Are You Sure You Want to Post This? Analyzing and Classifying Implicit Hate Speech

Ramón Carreño and **Chih-Ying Huang**
University of the Basque Country (UPV/EHU)
{rcarreno001, hchihying001}@ikasle.ehu.eus

Abstract

The increase of hate speech in social media platforms has motivated NLP practitioners, institutions and companies to develop systems aimed at detecting and classifying such content. However, much of this work tends to focus on hate speech that immediately stands out given its level of explicitness, overlooking what lies beneath the surface but still can inflict damage. The latter is what we refer to as implicit hate speech: content that conveys hateful messages in subtler ways - mostly devoid of explicit derogatory language. In this report we will describe in detail the nature of this implicit hate speech, how we trained models to identify and classify it in a more fine grained taxonomy, and analyze some of the challenges arising from this task.

1 Introduction

Amongst many other factors, internet anonymity inevitably gives way to a prevalence of hate speech all over social media. Recent advancements in text analysis and processing have allowed social media platforms to better detect harmful content by automatically banning users that post such content or directly deleting it, providing warnings indicating what they are going to post can be dangerous or misleading, et cetera.

This scenario prompts hateful users to make use of more subtle, *coded* language in hopes of getting their message across without being detected by such systems, effectively taking part in what is called *implicit hate speech*. This type of hateful content poses a new challenge in the current NLP landscape, since it frequently achieves their goal - most current state of the art models struggle understanding certain information (such as coded hate symbols) or ways of conveying it (using irony, for example) where humans easily can, making it easy for hate groups to spread dangerous content aimed at clear targets.

Using several datasets compiled by ElSherief et al. (2021), we will first analyze how is implicit hate speech different from the commonly analyzed hate speech. Subsequently, we will fine-tune a BERT-style model for two text classification tasks: first, a binary classification in order to check to what extent implicit hate speech can be discerned from non-hateful posts and secondly, a multi-class classification of the implicit texts according to the proposed taxonomy so it can be analyzed which formats of posts current models struggle detecting. As a bonus, we will analyze the challenge of irony detection and test alternative representations so language models can understand its nuances better. The language used in all of our test datasets is English.

2 Related Work

Although the concept of implicit hate speech was coined by Waseem et al. (2017), which provided annotation guidelines for distinguishing between implicit and explicit posts, most of our work is based on the findings of ElSherief et al. (2021). Their research not only provided us with data to work with but also served us as a reference point for our study.

ElSherief et al. (2021) scrapped 4,748,226 tweets from the most prominent hate groups in the US and attempted to filter out explicitly hateful content with the HateSonar classifier by Davidson et al. (2017) among some other criteria. After this filtering process, they compiled and annotated two of our datasets of choice and with them trained several models for implicit hate speech detection and classification.

The first dataset comprises labeled data intended for implicit hate detection, while the second dataset exclusively contains implicit hate posts annotated at a sub-category level, following a fine-grained taxonomy proposed by the authors. Additionally,

they also built a third dataset including the intended targets of each implicit hate speech post as the result of a natural language generation task using fine-tuned versions of GPT and GPT-2, aiming for explainability of implicit hate speech.

Other authors such as Kennedy et al. (2020) produced datasets annotated according to the implicit vs. explicit criteria as well, which we deemed less complete than our final choice. Hartvigsen et al. (2022) also proposed a machine-generated dataset for implicit hate detection. With the appearance of more works focusing on this task, recently Ocampo et al. (2023) presented a benchmark for hate speech detection including "implicit" and "subtle" amongst its labels.

3 Datasets

As introduced in the previous sections, we used three different datasets from the Implicit Hate Corpus (IHC) by ElSherief et al. (2021), all of them available in .tsv format: *implicit_hate_v1_stg1_posts.tsv* for implicit hate detection, *implicit_hate_v1_stg2_posts.tsv* for fine-grained implicit hate classification and finally *implicit_hate_v1_stg3_posts.tsv*[1] which includes generated implicit hate targets. The latter was not used for any training or evaluation tasks.

