An In-depth Analysis of Implicit and Subtle Hate Speech Messages

Nicolás Benjamín Ocampo[1], Ekaterina Sviridova[1], Elena Cabrio[1], Serena Villata[1]
[1]Universite Côte d'Azur, CNRS, Inria, I3S, France
nicolas-benjamin.ocampo@univ-cotedazur.fr, ekaterina.sviridova@inria.fr
elena.cabrio@univ-cotedazur.fr, villata@i3s.unice.fr

Abstract

The research carried out so far in detecting abusive content in social media has primarily focused on overt forms of hate speech. While explicit hate speech (HS) is more easily identifiable by recognizing hateful words, messages containing linguistically subtle and implicit forms of HS (as circumlocution, metaphors and sarcasm) constitute a real challenge for automatic systems. While the sneaky and tricky nature of subtle messages might be perceived as less hurtful with respect to the same content expressed clearly, such abuse is at least as harmful as overt abuse. In this paper, we first provide an in-depth and systematic analysis of 7 standard benchmarks for HS detection, relying on a fine-grained and linguistically-grounded definition of implicit and subtle messages. Then, we experiment with state-of-the-art neural network architectures on two supervised tasks, namely implicit HS and subtle HS message classification. We show that while such models perform satisfactory on explicit messages, they fail to detect implicit and subtle content, highlighting the fact that HS detection is not a solved problem and deserves further investigation.

1 Introduction

The rising mass of communication through social media further exacerbates harmful consequences of online hate speech. As a result, social media have faced mounting pressure from civil rights groups demanding to ramp up their enforcement of anti-hate speech policies, so that to monitor and limit this kind of content. In the latest years, numerous methods have been developed to automatically identify this type of utterances expressing hateful or abusive content on social media using Natural Language Processing methods. A variety of datasets have also been built, exemplifying various manifestations of this harmful content (Poletto et al., 2021). However, most of the research carried out so far on this topic has focused on overt forms of hate speech. Explicit hate speech is more easily identifiable by recognizing a clearly hateful word or phrase. Only recently, a few works (Hartvigsen et al., 2022; Wiegand et al., 2022, 2021a; ElSherief et al., 2021; Jurgens et al., 2019; Waseem et al., 2017) have started to focus on implicitness, where circumlocution, metaphor, or stereotypes are used to intentionally convey hatred towards a particular group. In those messages, hatefulness can be captured only by understanding their global meaning, as well as contextual information.

In this paper, we carry out an in-depth analysis of implicit HS in standard benchmarks for HS detection. Additionally, we define the notion of *Subtle HS* that puts forward hateful meanings elusively relying on human perception and through the use of complex syntactic structures. In our study, we collect messages from 7 available datasets for HS detection that cover different topics and are extracted from different social media platforms, and we enrich them with the following three-layer annotation: HS/non HS, Explicit/Implicit and Subtle/Non Subtle. We also provide a fine-grained annotation for implicit HS messages with 18 implicit properties such as irony, exaggeration, metaphor, and rhetorical question, among others. The newly created resource named ISHate (Implicit and Subtle Hate speech) provides a rich and variegate benchmark for pushing forward research on implicit and subtle hateful messages, and constitutes a challenging test-bed to evaluate computational approaches.[1] Additionally, we evaluate SOTA and competitive baseline classifiers to detect both implicit and subtle HS in ISHate, showing that current methods fail to effectively detect implicit and subtle HS messages due to their peculiar nature.

NOTE: This paper contains examples of language which may be offensive to some readers. They do

[1] The annotated corpora, and the accompanying annotation guidelines and software can be found at https://github.com/benjaminocampo/ISHate

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1997–2013
May 2-6, 2023 ©2023 Association for Computational Linguistics

not represent the views of the authors.

2 Related Work

In the latest years, there has been significant research on abusive language and hate speech detection using Natural Language Processing (NLP) methods (e.g., Xu et al. (2012); Dadvar et al. (2013); Poletto et al. (2021); Bohra et al. (2018); Corazza et al. (2020); Zampieri et al. (2019a); Caselli et al. (2020, 2021)). A few works focus on subtypes of HS, such as Warner and Hirschberg (2012) that tackles the recognition of antisemitism, or Waseem and Hovy (2016); Badjatiya et al. (2017); Gambäck and Sikdar (2017) that investigate predictive features to identify HS in the form of racism and sexism. In this context, several challenges and shared tasks have also been organized over the years, that made datasets and resources for multiple languages available (for a survey, see Poletto et al. (2021)). Research studies carried out so far have mostly focused on overt forms of hate speech, while very few works address the issue of implicit and subtle HS (ElSherief et al., 2021). However, several works show awareness of the problem. For instance, Warner and Hirschberg (2012) and Xu et al. (2012) discuss systems' limitations in identifying HS messages which are ambiguous, have patterns of emotional speech or lack context. Zhang and Luo (2018) and Corazza et al. (2020) highlight the complexity of recognizing hateful messages when the meaning is conveyed through sarcasm, stereotypes, complex syntactic structure, or non-explicit lexical patterns.

Among the few studies that attempted to address the issues of implicit and subtle detection, Caselli et al. (2020) defines a shared task to detect implicit and explicit abusive messages from AbusEval, a reannotated dataset based on OLID/OffensEval (Zampieri et al., 2019a). Benikova et al. (2018) paraphrases German HS tweets obtaining implicit and explicit messages to study classification methods. Dadvar et al. (2013) shows how taking user context improves cyberbullying detection with neither explicit profanities nor apparent neutral emotions. Jurgens et al. (2019) and Waseem et al. (2017) explain why explicitness, implicitness, and subtlety are typologies of abusiveness and encourage researchers to develop proactive technologies in this area. ElSherief et al. (2021) introduces a taxonomy of implicit hate speech and a benchmark corpus with fine-grained labels for each message. Hartvigsen et al. (2022) proposes a large-scale approach to automatically generate benign and implicit HS statements through the language model GPT3. Wiegand et al. (2021a, 2022) proposes resources to tackle implicitly abusive comparisons and abusive remarks, which are two subtypes of implicitness. Wiegand et al. (2021b) explains the key issues in the implicit abuse detection, as well as possible future directions to explore. Even though these studies set the basis to explore and model the issue of implicit HS, there is still large room for improvement, both in terms of creating adequate resources to investigate these aspects and in terms of computational approaches to address them.

