical analysis towards this toxicity.

Our initial aim was to boost the model performance using sentiment score, however, previous work has suggested that sentiment analysis has not been very useful in detecting the type of toxicity on social media. The paper by Phadke and Mitra (5) mentions that even positive framing of negative entity can be problematic. many challenges have been discussed by Zhang et al. in (6) to understand the future scope of this problem. We still try to understand the spread of fine-tuned sentiment score (12) across implicit and explicit labels that we gather to validate whether the initial findings still

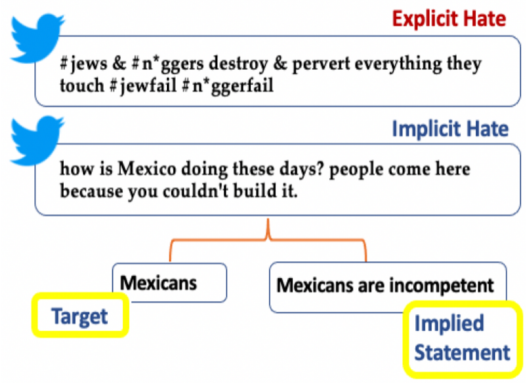


Figure 1: Example of Latent Hatred Dataset

Dataset	Training Dataset		
Labels	Implicit	Explicit	Not Hateful
Total Posts	7100	1089	13291
Avg Character Length	95.3	88.08	86.41
Max Character Length	801	363	322
Min Character Length	7	7	4

Figure 2: Statistics for the Latent Hatred Dataset

stand with implicit hate detection.

Benikova et al. discusses the sensitivity of automatic hate speech detection in (7), which shows the relevance of our study. The paper by Mathew et al. (8) explained how hate has a temporal aspect as well. The paper by ElSherif et al (1) was the benchmark study upon which we implemented our project. This study discussed the importance of implicit hate detection. We pulled the Implicit Hate Dataset from this study with the data labels: Implicit, Explicit and Not Hateful. To further evaluate how well our model can classify the posts, we used the dataset from the paper by Mollas et al. (2) which provided a textual dataset with two variants: Binary and Multi-label. We used the binary hate and not hate labelled data to validate our binary classifier. Next we aimed to understand how efficient our model is to identify implicit and explicit hate in this and compared these new labels with the existing ones.

Methods

We train binary classifier over standard hate data from (1). We evaluate binary classification performance over such in-domain and cross-domain data from (2) and compare it to benchmark models. Then we use our learning to develop a three-class classifier and perform inference over unlabelled cross domain data collected from Reddit.

Standard Datasets

For training and testing our binary classifier, we use data from Latent Hatred paper (1) which is a collection of twitter data annotated and cleaned for the purpose of implicit and explicit hate speech detection. It is a dataset of over

Dataset	Testing Dataset	
Labels	Hate	Not Hate
Total Posts	359	639
Avg Character Length	114.63	110.31
Max Character Length	3347	228
Min Character Length	12	10

Figure 3: Statistics for the Ethos dataset.

twenty one thousand tweets labelled for Implicit, Explicit and Non Hate categories. Examples of implicit and explicit hate present in this dataset are shown in Fig 1 and detailed summary statistics for this dataset are shown in Fig 2. This includes the average, minimum and maximum character length for all posts.

We also make use of another standard binary hate classification ETHOS dataset (2). It is a collection of 998 Youtube and Reddit's comments which are hateful in nature. It is used to see the application of our model in a cross-domain environment over hate classification task. Its summary is presented in Figure 3.

Dataset	r/Europe	e/Politics	r/ChampionsLeague
Dates	15.02.22 - 05.03.22	8.08.20- 15.01.21	15.02.22 - 5.03.22
Number of Days	18	158	60
Avg Character Length	96	88.18	86.55
Max Character Length	432	363	297
Min Character Length	14	7	8

Figure 4: Statistics for Reddit Data Collection

Data Collection

For 3-way classification, apart from ETHOS data, we collected our own Reddit data using PushShift (9) APIs for three different subreddits: r/politics, r/europe and r/championsleague. For each of the subreddits, we collected top 100 posts each day which are determined by the total number of interactions (upvotes and comments) over each post. A summary of statistics for them is shown in Fig 4 with examples given in Fig 5. We collect data for all three subreddits around respective global events.