All of the datasets consist of pairs of tweets with their respective class annotation. Insights stemming from our exploratory data analysis within each are presented thereupon.

3.1 IHC1 for Detection

This dataset comprises a total of 21,480 tweets annotated as either *not_hate* (with 13,291 examples), *implicit_hate* (7,100) or *explicit_hate* (1,089). Even though the filtering process discussed in 2 intended to eliminate the most overt hate posts, inevitably retained some content requiring annotation as explicit hate speech, following the guidelines outlined by Waseem et al. (2017).

Following the extremely imbalanced class distribution, especially owing to the underrepresentation of *explicit_hate*, we removed all the posts belonging to this class for training as it would only add noise in the classification. However, as a means to clearly illustrate the differences between implicit and explicit hate, we kept them for our corpus analysis with the Scattertext tool (Kessler, 2017). Available as a Python library, Scattertext presents in a scatterplot the prevalence of terms according to rank-frequency for each of two categories. Having three categories again, we compared only the combinations of two we deemed more interesting: naturally, *implicit_hate* vs *explicit_hate* and for further understanding of the ICH corpus, *implicit_hate* vs *not_hate*.[2]

The latter showed how most of the data is of political nature, owing to the authors' filtering strategy, since the top non-hateful terms are concepts associated with politics (american in particular) such as "alt right", "white house", "supremacists", "trump", hinting a possible origin from posts citing news sources rather than user opinions. On the other hand, the words mostly associated with implicitly hateful posts are either names of target groups - rarely any slurs (although we might consider some subtle exceptions, such as the frequent mention of "savages") or explicitly hateful denominations. One of the most *explicit* terms in this category is "white genocide" which is usually related to certain hateful groups, although arguably still requires a baseline level of political context to understand its usage connotations.

Regarding implicit versus explicit comparison, we observed again political terms for the implicit part ("illegal", "country") and interestingly enough, certain stopwords that can be associated with identity (e.g. "my", "our"). A first quick glance at the top explicit terms already reveals a lot - it is difficult to find something that does not qualify as an insult or that is not inappropiate language.

3.2 IHC2 for Fine-grained Classification

IHC2 consists only of 6,346 implicit hate tweets, annotated according to the taxonomy described in the IHC paper. The most prevalent category is *white_grievance* with 1,538 examples. It refers to the frustration over a minority's group perceived privilege and casting majority groups as the real victims, for instance the post *'They say they are anti-racist but what they really are is anti-white!'*. Closely, *stereotypical* and *incitement* posts follow up in representation, whose names are self-explanatory.

The remaining categories, namely *irony*, *threatening* and *incitement* are certainly less represented, but not to the point to where a significant imbalance

[1] From now on, we will refer to each of them as IHC1, IHC2 and IHC3, respectively.

[2] Figures were only included in the associated Notebook, since simply hovering over them it can reveal way more information than a static screenshot.

must be compensated - exceptuating 80 tweets pertaining to the *other* category, which were stripped from training as it only includes noisy posts that did not fit well any of the proposed classes.

3.3 IHC3 for Target Insights

Finally, we analyzed the last dataset: implicit hate tweets annotated with the perceived target using a bubble plot. Since annotations came from a generative model, they were often inconsistent, so we organized them based on their most semantically related target groups after converting every target to lowercase. For instance, terms such as "black people", "black men", "black folk", "black", "black folks", "blacks," and "africans" were all aggregated into the *black people* category.

We observed *white people* emerged as the most mentioned target group. This might seem surprising, but it is explained by the fact that in this implicit context *target* does not necessarily mean *target of actual hate*. Upon further examination of the data, it can be seen they are typically not the direct recipients of hate; rather, they are often mentioned within hateful comments - i.e. *'We must secure the existence of our people and a future for white children'*.

Additional commonly mentioned targets identified in the data include minority groups such as immigrants, black people, jews, women, and LGBTQ+.