3 HS, Implicitness and Subtlety

Hate Speech is defined as a direct attack against people – rather than concepts or institutions – based on protected characteristics (PC): race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and severe disease (Meta, 2022). We encompass in the concept refugees, migrants, immigrants, asylum seekers from the most severe attacks, and occupations when they are referenced along with PC, though commentary and criticism of immigration policies are excluded from HS. Following this definition, HS differs from related concepts mainly since it is specifically oriented to groups of people with PC as the main target (Poletto et al., 2021). In the following, we provide clear and operational definitions of explicit, implicit and subtle HS. Reported examples are extracted from the White Supremacy Forum (WSF) (de Gibert et al., 2018) and the CONAN (Chung et al., 2019) datasets, described in Section 4.

3.1 Explicit Hate Speech

Explicit HS is unambiguous in its potential to be abusive/hateful, such as language containing racial or homophobic slurs. Explicit HS uses words whose literal definition (taken from the dictionary) is hateful (ElSherief et al., 2021; Waseem et al., 2017; Caselli et al., 2020), as in Example 1.

1. *Negros are so dumb* . (WSF)

3.2 Implicit Hate Speech

According to the definitions of ElSherief et al. (2021), implicit HS does not immediately denote abuse/hate. Implicitness goes beyond word-related meaning, implying figurative language use such as irony, sarcasm, etc. (ElSherief et al., 2021; Benikova et al., 2018), generally hiding the real meaning, making it more difficult to grasp and undermining the collection of hateful messages (Hartvigsen et al., 2022; Waseem et al., 2017). Therefore, we deal with figurative meaning intended by the author rather than literal meaning of words that may be extracted from dictionaries.

In our extensive study on hate speech datasets, we identified 18 typical properties of implicitness relying on linguistics features (listed in Table 2). Among the most significant ones we have irony (Frenda et al., 2022), sarcasm (Potamias et al., 2020), black humor, metaphor (Choi et al., 2021; Gao et al., 2018), exaggeration (Troiano et al., 2018), rhetorical question, sentiment (Li et al., 2021), inference, lack of context (Dadvar et al., 2013), and lack of extralinguistic knowledge. Most of the implicit HS messages contain combinations of several properties, as in:

2. *Global Jewry did not want Brexit and are doing everything they can to sabotage it.* (CONAN)

Example 2 illustrates an interlacement of irony and exaggeration. Global Jewry represents a wordplay mixing Global Jury and Jews adding humorous content to the perception of Jewish people as a nation, while doing everything is hyperbolic pointing out that Jewish are extremely occupied by Brexit. The whole message is ironical as it emphasizes that Israel has its own interest in other country's affairs.

3. *I also , unfortuantely , spent 7-8th grade in a 95 % monkey house* (WSF)

In Example 3 there is an overlap of irony, exaggeration, and metaphor (underlined). *95 %* looks more like an exaggeration than a fact, while monkey house refers to the group of students and in the metaphor due to their skin color they are compared with monkeys.

4. *Are you sure that Islam is a peaceful religion?* (CONAN)

Example 4 interlaces rhetorical question, inference, and sentiment. The rhetorical question doubts about the peacefulness of Islam, expressing at the same time the author's take on it. As well as that, we may read negative inference between the lines.

3.3 Subtle Hate Speech

Subtle Hate Speech concerns hateful messages that are so delicate or elusive as to be difficult to analyze or describe, and that depend on an indirect method to deliver the meaning (Merriam-Webster, 2022). However, literal meanings (Roland et al., 1968; Bertrand, 1905) are of prime importance in subtle messages by contrast to implicit messages where we go beyond literal meanings (Waseem et al., 2017). Although implicitness and subtlety differ a lot at this point, we still rely on language users' perception (Grice, 1967, 1989; Korta and Perry, 2020) to understand implicit and subtle messages (Benikova et al., 2018). Despite the challenges in characterizing human perception schematically, in our study, the use of elements such as negations with positive clauses, conditionals, connectors, unrelated constructions, word order, and circumlocution can greatly affect the subtlety of a message. Cases of HS that convey the meaning directly and straightforwardly are non-subtle. As subtlety is even less studied than implicitness in the domain of HS (Breitfeller et al., 2019), and only considered as a related concept of implicitness in previous works (Jurgens et al., 2019; ElSherief et al., 2021; Hartvigsen et al., 2022), we analyze their interrelation, particularly whether one of these concepts is an umbrella term for the other.

First, we study whether explicit HS can convey the meaning in a subtle way, to accept or reject the idea that implicitness is a general term for subtlety. As can be seen from the example below, a message can be explicit and subtle at the same time:

5. *As a brit my knowledge of american law is somewhat lacking but even i know that this holder groid has committed treason* (WSF)

Example 5 is explicit as it literally conveys hateful meanings with the expression holder groid, but it is subtle as the intended hateful meaning is disguised by the first and second underlined constructions, pushing out the attention from the explicit hateful expression.

On the other side, we investigate whether subtlety is an umbrella concept for implicitness (whether all implicit messages are subtle by default). The following example shows an implicit non-subtle HS message:

6. *I 'm either in North Florida or Nigeria sometimes i can't tell the difference.* (WSF)

According to the definitions of explicitness and implicitness, this message is not explicit since it does not contain any hateful words, thus, it does not directly convey hateful meaning. However, it is hateful despite being more elaborated, as it includes certain figures of speech that trigger hateful meaning. The underlined expression in Example 6 represents antithesis, while the message itself is ironical. This message also lacks contextual information and extralinguistic knowledge. Although being implicit, this message is not subtle as the meaning is delivered plainly without any delusion or delicacy. Based on these considerations, we consider implicitness and subtlety as two separate concepts. To illustrate the difference between implicit non-subtle messages presented above and implicit subtle cases, consider the following example:

7. *I think it is a bit late to think to look after the saety and the future of white people in Southfrica* . (WSF)

We label it as implicit HS as it does not express hateful meaning explicitly via offensive words and it is ironic. It is also subtle as it starts with an introductory construction that makes the message more sophisticated and adds extra weight to it.