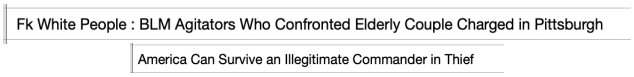


Figure 5: Example Explicit and Implicit Hate in Reddit Data

Figure 6: Word Graph obtained from r/Europe.

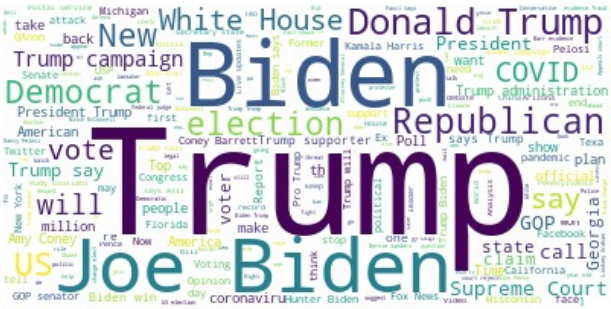


Figure 7: Word Graph obtained from r/Politics.

Analysis Setup

To analyze the change in hate with time, we have collected data from r/europe for several days around 21st February 2022, when the Ukraine-Russia conflict started. A word graph for the same is shown in Fig 6 which shows the relevancy of such data to analyzing the hate due this conflict.

Similarly we collected data for around five months surrounding 8th November 2020 from r/politics and for 2 months around 29th May 2021 for r/championsleague. A quick look at the word graphs as shown in Fig 7 and Fig 8 shows that US Presidential Elections and the Champions League finals were the most important topic of discussion in respective subreddits. We analyze the amount and type of hate in these subreddits and how it changes over time.

To train and test a binary hate classification model, we first use the (1) data. Since the overall dataset is heavily class imbalanced, we use a smaller subset of it with a $2:2:1$ split between the Implicit Hate : Explicit Hate : Non-Hate posts as shown in Fig. 9. A brief summary of this subset of data is given in Fig 10

Hypotheses

To further run the experiments on our dataset collected we need a model which will be trained on a standard dataset for 3 way classification - implicit, explicit and not hate. For this we intend to use a pre-trained BERT model. However, before running the analysis on any data, we laid a set of hypotheses. These hypotheses are laid in accordance with the Word-graphs and the world events at the time. Please note that the data has been collected during the time that specific events were happening. The timeline of this data might not

Figure 8: Word Graph obtained from r/ChampionsLeague.

be the same for all the collected data, but the interval for the collected data is similar (approximately 1 month). The next paragraphs describe our Hypotheses, before running the model.

Champions League tournament contains teams with zealous supporters of these different clubs. These supporters defend their teams furiously and would not care to post explicitly toxic content against the haters of this club. Also, the subreddit is followed by these sports fans who might not have put any moderation in their platform. Finally, this reddit is followed by people from all over the world, who speak different languages, and are explicitly toxic in expressing their views in those languages. Hence, we expect to see a more explicit hate than implicit for r/championsleague.

Similarly, more implicit hate is expected in r/politics as these accounts is also followed by diplomats. We expect the moderation policies in this to be very proactive. This account is not only followed by politicians but also famous celebrities, and elderly people. Also, it might act as a source of news to people, who are only connected to the world through Reddit. This would lead to the content on this subreddit being monitored, and any explicitly hateful content being removed as soon as it is detected.

For r/europe, we would expect a mix of implicit and explicit hate. This subreddit is followed by people who reside

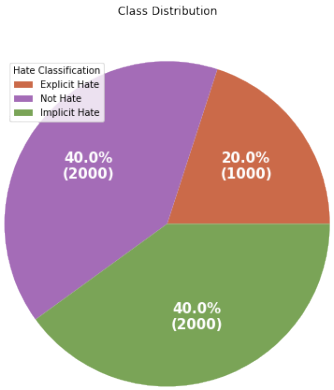


Figure 9: Training Dataset Distribution

Dataset	Training Dataset		
Labels	**Implicit**	**Explicit**	**Not Hateful**
Total Posts	2000	1000	2000
Avg Character Length	96	88.18	86.55
Max Character Length	432	363	297
Min Character Length	14	7	8

Figure 10: Training Data Statistics

in European countries and it has been very active since the Russia-Ukraine conflict. As visible in the word-graph, the conflict is the event that is being talked about, the most. Moreover, we expect less hate in general, and people being more supportive towards distressed Ukrainians in the conflict. Most of the countries are extending help towards the people in danger, providing them shelter, food, etc. and taking care of their needs till it is safe to go back.