4 Classification Tasks

After our preiliminary data analysis, we carried out the two text classification tasks. Initially, a data cleaning stage was proposed, removing parentheses, quotes, emojis, hyperlinks and other kinds of problematic or difficult to parse symbols. Ultimately, we proceeded to do the training without any preprocessing other than removing unbalanced noisy columns (*explicit_hate* in IHC1 and *other* in IHC2) since certain features can end up being predictive of a certain category in this implicit context.

The model used for both classification instances was HuggingFace's DistilBERT (Sanh et al., 2020), a lightweight version of the ubiquitous encoder-only transformer model BERT, used for many common NLP classification tasks. Fine-tuning DistilBERT conveniently fits our time and resource limitations without giving up too much on performance from the original checkpoint.

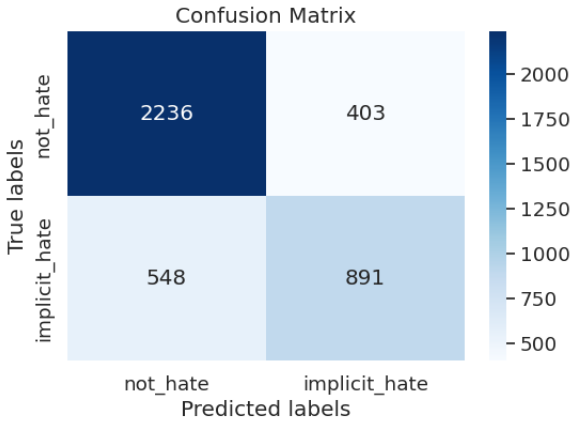

Figure 1: Confusion matrix obtained after binary classification

4.1 Implicit Hate Speech Detection

IHC1 (see 3.1) was used for this task. We imported the *distilbert-base-uncased* checkpoint of DistilBERT and conducted a preliminary tokenization of the entire dataset. This approach was adopted to save up on memory resources and reduce training time, considering that the default maximum token limit for BERT is 512. Anticipating that the average tweet would not require such a high token count, we determined that the longest tweet indeed contained 189 tokens. Therefore, we rounded up to the nearest power of 2, establishing a *max_length* of 256.

Then, we ran the definitive tokenization and instantiated a TensorDataset object with the resulting identifiers, attention masks and class labels, which was divided using a 60-20-20 split as our reference authors did.

Moving on to the training process, the train split was divided into 765 batches of 16 examples each with a DataLoader to further save up on computational resources. Moreover, an AdamW optimizer with a learning rate of 2e-5 was implemented. We found that just after 4 epochs, our data was overfitting on the training set, even though validation (and consequently) test accuracy did not improve after the first epoch. This reflects how little data is necessary to fine-tune a pre-trained transformer model for a text classification task.

Looking at the resulting confusion matrix in Figure 1, we can see the model clearly struggles detecting actually implicitly hateful tweets; almost two fifths of them were misclassified, much less for the opposite class. Table 1 directly compares our fine-tuning results with those achieved by ElSh-

erief et al. (2021), being slightly worse but still acceptable.

It's worth noting that we intended to illustrate the relationship between IHC1 and IHC2 more cohesively by showing a distribution with the implicit hate categories of the actual posts that were incorrectly predicted as either *implicit_hate* or *not_hate*, hoping for remarkable correlations to arise. Unfortunately, due to limitations with Google Colab GPU, we were unable to carry out this additional analysis.

Models	P	R	F1	Acc
ElSherief et al.	72.1	66.0	68.9	78.3
Our DistilBERT	68.9	61.9	65.2	76.7

Table 1: Comparison of our binary classification results vs. those of the original IHC paper

4.2 Categorizing Implicit Hate Speech

For this second classification task, IHC2 (see 3.1) was used, with the same original *distilbert-base-uncased* checkpoint. Here, we employed a slightly different approach by training the model directly with the HuggingFace Trainer. This involved converting all data into a HuggingFace Dataset object, similar to what we did in Lab 7 of this course. With respect to the hyperparameters, this time we proceeded again with a learning rate of 2e-5 but only trained for 3 epochs.