As can be expected, it is also possible to find non-hateful subtle or implicit sentences (such as ironic or sentimental texts). However, our work focuses on exploring implicitness and subtlety in the context of hate speech only, therefore those more general cases are not investigated.

4 The ISHate Dataset

Relying on the fine-grained definitions of HS provided in the previous section (explicit, implicit and subtle HS), we collect and enrich 7 available standard datasets for HS detection. As a result, we create the first benchmark for implicit and subtle HS detection on social media messages extracted from different sources.

4.1 Data Collection

Nearly all available resources of user-generated HS content are retrieved with a keyword-based approach, and mainly relying on a list of words with negative polarity (Poletto et al., 2021). However, with this strategy it is possible to extract mainly explicit HS expressions (as in the AbusEval dataset, Caselli et al. (2020)). Given that our study focuses on implicit and subtle HS, we prefer to explore resources collected from communities of users that are potentially prone to hate speech, or resources manually created using a systematic approach. In the following, we list the considered resources:

White Supremacy Forum Dataset (WSF) (de Gibert et al., 2018), that contains HS messages from Stormfront, scraped from the most influential white supremacist forum on the Web. The database is arranged in sub-forums and conversation threads.

HatEval (Basile et al., 2019), which is among the most well-known benchmark for HS detection. A combined approach is applied to collect hateful and misogynous tweets by monitoring potential victims of hate accounts, downloading the history of identified haters, and filtering Twitter streams with both neutral and derogatory keywords.

Implicit Hate Corpus (IHC) (ElSherief et al., 2021), annotated with explicit HS, implicit HS, and non-HS labels obtained from online hate groups on Twitter. The authors focused on eight ideological clusters of U.S., as Black Separatists, White Nationalist and Neo-Nazi. From this dataset we only extracted messages labeled as implicit HS, as it is one of our target categories.

ToxiGen (Hartvigsen et al., 2022), a dataset with benign and implicit toxic messages against minority groups. ToxiGen is machine-generated through the GPT3 language model and prompt programming. Similarly to IHC, we only extracted messages which were automatically labeled as implicit HS and human-validated as toxic by the authors. We did not consider unfinished generated sentences which make a part of implicit messages.

YouTube Video Comments Dataset (YouTube) (Hammer, 2017), that consists of YouTube comments posted under videos related to religion and politics. Differently from the other resources, the messages are annotated as “violent” or “clean”.

CONAN (Chung et al., 2019), a dataset of HS messages and counter-narratives (CN) pairs for CN generation. Two native English speakers were asked to write 50 prototypical short texts, which NGO could later use to write their hate texts and counter-narratives. We believe that messages for which a CN can be provided might be richer in implicit content since a slur-based explicit HS message might produce very poor argumentative CN.

Multi-Target CONAN (MCONAN) (Fanton et al., 2021), a dataset of English HS/CN pairs comprising several hate targets. It is collected using a Human-in-the-Loop approach. A generative language model is refined iteratively by using data from the previous loops to generate new samples that NGOs experts review.

Before starting the annotation process with the fine-grained annotations (Explicit, Implicit and Subtle HS), we had to make sure that the definition of HS originally used to annotate such resources is consistent with ours. In the first annotation round, we checked the messages originally annotated as HS, and discarded the few ones that did not correspond to the definition of HS reported in Section 3. For the YouTube dataset, we also added the HS labels. While all the messages annotated as HS are directed to PC, it should be noted that the topics distribution and the writing quality might be different, given the heterogeneity of the selected resources. HS messages mostly target Islamism, Judaism, misogyny, multi-culturalism, racism, immigration, and refugees. Regarding time creation, WSF is made from threads posted between 2002 and 2017, ToxiGen's LM was trained with messages from 2016 to 2019, Hateval consists of messages of 2018, the YouTube comments were collected in 2017, while the IHC contains tweets from U.S. ideological clusters from 2015 to 2017.

4.2 Annotation Procedure

Following the annotation scheme described in Section 3, four graduate-level annotators with linguistics and computational linguistics competences carried out a pilot annotation study on a sample of 100 messages extracted from each of the above mentioned resources to converge to non-ambiguous annotation strategies. We calculate the Inter Annotator Agreement (IAA) on this sample, resulting in Cohen's κ=0.793 (Cohen, 1960) for the implicit layer (binary annotation Explicit/Implicit) and 0.730 for the subtlety layer (binary annotation Subtle/Non-Subtle). We also compute the IAA considering both layers simultaneously, that is, considering one layer of 4 classes (Implicit, Explicit, Subtle, Non-Subtle), obtaining a Cohen's κ of 0.734. In the reconciliation phase, we notice that most of the disagreements are due to the interlacement of subtlety and implicitness. For that reason, we also calculate an ordered weighted disagreement using Krippendorff's α to penalize less when the annotators agree at least on one of the layers (Artstein and Poesio, 2008). The Krippendorff's α is 0.757. Despite the complexity of the annotation task, obtained results are considered as strong agreement in a two-annotators setting. The rest of the annotations has then been carried out by two of the annotators mentioned above, which were provided with the final version of the annotation guidelines (containing the definitions of the target classes, i.e., subtlety and implicitness, and a discussion about borderline cases), together with a set of labeled examples.

Finally, the implicit properties annotations are added on top of the messages labeled as implicit as an additional annotation layer to highlight 18 linguistic features that implicitly convey hateful meaning. For this layer, annotations are carried out by one expert linguist.

4.3 Data Statistics

Table 1 shows statistics of the final dataset, reporting on the number of annotated HS messages for each resource and for the three annotation layers.

The ISHate collection consists of a total of 29116 messages, where 11247 are HS (further annotated with the Explicit/Implicit and Subtle/Non-subtle labels). For computational purposes, we provide a dataset split in three subsets, i.e., *train* (70%), *validation* (15%), and *test* (15%) sets. Each of the partition respects the distribution of all the annotation layers using stratified splitting. As can be seen, classes are unbalanced, each resource providing only a reduced number of implicit and subtle messages - as expected. Note that CONAN and MCONAN do not contain Non-HS messages, because their main objective is CN generation. As for IHC and ToxiGen, we only look through previously annotated implicit HS messages disregarding non hateful ones. Note also that ToxiGen claimed to contain only implicit adversarial messages, but according to our definitions and annotation guidelines many messages are considered as explicit and non-subtle by our annotators.