Temporal analysis should highlight these world events that we have targeted. It is to be noted that we do not intend to provide any improvements over the State-of-the-Art models. Our aim is just to perform a qualitative and quantitative analysis on Hate speech.

Models

We have used a BERT based model and fine tuned it to classify hate and compare its performance to the benchmark models presented in (1). A detailed Flowchart describing the step by step training Pipeline for developing our model is shown in Fig 11.

We fine tune the pre-trained BERT model with 'Bert-base-uncased' configuration for the number of labels that we want to classify it for. The fine-tuning process takes input of 60% and 20% of the dataset as train and validation data-loaders. We calculate categorical accuracy for the different labels that exists. Further we also calculate the macro and average F1-score to ensure high precision and recall. We run this experiment for 10 epochs and save the model using torch.save() API, which has the best macro-F1 score in that particular epoch. The saved model is then used to calculate the test accuracy on 20% of the held-out dataset. This test accuracy and F1-score is what we share in our results.

Two-Way Classifier

Our benchmark paper clearly state that it was able to achieve 78% accuracy with 0.68 F1 score on the binary classification for Implicit vs Not Hate dataset. We intend to train our model as close as possible to the benchmark research paper for Latent Hatred (1). Hence, we tune our BERT model for this 2 way classification and reuse the hyper parameters used to fine tune BERT model for 3 way full scale classification.

The pipeline followed for binary classification is same as for three-way classification. Here, we run 10 epochs with our training and validation dataset. When we see a better macro-f1 score we save that model for future use. We have seen the model provide best results around 5-6 epochs. We have tuned learning rate and batch size to achieve the best

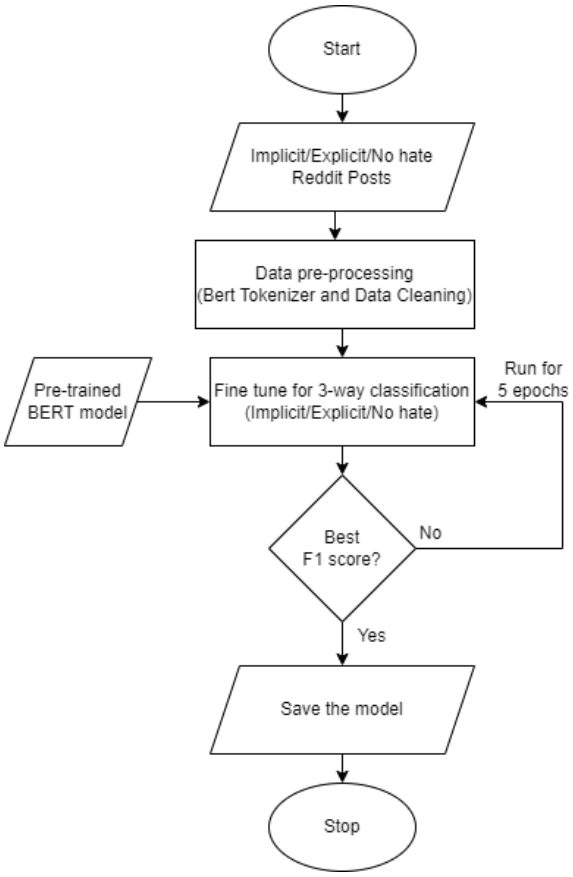


Figure 11: Model Architecture for 3-way classifier

possible results. Learning rate was experimented with three values - [1e-5, 3e-5, 5e-5] and batch size - [8,16,32,64].

Further, as mentioned, to handle the unbalanced dataset we have sub-sampled from the training dataset to maintain 2000 Implicit posts, 1000 explicit posts and 2000 Not Hate posts.

Three-Way Classifier

Once we achieve the hyper parameters after tuning the binary classifier for best macro F1 score. We again use these values to now train the BERT model. This model is then trained on sub-sampled balanced dataset. When introducing the balanced data we aim to achieve a higher F1 score with higher accuracy.

To further make sense of the implicit and explicit labels against the sentiment score that could be provided we used VADER to calculate the how positive and negative a post is. The output shows the probability of the post being positive, negative or neutral.