Results were again similar to our reference (Table 2), achieving a slightly worse accuracy but slightly better (macro) precision, recall and F-scores. Confusion matrix in Figure 2 revealed most notably the model sometimes struggles discerning *white_grievance* with *incitement* and *inferiority* language, which is understandable given the complexity of disambiguating among the proposed taxonomy, even for humans.

For instance, consider the sentence *'I think white people are waking up in unprecedented numbers'*. The sentence was originally annotated as *white_grievance* since it implies white people are the current victims of something or are (exclusively) experiencing hardship. However, in this same sentence there is also a very clear subtext of incitement (*'waking up'*).

Models	P	R	F1	Acc
ElSherief et al.	59.1	57.9	58.9	62.9
Our DistilBERT	61.0	59.8	60.2	60.1

Table 2: Comparison of our implicit hate classification results vs. those of the original IHC paper

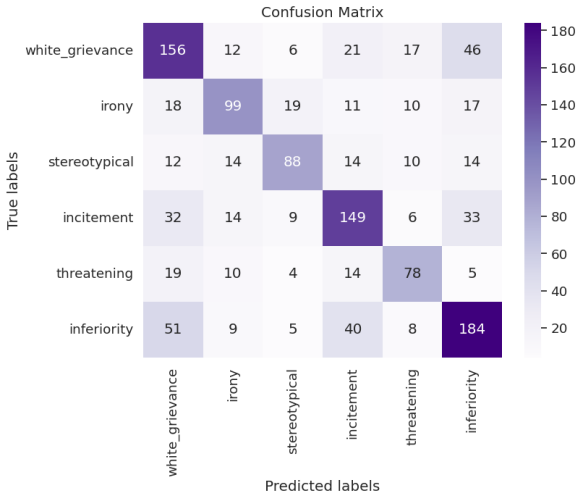

Figure 2: Confusion matrix obtained after implicit hate classification

5 Analyzing and Improving Irony Detection

By looking at the errors commited by the model, ElSherief et al. (2021) also reported several challenges to face when dealing with implicit hate speech.

Inspired by the latter, we decided to test for inference a checkpoint of RoBERTa (Liu et al., 2019) fine-tuned for irony detection (HuggingFace link). For this task, we predicted irony level using 797 posts from ICH2 tagged as irony, and observed only 45% of them were classified as more ironic than unironic. Interestingly, we observed the sentences where RoBERTa usually had more trouble detecting irony where those posed as direct questions.

After transforming to affirmative all the sentences that were incorrectly detected as *not_irony* and rerunning the inference, we found that 49% of them were now correctly detected as *irony*. We also observed removing question mark symbols "?" already improved irony detection. Such results can possibly shed light on the failure to detect some implicitly hateful content. Training with more examples that contain implicitly hateful content in direct question form could possibly improve detection.

6 Conclusions

Implicit hate speech, although not a novel concept, presents a significant challenge in combating hateful content on social media platforms. Often overlooked in the literature, in this report we provide a comprehensive analysis of its nature.

Due to its subtle nature and limited representation in training data, even current text classification systems encounter difficulties in distinguishing implicitly hateful text from benign content. Regardless, we have demonstrated it is possible to fine-tune state of the art models with consumer grade hardware and still obtain promising results.

By examining model errors, we can outline directions for future work based off challenges inherent to this task, such as achieving positive detections in spite of the presence of irony or coded symbols commonly making appearance in this type of content.

References

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Koombs, Shreya Havaldar, G J Portillo-Wightman, Elaine Gonzalez, Joe Hoover, Aida Azatian, Alyzeh Hussain, Austin Lara, Gabriel Olmos, Adam Omary, Christina Park, Clarisa Wang, Xin Wang, and Morteza Dehghani. 2020. The gab hate corpus: A collection of 27k posts annotated for hate speech.

Jason Kessler. 2017. Scattertext: a browser-based tool for visualizing how corpora differ. In *Proceedings of ACL 2017, System Demonstrations*, pages 85–90, Vancouver, Canada. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Nicolás Benjamín Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. An in-depth analysis of implicit and subtle hate speech messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1997–2013, Dubrovnik, Croatia. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.