Table 2 shows the full distribution of the implicit properties relative to the implicit messages in ISHate. As it can be seen, Inference (58%), Context (48%), Sentiment (45%), Exaggeration (28%) and Irony (22%), are the most frequent properties of implicit HS messages, whereas Euphemism (4%), Circumlocution (3%), Metonymy (0.4%) and Synecdoche (0.08%) are the least recurrent. Note that one implicit message can be labeled with more than one property.

	Train		Dev		Test		CONAN	HatEval	IHC	MCONAN	ToxiGen	WSF	Youtube
Label	#	%	#	%	#	%							
Non-HS	12508	.614	2680	.614	2681	.614	0	7421	0	0	0	9342	1106
Explicit HS	7007	.344	1501	.344	1501	.344	324	3107	317	3344	183	987	1747
Implicit HS	866	.042	186	.043	186	.043	81	110	300	295	170	173	109
Non-HS	12508	.614	2680	.614	2681	.614	0	7421	0	0	0	9342	1106
Non-Subtle	7691	.377	1648	.377	1648	.377	393	3191	614	3595	348	1018	1828
Subtle	182	.009	39	.009	39	.009	12	26	3	44	5	142	28

Table 1: Statistics on the annotated dataset (resources and label distributions for the two tasks)

	Implicit HS	
Implicit Properties	#	%
Inference	729	58.885
Context	602	48.627
Sentiment	569	45.961
Exaggeration	359	28.998
Irony	275	22.213
Extralinguistic knowledge	193	15.590
Black humor	144	11.632
Rhetorical question	134	10.824
Visual signs	122	9.855
Humiliation	115	9.289
Antithesis	97	7.835
Metaphor	93	7.512
Sarcasm	85	6.866
Fallacy	74	5.977
Euphemism	56	4.523
Circumlocution	41	3.312
Metonymy	6	0.485
Synecdoche	1	0.081

Table 2: Statistics on implicit properties distribution.

4.4 Data Augmentation

To overcome the problem of the unbalanced dataset, we propose oversampling and data augmentation (DA) methods for the minority classes using adversarial methods and generative models following (Mayer et al., 2020; Wei and Zou, 2019), and the GPT2 language model (Radford et al., 2019).

Replace Named Entities (RNE). It replaces a named entity (PER, LOC, ORG, and MISC) in the input sentence. A candidate NE in a sentence is replaced by another one according to a previously collected list of NEs (Mayer et al., 2020). Then, the most similar NE is selected by using pre-trained FastText embeddings (Bojanowski et al., 2016). In our use case, we notice that the number of NEs PER, LOC, and ORG are very few compared to MISC. This might be due to the fact that HS messages in our collection mostly target groups and not individuals. However, expressions like *muslims*, *jews*, or *blacks* are present in the MISC category and replaced, as in Example 8. As it can be noticed, the expression preserves its meaning, that is, the use of the rhetorical question to convey that Muslims are not considered as a part of society.

8. **Original**: *Have Muslims ever made a contribution to our society?* (CONAN)
   Augmented: *Have Moslem Arabs ever made a contribution to our society?*

Replace Scalar Adverbs (RSA). It replaces emphasising adverbs like *considerably* or *largely*, with another scalar adverb that might increase or decrease the emphasis of an adjective/verb.

9. **Original**: *I do like the outfit they dressed the simian in ; he looks so regal* (WSF)
   Augmented: *I do like the outfit they dressed the simian in ; he looks incredibly regal*

Example 9 is another implicit racist HS message that uses irony by combining *simian*, *dressed*, and *regal*. RSA replaces the scalar adverb, slightly increasing the emphasis of *regal* in the sentence.

Add Adverbs to Verbs (AAV). It adds modifiers to verbs to accentuate them. In this case, we use speculative adverbs like *certainly*, *likely*, *clearly*.

10. **Original**: *I wish we lived in a world in which only men could be doctors, lawyers, firemen, and teachers.* (MCONAN)
    Augmented: *I wish we lived in a world in which only men could plausibly be doctors, lawyers, firemen, and teachers.*

In Example 10, the meaning is made elusive by the use of the expression *I wish we lived*, the addition of *plausibly* alters the possibility of the verb "to be", but keeps the misogyny meaning.

Replace Adjectives (RA). It replaces adjectives with their synonyms, as in:

11. **Original**: *If you look at an Alaskan malamute and a timber wolf that are two different species , they look a lot more alike than say a Congolese and a Swede .* (WSF)
    Augmented: *If you look at an Alaskan malamute and a timber wolf that are two unlike species , they look a lot more alike than say a Congolese and a Swede .*

Replace In-Domain Expressions (RI). It replaces a list of manually-crafted expressions often used in HS messages (not captured by the RNE) with other semantically similar expressions, as in Example 12. Here the word "migrants" is replaced by "foreigners" without adding hateful connotations. It would have been different to replace it with *aliens*, which is a derogatory term to refer to migrants.

12. **Original**: *migrants are entering the u.s. just to obtain welfare and other benefits .* (MCONAN)
    Augmented: *foreigners are entering the u.s. just to obtain welfare and other benefits .*

Easy Data Augmentation (EDA). Given an input sentence, EDA randomly: *i)* replaces a non-stop word expression with a synonym using Wordnet; *ii)* inserts a synonym of a non-stop word in a random position; *iii)* chooses two words of the sentence and swaps their positions; *iv)* removes each word in the sentence with a certain probability. Only one of the four operations at a time is applied to a sentence.

13. **Original**: *A lot of white women are trying to create danger by drawing these other people into our countries .* (WSF)
    Augmented: *A lot of our women are trying to create danger by drawing these other people into white countries .*

A parameter α expressing the percent of the words to be changed in a sentence is specified (in our setting, $\alpha = 0.1$ as in Wei and Zou (2019)).

Back Translation. It translates an input message into a different language to translate it back into the original language. We use the translation from English to Russian following (ElSherief et al., 2021).