Some extensions to our model to boost the performance could have be to use transfer learning techniques. Our data was sparse and has a very low count of explicit posts, this hinder with our analysis for implicit vs explicit annotations. So, if we pick a fine-tuned BERT model which is aware of the standard binary hate classification technique might be

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

Figure 12: Model accuracy with balanced data & fine-tuning

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

Figure 13: Results for classification on ETHOS Dataset

able to boost the performance aiding the identification of explicit hate. Another way which we could have used to optimise the accuracy was by data augmentation techniques - as suggested by benchmark paper.

Results

Experimental Setup

The balanced training Data of 4K posts for Binary classification was broken into an 60:20:20 Train : Validation : Test Split along with a learning rate of 5e-5 and batch size of 64. Our binary classifier accuracy improved from 68% to 72% using balanced data and our F1 score of 0.71 outperformed the benchmark model in (1) as seen in Figure 12.

Fig 14 shows three different learning curves for different learning rates mentioned above. We can see that the best macro F1 score is for learning rate 5E-5.

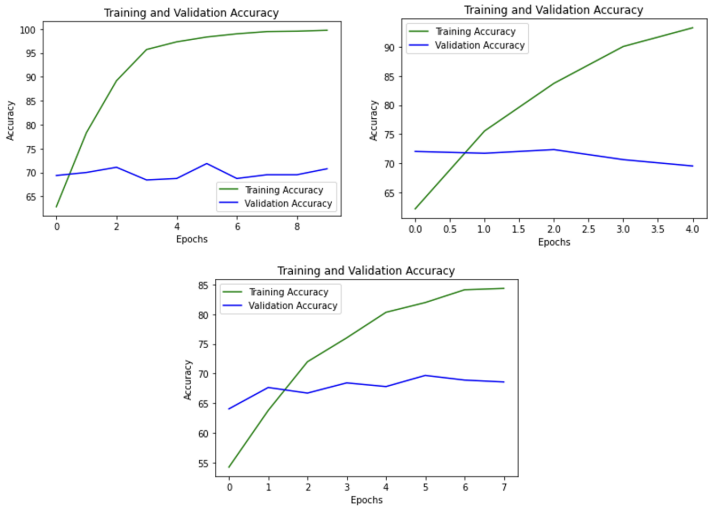
We used our learning from the binary classifier and develop a BERT based 3-way classifier over the 5K training data from (1) with the same Hyper parameters and were able to obtain 65% accuracy and 0.64 F1-score.

Manual Inspection

For verifying results of our inference over self collected Reddit data, we took approximately equal number of non-hate and hateful posts as shown in Fig 15. After deciding on a strict annotation manual, three annotators performed the labelling over 275 posts and were able to obtain a Fleiss Kappa Metric of 0.61 over three annotator conformance. The process used for manual annotation and analysis is shown in detail as a flowchart in Fig 16.

Qualitative Analysis

Fig. 13 shows the distribution of three way classification for a standard dataset which has been used by (2). Here, we are trying to evaluate how well our model is able to capture the implicit hate when we already know the hate and not hate labels. We were able to achieve a 70.3% accuracy for this dataset.



Learning Rate (AdamW)	Not Hate (F1 score)	Implicit Hate (F1 score)	Accuracy (test split)	Macro F1 score (test split)
5.00E-05	0.71	0.74	0.72	0.73
3.00E-05	0.69	0.74	0.72	0.72
1.00E-05	0.7	0.71	0.71	0.71

Figure 14: Learning curve for different learning rates - [5e-5, 3e-5, 1e-5] (left to right). Summary of test metrics is given in table below.

The results obtained clearly shows that the model has performed really well in identifying the implicit hate which was earlier labelled as not hateful. Out of not hateful posts 46% were identified as implicit. Similarly, out of all hateful posts 46% were identified as explicit and 48% as implicit, but only a very few 5% were identified as not hateful. We understand the model still has the scope for improvement in accuracy.

Further, we focused on analysing that what examples were classified as implicit hate but had a not hate label previously. We performed a word frequency analysis and found the most popular words as - *"people", "women", "white", "right", and even "f***"*. We were also able to find a few examples which correlated with the most found words. A few examples are - *"The world before **white** people. 1 word 8 letters. PEACEFUL"* - This is a hateful post which might not depict any hate by just looking at words but has a deeper meaning which was offensive to a target community. Our model was able to identify it as implicit hate. Similarly, another example was - *"He is complaining his **rights** are violated well maybe he shouldn' have come uninvited in our country"* - This example again is offensive for immigrants of a nation but does not use any explicit offensive words.

So, these results increase the confidence in our model, where we were able to successfully establish the validaity of our three-way classification.