14. **Original:** *As a nation , we are in desperate need for an immigration time out and a zero tolerance policy for those here illegally .* (WSF)
    Augmented: *As a nation, we desperately need an immigration timeout and a zero-tolerance policy for those here illegally.*

Generative Models (GM). We fine-tune autoregressive generative language models with instances from our minority classes, i.e., explicit subtle, implicit non-subtle, and implicit subtle messages. To do so, we prefix this label on the text as a prompt. Then, language models are asked to generate messages starting with one of our fine-tuned prompts, as in Example 15. We use GPT2 (Radford et al., 2019) as a language model, fine-tuned for 4 epochs using learning rate of 3e-5, and batch size of 32. Additionally, we implement a human-in-the-loop approach revising the generated examples and reannotating them in case the original label is nomore appropriate for the message.

15. **Input:** *Explicit Subtle HS:*
    Augmented: *Explicit Subtle HS: In the end, it comes down to what women want from a man... If they want to play with whores, they can stay at home and have babies...*

Except for GM and BT, the same strategy is applied to augmentation methods to produce new messages. Preprocessing (e.g., Parts-of-Speech tagging and Named Entities Recognition) is carried out using Flair (Akbik et al., 2019) and NLTK (Bird and Loper, 2004) models, and allows to recognize possible candidate phrases to perform a replacement/addition. Then, a candidate phrase is perturbed by another one according to a list of adverbs, NEs, or adjectives based on domain data. We rely on FastText and WordNet Synsets to maintain the semantics of the augmented sentences with respect to the original one. The number of candidates to perform a replacement/addition and the number of replacement/additions per candidate are provided as parameters to these methods.

5 Evaluation

To show that implicit and subtle HS detection is still a very challenging task, we evaluate a set of state-of-the-art models for HS detection on the ISHate dataset. We propose two 3-label classification tasks:

- Task A (Non-HS/Explicit HS/Implicit HS)
- Task B (Non-HS/Non-Subtle HS/Subtle HS)

To this goal, we consider the following models:

Universal Sentence Encoder (USE) + SVM (Indurthi et al., 2019). First-ranked model on the HatEval benchmark (Basile et al., 2019). The USE (Cer et al., 2018) is a sentence embedding that encodes text into high dimensional vectors of 512 dimensions, trained on large data sources to provide an encoding method that works for various NLP tasks. An SVM classifier with RBF kernel and default parameters is then used for classification.

DeBERTa V3 (hate_speech18). SOTA model on the WSF dataset (de Gibert et al., 2018). For classification, a default HuggingFace implementation of a one-layer Feed Forward network is used on top of DeBERTa (He et al., 2021a,b), a transformer-based model. The model is later fine-tuned for 4 epochs (learning rate of 2e-5, batch size of 32).

BERT (Devlin et al., 2018). We use this language model to encode text sequences and classify them by adding a Feed-forward neural network on top.

HateBERT. A re-trained BERT model using over 1 million posts from banned communities on Reddit (Caselli et al., 2021) and then fine-tuned on our dataset. HateBERT obtained very promising results in the benchmarks HatEval, OffensEval (Zampieri et al., 2019b), and AbusEval (Caselli et al., 2020).

As for preprocessing, we replace long non-space character chains for only one occurrence, and delete digits, special symbols, and URLs.

5.1 Results

Table 3 reports on the results of the different models on the two tasks. On both tasks, all models show satisfactory performances when detecting overt forms of HS (Explicit HS and Non-Subtle HS classes), with DeBERTa outperforming the other models. The results obtained by all models for the Implicit HS and Subtle HS classes are much lower, and comparable to those obtained by ElSherief et al. (2021) (F1-score=.586) on the implicit class.

As a follow-up experiment, we apply the oversampling techniques (Section 4) on the minority classes of tasks A and B until balancing them with respect to the Explicit HS and Non-Subtle HS categories. The oversampling is performed on the training set only. The test set is the one of the original dataset, and is therefore unbalanced in order to evaluate the system on real class distribution and to avoid information leakage from train to test through augmentation methods. Tables 4a and 4b show the number of additional generated implicit/subtle messages and the resultant training set distribution per augmentation method, respectively.

Among all tested models, only HateBERT significantly improves its performance for detecting implicit messages combining all augmented data (ALL) (see Table 3). We also highlight that back translation (BT) better contributes to the performance on the implicit hate class for BERT, DeBERTa, and USE+SVM [2]. Performances surprisingly increase for the subtle class with USE+SVM+BT showing that back-translated messages provide diversity by rephrasing subtle examples without altering their meaning. Data generated with simpler augmentation methods as BERT+RNE and DeBERTa+RI also show slight improvements for subtlety. However, performances decrease on the implicit class when applying data augmentation strategies GM and GM+Revised, and only slightly improve on the subtle class.

5.2 Error Analysis

To gain insights into the models' behaviours, we manually analyse the classification errors of the best performing approaches, i.e., HateBERT+ALL and USE+SVM+BT for both tasks A and B. For the Non-HS/Explicit/Implicit classification, it is harder for HateBERT+ALL to differentiate implicit messages rather than explicit ones. Figure 1 in Appendix A shows the resultant embedding through t-SNE (van der Maaten and Hinton, 2008) approximation of all implicit messages of the test set, and two samples of the other two classes, showing how HateBERT+ALL is capable of separating the space in three different blobs for classification (Figure 1a), embedding well the human-annotated Explicit HS and Non-HS classes (Figure 1b), but badly encoding implicit annotated messages (instances spread over the explicit and non-hateful blobs). We also considered the F1-scores of the (Non-HS vs Implicit HS) and (Non-HS vs Explicit HS) classes, as if we had binary labels, obtaining an (F1-score=0.670) and (F1-score=0.850) on the target categories respectively, showing that implicit cases are harder to detect than explicit ones.

Among misclassified examples, we can find messages where the PC are not explicitly named (Example 16). Contextual knowledge is needed there to grasp that the target of the message is black people.

16. *I can picture him as that baby monkey on the ' save the rainforest comercial or whatever*

Additionally, from the misclassified messages we extract the most frequent implicit properties not captured by the classifier[3]. They concern Inference (53%), Context (41%), Sentiment (40%), Exaggeration (24%), Extralinguistic knowledge (24%).