Type of Posts	Number of Posts
Not Hateful	132
Implicit Hate	127
Explicit Hate	16
Total	275

Figure 15: Summary of Data for Manual Labelling.

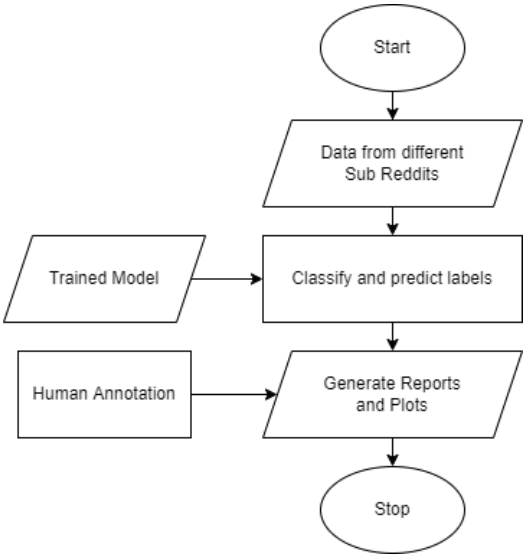


Figure 16: Flowchart for data analysis techniques.

Results Analysis and Comparison

Here we aim to analyse the implicit vs explicit hate content within subreddits, and understand the cycle of each hate category with time. We hypothesized that r/politics being focused on political issues should be rich in implicit hate posts. Similarly, a game event where fans support their preferred team need not hide their hate through indirect comments, hence, r/championsleague should have explicit hate rich content. We picked r/europe to understand hate changing through the time due to recent Russia-Ukraine conflict.

We expected the sentiment analysis of a specific post to depend on the type of hate displayed on the post. We conducted the analysis expecting a higher positive sentiment for implicitly hateful posts and a higher negative sentiment for explicitly hateful posts. To our dismay, the results from the sentiment analysis did not concord with the type of hate. This means that posts could receive positive sentiment, regardless of them being explicitly or implicitly hateful. However, as shown in fig 17 we were still able to validate previous results for implicit analysis in order to confirm the model performance.

	implicit	explicit	nothateful
neg	13	9	25
pos	41	77	41
neu	1946	914	1934

Figure 17: Distribution of sentiment labels on training data

Temporal Analysis - r/europe

As expected, r/europe did show a mix of both with the hate increasing near February 21, 2021. This is when the Ukraine-Russia conflict gained heat. The graph in Fig. 18 for the amount of Hate observed in r/europe over the 2K posts has been very sporadic. before the model was fine-tuned, the explicit hate had been observed at only a specific point during the analysis.

Now, after fine-tuning the model, we could observe that even the Explicit hate has been distributed equally. Still, the overall toxicity of the subreddit remains very high with the peak being the beginning of the Ukraine-Russia conflict. Moreover, there was a reduction in the combined hate of the subreddit as people would have come forward to help the people in distress.

Listed below are some examples listed by our model. The explicit example is a simple music videos shared on the group, however, due to the meaning of Perkele(evil spirit), the post has been classified as hate. This is a shortcoming of explicit hate detection. The example of implicit hate here is very appropriate for the situation of war in Ukraine. Here the author indicates support for the war.

Explicit Hate - ”Perkele- Yellow and Blue p**k”
Implicit Hate - ”Raise your voice I want the liberation of Palestine I want the liberation of Kashmir I want the liberation of Ukraine”

Temporal Analysis - r/politics

The amount of implicit hate in r/politics was enormous, and the explicit hate was rare. Both kinds of hate decreased after 8th Nov 2020 which was when the elections got over. In the Fig. 19, we can see the inference results over the 15K posts from r/politics. This shows how hateful posts are higher in count before the election, that was held on 8 Nov 2020. Towards the left, we can see our analysis before fine-tuning the model, and towards the right, they analysis after fine-tuning the model.

As we can see there is a significant increase in the number of hateful posts detected by our model. Even still, the trend in these hateful posts remains the same. Most posts containing implicit hate have been trending around August, 2020 and there has been a drastic decline in the toxicity. The drastic decline could be devoted to the end of elections, and not enough activity (in general) on the subreddit. This implies that implicit hate has been propagated by the political parties in order to persuade the voters in a specific directions and manipulate their decisions.

Listed below are some examples listed by our model. These examples are visibly correctly labelled as explicit hate post have some disturbing words, where as implicit hate posts have indirect accusations.

Explicit Hate - ”Honestly I m glad we don t have that n***er Obama as president anymore God just ban me already”, or ”N***er”

Implicit Hate - ”Graham declares that Black people can go anywhere in South Carolina. Just need to be conservative”.

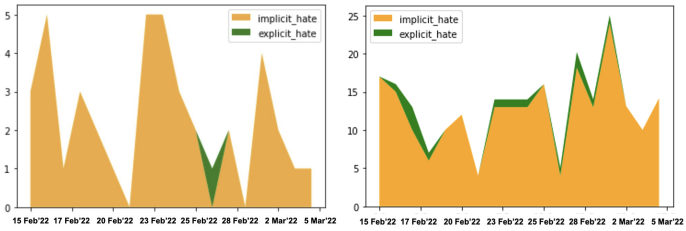


Figure 18: Implicit vs Explicit Hate for r/Europe over Time. Left: Before model tuning, Right: After model tuning.

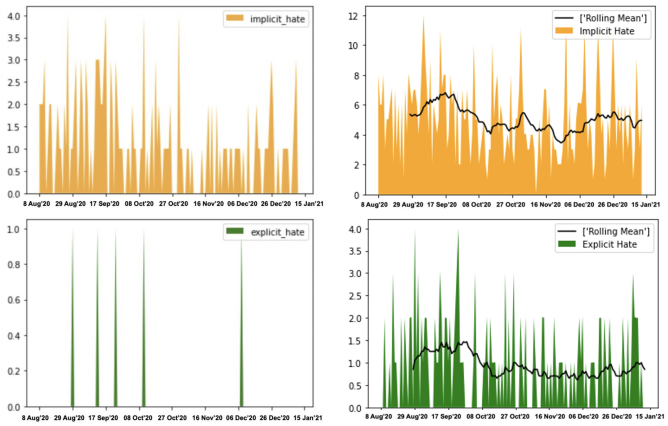


Figure 19: Implicit vs Explicit Hate for r/Politics over Time. Left: Before model tuning, Right: After model tuning. Moving average in right image is for 20 days.

Temporal Analysis - r/championsLeague

As expected, r/champions league demonstrated a lot of explicit hate out of all the subreddits. The amount of hate was the most near 29 May 2021 as that was the day of the tournament finals. This analysis had been conducted on over 6k posts and the results have been collected before the model was fine-tuned and after it was fine-tuned.

We can see that in Fig. 20, our previous analysis stated that the Explicit hate has been non-existent until a specific point in time (May 29). But, through our fine-tuned model, we were better able to identify the hate on the subreddit. According to this analysis, the hate has been present throughout the timeline of our dataset, but has been at its peak near the finals where both the competing teams English teams

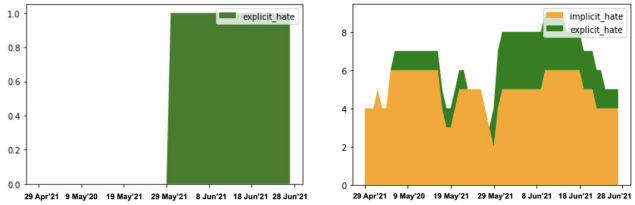


Figure 20: Implicit vs Explicit Hate for r/ChampionsLeague over Time. Left: Before model tuning, Right: After performing model tuning

with prior rivalry. As the results were announced the losing team’s supporters might not be satisfied with them, which could have also led to an increase in Hate comments.

Listed below are some examples listed by our model. There is only explicit hate example as no post was classified as implicit hate in the dataset. This again is clearly accusing the players for being defeated.

Explicit Hate - ”Posh c**ts can’t finish drink between’em.”

Work Division

All of the three team members have contributed equally to the overall tasks. Each one has contributed equally to literature surveys, participated actively in brainstorming sessions and has performed manual annotations for the data subset. Members have pair-programmed with other members, for their assigned tasks.

Abhay and Himanshu worked on setting up the code repository and resources, along with data collection through PushShift APIs as well as data cleaning.

Aarushi and Himanshu have worked on Feature Extraction, Model Training for a binary classifier and Validation of its results over (1) and (2) standard datasets.

Aarushi and Abhay have worked on developing the three-way classifier, performing regularization and evaluating its performance. Himanshu and Aarushi performed then performed inference over Reddit data and performed temporal analysis over them of implicit and explicit hate distribution across subreddits.