For Subtle/Non Subtle message classification, we also plotted the USE embedding for the best

[2] The table reporting the obtained results by all models on different types of augmented data is in the Appendix.

[3] The full table can be found in the Appendix.

	Task A									Task B								
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Model	Non-HS			Explicit HS			Implicit HS			Non-HS			Non-Subtle HS			Subtle HS		
USE+SVM	.888	.866	.877	.766	.803	.784	.399	.382	.390	.891	.868	.879	.783	.832	.807	.667	.103	**.178**
BERT	.903	.893	.898	.81	.833	.821	.394	.371	.382	.902	.891	.897	.819	.846	.832	.250	.103	.145
HateBERT	.904	.89	.897	.811	.849	.829	.447	.382	.412	.903	.890	.897	.814	.850	.831	.143	.026	.043
DeBERTa	.927	.899	**.913**	.825	.880	**.851**	.467	.419	**.442**	.920	.893	**.906**	.823	.877	**.849**	.375	.077	.128
HateBERT+ALL	.903	.896	.899	.827	.827	.827	.502	.559	**.529**	.903	.881	.892	.816	.844	.830	.391	.462	.424
BERT+BT	.909	.887	.898	.824	.826	.825	.459	.608	.523	.898	.900	.899	.839	.832	.835	.304	.359	.329
DeBERTa+BT	.919	.885	.902	.830	.857	.844	.428	.543	.479	.920	.897	**.908**	.835	.876	**.855**	.385	.256	.308
USE+SVM+BT	.897	.856	.876	.782	.787	.785	.403	.645	.496	.892	.868	.880	.789	.831	.809	.739	.436	**.548**
BERT+RNE	.897	.897	.897	.807	.829	.818	.455	.349	.395	.899	.895	.897	.826	.839	.833	.400	.256	.312
DeBERTa+RI	.922	.894	**.908**	.821	.878	**.849**	.460	.398	.427	.910	.894	.902	.828	.860	.843	.364	.205	.262
HateBERT+GM	.901	.898	.899	.824	.827	.825	.414	.425	.419	.899	.898	.899	.831	.834	.832	.250	.231	.240
HateBERT+GM+R.	.905	.891	.898	.816	.835	.826	.408	.419	.414	.894	.898	.896	.826	.826	.826	.192	.128	.154

Table 3: Obtained results on tasks A and B.

Aug. method Label	RSA	AAV	RNE	RI	RA	EDA	BT	GM	GM+Revised	ALL
Implicit HS	6848	7032	828	817	467	6935	748	200	82	23957
Subtle HS	3192	3136	480	210	172	2912	179	200	204	10685

(a) Number of additional implicit/subtle messages generated by each augmentation method.

Aug. method Label	ORIG	RSA	AAV	RNE	RI	RA	EDA	BT	GM	GM+Revised	ALL
Non-HS	.614	.459	.456	.59	.590	.600	.458	.592	.608	.611	.282
Explicit HS	.344	.257	.256	.33	.331	.336	.257	.332	.340	.342	.158
Implicit HS	.042	.283	.288	.08	.079	.064	.286	.076	.052	.046	.560
Non-HS	.614	.531	.532	.600	.607	.609	.537	.608	.608	.608	.403
Non-Subtle HS	.377	.326	.327	.369	.374	.374	.330	.374	.374	.374	.248
Subtle HS	.009	.143	.141	.032	.019	.017	.133	.018	.019	.019	.350

(b) Train set distribution (%) per augmentation method (ORIG corresponds to the original train distribution).

Table 4: Statistics on the train set with data augmentation.

model on this task (Figure 2 in Appendix A). However, it can be seen that USE+SVM+BT could not differentiate correctly on the subtle notion despite of the results reported in Table 6b. Example 17 is not predicted as subtle. It shows how the word order may influence our understanding. At a first glance, the part how stupid the Jews seems to have a different meaning from what the phrase actually conveys if we read it entirely. We may also notice a circumlocution in the second part of the message.

17. *I am insulted by how stupid the jews think we are until i see what they see by reading the posts amongst our so called , ‘ ‘ awakened brethren .*

6 Conclusions

In this paper, we have presented ISHate, the first benchmark dataset annotated with both implicit and subtle HS labels, which represents a challenging test-bed to evaluate computational approaches. We also provide a fine-grained annotation for implicit HS messages with 18 implicit properties which represent the relevant features that HS classifiers should possess to improve implicit HS detection. It has been created enriching 7 existing datasets for HS detection over different topics and from different social media. We have shown that current SOTA models fail to properly detect implicit and subtle HS messages as peculiar features connected to Sentiment, Inference, Context and Irony, as well as complex syntactic structure, cannot be properly understood. We also investigated data augmentation strategies to increase the number of instances for the minority classes. We show that - while they cannot be the ultimate solution to the lack of implicit and subtle examples - they still play a role in improving the systems' performances, in line with ElSherief et al. (2021). As for future work, we plan to propose alternative large-scale methods to collect implicit and subtle messages by targeting "hateful" users, manual creation (Wiegand et al., 2021a, 2022) or refining human-in-the-loop generative methods as in (Hartvigsen et al., 2022). Also, we will investigate features modeling implicit properties (Wallace et al., 2014; Troiano et al., 2018; Frenda and Patti, 2019) and new model architectures for HS detection (Nejadgholi et al., 2022).

Limitations

The main limitation of this paper lies in the intrinsic difficulty to provide a clear definition of the notions of Implicit HS and Subtle HS (given the limited number of definitions available in the literature for these notions), and, as a consequence, to build annotated resources. Enhancing the ISHate dataset with new instances requires future annotators to be experts in computational linguistics trained on our annotation guidelines through pilot annotations to keep the same level of agreement. This restricts crowdsourcing-like options, making the resource building process more expensive. Moreover, the complexity of the messages and of the considered categories makes the process time-consuming (i.e., a trained annotator requires 30sec. for explicit messages and 1.30min. for implicit/subtle messages on average). Even opting for generative and synthetic data augmentation approaches, they still require human-in-the-loop intervention and high computational resources to generate Implicit/Subtle HS messages on a big scale.

Ethics Statement

This paper contains examples of HS from existing linguistic resources for HS detection and which do not reflect the authors' opinions.

While our purpose is to prevent and curate social media resources from HS, the release of this dataset might still pose a potential misuse case. However, we still consider that effective classifiers for this task are necessary to tackle implicit and subtle online hate on scale and prevent the spreading of this harmful content online. Our work aims at making a step towards that objective and encourages the scientific community to investigate these aspects.

Acknowledgements

This work has been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR- 19-P3IA-0002.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. *CoRR*, abs/1706.00188.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Darina Benikova, Michael Wojatzki, and Torsten Zesch. 2018. What Does This Imply? Examining the Impact of Implicitness on the Perception of Hate Speech. In *Language Technologies for the Challenges of the Digital Age*, pages 171–179, Cham. Springer International Publishing.

Russell Bertrand. 1905. On denoting. *Mind*, 56(14):479–493.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of Hindi-English code-mixed social media text for hate speech detection. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 36–41, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. 2016. Enriching word vectors with subword information. *CoRR*, abs/1607.04606.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding Microaggressions in the Wild: A Case for Locating Elusive Phenomena in Social Media Posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, Hong Kong, China. Association for Computational Linguistics.

Miguel Casas Gómez. 2009. Towards a new approach to the linguistic definition of euphemism. *Language Sciences*, 31(6):725–739.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I Feel Offended, Don't Be Abusive! Implicit/Explicit Messages in Offensive and Abusive Language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *CoRR*, abs/1803.11175.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor Detection via Contextualized Late Interaction using Metaphorical Identification Theories. Number: arXiv:2104.13615 arXiv:2104.13615 [cs].

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. A multilingual evaluation for online hate speech detection. *ACM Trans. Internet Technol.*, 20(2).

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. In *Proceedings of the 35th European Conference on Advances in Information Retrieval*, ECIR'13, page 693–696, Berlin, Heidelberg. Springer-Verlag.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate Speech Dataset from a White Supremacy Forum. *arXiv:1809.04444 [cs]*. ArXiv: 1809.04444.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Collins Dictionary. 2022. Collins dictionary.

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.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-Loop for Data Collection: a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Jane Frank. 1990. You call that a rhetorical question?: Forms and functions of rhetorical questions in conversation. *Journal of Pragmatics*, 14(5):723–738.

Simona Frenda, Alessandra Teresa Cignarella, Valerio Basile, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2022. The unbearable hurtfulness of sarcasm. *Expert Syst. Appl.*, 193(C).

Simona Frenda and Viviana Patti. 2019. Computational Models for Irony Detection in Three Spanish Variants. page 13.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90, Vancouver, BC, Canada. Association for Computational Linguistics.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural Metaphor Detection in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 607–613, Brussels, Belgium. Association for Computational Linguistics.

Herbert Paul Grice. 1967. Logic and conversation. In Paul Grice, editor, *Studies in the Way of Words*, pages 41–58. Harvard University Press.

Herbert Paul Grice. 1989. *Studies in the Way of Words*. Cambridge: Harvard University Press.

Hugo Lewi Hammer. 2017. Automatic Detection of Hateful Comments in Online Discussion. In *Industrial Networks and Intelligent Systems*, pages 164–173, Cham. Springer International Publishing.

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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Vijayasaradhi Indurthi, Bakhtiyar Syed, Manish Shrivastava, Nikhil Chakravartula, Manish Gupta, and Vasudeva Varma. 2019. FERMI at SemEval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 70–74, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Florence, Italy. Association for Computational Linguistics.

Kepa Korta and John Perry. 2020. Pragmatics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Spring 2020 edition. Metaphysics Research Lab, Stanford University.

Zhengyan Li, Yicheng Zou, Chong Zhang, Qi Zhang, and Zhongyu Wei. 2021. Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 246–256, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tobias Mayer, Santiago Marro, Elena Cabrio, and Serena Villata. 2020. Generating adversarial examples for topic-dependent argument classification. In *Computational Models of Argument - Proceedings of COMMA 2020, Perugia, Italy, September 4-11, 2020*, volume 326 of *Frontiers in Artificial Intelligence and Applications*, pages 33–44. IOS Press.

Merriam-Webster. 2022. Dictionary.

Meta. 2022. Facebook: Hate speech policies.

Isar Nejadgholi, Kathleen Fraser, and Svetlana Kiritchenko. 2022. Improving generalizability in implicitly abusive language detection with concept activation vectors. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5517–5529, Dublin, Ireland. Association for Computational Linguistics.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Rolandos Alexandros Potamias, Georgios Siolas, and Andreas Georgios Stafylopatis. 2020. A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, 32(23):17309–17320.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Barthes Roland, Lavers Annette, and Smith Colin. 1968. *Elements of semiology / Roland Barthes ; translated from the French by Annette Lavers and Colin Smith*, 1st american ed. edition. Hill and Wang New York.

Enrica Troiano, Carlo Strapparava, Gözde Özbal, and Serra Sinem Tekiroğlu. 2018. A computational exploration of exaggeration. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3296–3304, Brussels, Belgium. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Byron C. Wallace, Do Kook Choe, Laura Kertz, and Eugene Charniak. 2014. Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 512–516, Baltimore, Maryland. Association for Computational Linguistics.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada. Association for Computational Linguistics.

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.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. Technical Report arXiv:1901.11196, arXiv. ArXiv:1901.11196 [cs] type: article.

Michael Wiegand, Elisabeth Eder, and Josef Ruppenhofer. 2022. Identifying Implicitly Abusive Remarks about Identity Groups using a Linguistically Informed Approach. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, pages 5600–5612, Seattle, United States. Association for Computational Linguistics.

Michael Wiegand, Maja Geulig, and Josef Ruppenhofer. 2021a. Implicitly abusive comparisons – a new dataset and linguistic analysis. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 358–368, Online. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021b. Implicitly abusive language – what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587, Online. Association for Computational Linguistics.