Conclusion

Through our project, we realized that most standard hate-not hate classifiers will label implicit hate as not-hate. This is because in order to spread hate implicitly, people use means such as sarcasm, white grievances and superiority complex, among others. Also, some people might frame a negative entity in a positive way which would be difficult to discern as hateful without knowledge of context. This is alarming as modern interactions on social media generally contain some positive hate, which might lead to conflicts and disagreements. From the graphs and our analyses, we could conclude that Explicit hate was always followed by a peak in implicit hate. This could be due to any real-world event inciting negative feelings in the masses, and them leading to expressing their hate online.

Moreover, most major events had led to a spike in hateful posts. This was observed in the analysis of all subreddits and could be derived by checking at the date of these spikes. For example, in r/championsleague, a spike in the toxicity was seen as that was the time near finals of the tournament. Furthermore, we could see that diplomatic platforms had more implicit hate than explicit hate. This could be due to moderation present on these platforms, which filters explicit hate as soon as it is posted.

Future Scope

Our work has been able to validate how the classification of hate can help with better understanding of toxicity on social platforms. Future work may include more analysis to comprehend the difference in implicit vs explicit ratio between conversations where authority lies on one hand. For example, in Wikipedia, when a request is made to publish a new article or make changes to an existing one, how do they check for Hate speech? Moreover, our cross-domain analysis shows that this dataset can perform well on other platforms as well. Also, data from different subreddits such as r/gaming can be included in the analysis. r/gaming is a subreddit where the people involved in video games post about their favorite games and setup, and others comment on it. This subreddit could contain more implicit hate as players would use sarcasm and metaphor to insult a specific video game, setup, etc.

The paper by ElSherief et al (1) categorized implicit hate into 6 categories. These 6 categories later helped in deriving a taxonomy for the Implicit Hate speech and its classification. Our project can be further run on each of these 6 categories and perform temporal analysis on all the subreddits discussed so far, to provide further insights. Understanding of temporal analysis shows that explicit hate always follows implicit hate. Therefore, one can just monitor the amount of implicit hate present on different platforms and expect explicitly hateful posts. Finally, one can collect data from some other subreddits which are a global platform for people with different (and even conflicting) interests and continue work in the same direction.

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. https://arxiv.org/pdf/2109.05322.pdf.

[2] Mollas, I., Chrysopoulou, Z., Karlos, S., Tsoumakas, G. (2020). Ethos: an online hate speech detection dataset. arXiv preprint arXiv:2006.08328.

[3] Mozafari, M., Farahbakhsh, R., Crespi, N., (2019) A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media. Social and Information Networks.

[4] Warner, W., Hirschberg, J. (2012, June). Detecting hate speech on the world wide web. In Proceedings of the second workshop on language in social media (pp. 19-26).

[5] Phadke, S., Mitra, T. (2020, April). Many faced hate: A cross platform study of content framing and information sharing by online hate groups. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-13).

[6] Zhang, Z., Luo, L., (2018) Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter. arXiv:1803.03662v2 [cs.CL]

[7] Benikova D., Wojatzki M., Zesch T. (2018) What Does This Imply? Examining the Impact of Implicitness on the Perception of Hate Speech. In: Rehm G., Declerck T. (eds) Language Technologies for the Challenges of the Digital Age. GSCL 2017. Lecture Notes in Computer Science, vol 10713. Springer, Cham. https://doi.org/10.1007/978-3-319-73706-5_14

[8] Mathew, B., Illendula, A., Saha, P., Sarkar, S., Goyal, P., Mukherjee, A. (2020). Hate begets hate: A temporal study of hate speech. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW2), 1-24.

[9] Push-Shift API documentation: https://github.com/pushshift/api

[10] Gelber, K. (2017). Hate Speech-Definitions Empirical Evidence. Const. Comment., 32, 619.

[11] Gao, L., Kuppersmith, A., Huang, R. (2017). Recognizing Explicit and Implicit Hate Speech Using a Weakly Supervised Two-path Bootstrapping Approach.

[12] Sun, C., Qiu, X., Xu, Y., Huang, X. (2019, October). How to fine-tune bert for text classification?. In China national conference on Chinese computational linguistics (pp. 194-206). Springer, Cham.

[13] Bauwelinck, N., Lefever, E. (2019). Measuring the impact of sentiment for hate speech detection on Twitter. no. c, 17-22.