Ulrike Willinger, Andreas Hergovich, Michaela Schmoeger, Matthias Deckert, Susanne Stoettner, Iris Bunda, Andrea Witting, Melanie Seidler, Reinhilde Moser, Stefanie Kacena, David Jaeckle, Benjamin Loader, Christian Mueller, and Eduard Auff. 2017. Cognitive and emotional demands of black humour processing: the role of intelligence, aggressiveness and mood. *Cogn. Process.*, 18(2):159–167.

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 656–666, Montréal, Canada. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.

Ziqi Zhang and Lei Luo. 2018. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *CoRR*, abs/1803.03662.

A Performance Details in Data Augmentation

Inspired by the ranked one augmentation strategy in ElSherief et al. (2021), i.e., a back-translation approach, we also test SOTA models on our dataset, ISHate, with the augmentation techniques described in Section 4.4. Each model is trained with the originally collected data described in Section 4.3 and additional data obtained from one augmentation strategy. At the end, we also evaluate each model using only non-augmented test data. Tables 6a and 6b show the experiments' results on tasks A and B.

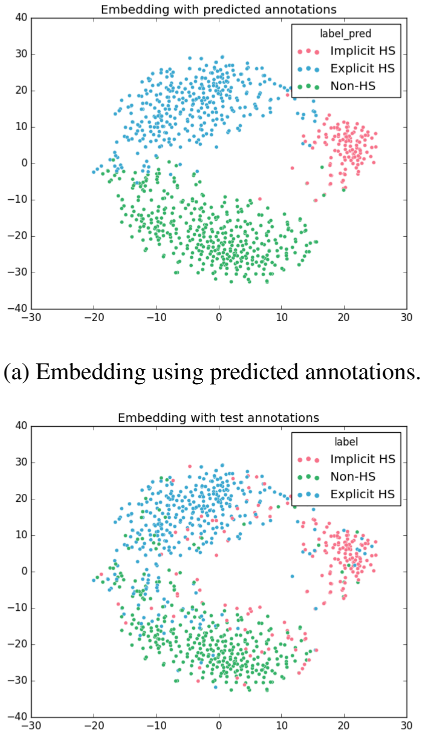
We further analyse the errors committed by the best performing model on task A. We took from Table 6a HateBERT+ALL and the third annotation layer described in Sections 3 and 4 to identify which are the most frequent implicit properties on task A miss-classified messages. Table 5 shows how Inference, Context, Sentiment, Exaggeration, and Extralinguistic knowledge are the most recurrent not captured devices.

	Implicit HS	
Implicit Property	#	%
Inference	44	53.659
Context	34	41.463
Sentiment	33	40.244
Exaggeration	23	28.049
Extralinguistic knowledge	20	24.390
Irony	17	20.732
Black humor	12	14.634
Visual signs	11	13.415
Metaphor	9	10.976
Rhetorical question	8	9.756
Antithesis	6	7.317
Humiliation	5	6.098
Sarcasm	5	6.098
Circumlocution	4	4.878
Fallacy	4	4.878
Euphemism	3	3.659

Table 5: Implicit properties of the messages that are not captured by HateBERT+ALL

.

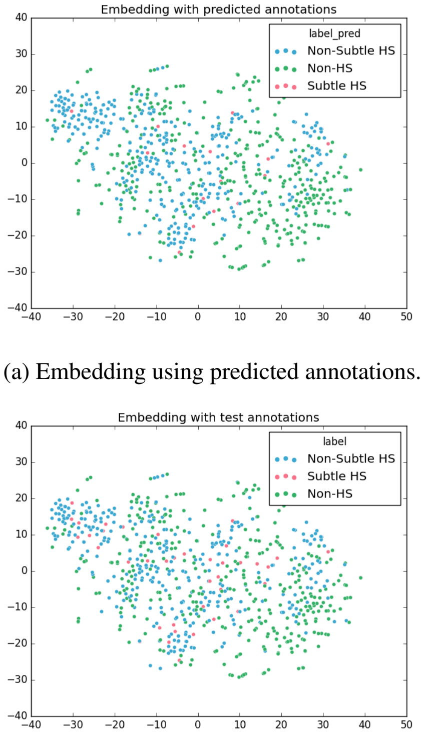
We also analysed the embeddings of our best-performing models in tasks A and B (HateBERT+ALL and USE+SVM+BT, respectively) through t-SNE (van der Maaten and Hinton, 2008). Figures 1 and 2 show the text embeddings for sentences of the test set, labeled by both classifiers and annotators, for the implicit and subtle tasks.



(a) Embedding using predicted annotations.

(b) Embedding using manual annotations.

Figure 1: Embedding of HateBERT + ALL in the test set of task A



(a) Embedding using predicted annotations.

(b) Embedding using manual annotations.

Figure 2: Embedding of USE+SVM+BT in the test set of task B

Task A	Non-HS			Explicit HS			Implicit HS		
USE + SVM	P	R	F1	P	R	F1	P	R	F1
RSA	.887	.875	.881	.768	.830	.798	.515	.285	.367
AAV	.888	.875	.881	.764	.825	.793	.491	.280	.356
RNE	.887	.867	.877	.770	.802	.786	.386	.382	.384
RI	.888	.865	.876	.769	.803	.786	.371	.371	.371
RA	.888	.867	.877	.768	.803	.785	.398	.387	.392
EDA	.892	.862	.877	.780	.797	.789	.339	.441	.383
BT	.897	.856	.876	.782	.787	.785	.403	.645	**.496**
GM	.887	.862	.874	.771	.794	.782	.352	.409	.378
GM+Revised	.887	.863	.875	.769	.803	.786	.385	.403	.394
ALL	.889	.879	**.884**	.796	.809	**.803**	.421	.430	.426
BERT	P	R	F1	P	R	F1	P	R	F1
RSA	.894	.899	.896	.805	.835	.819	.487	.301	.372
AAV	.896	.896	.896	.820	.817	.819	.387	.398	.393
RNE	.897	.897	.897	.807	.829	.818	.455	.349	.395
RI	.909	.897	**.903**	.812	.850	**.831**	.458	.376	.413
RA	.894	.905	.899	.822	.825	.823	.473	.382	.423
EDA	.900	.894	.897	.807	.836	.821	.416	.333	.370
BT	.909	.887	.898	.824	.826	.825	.459	.608	**.523**
GM	.898	.901	.900	.824	.821	.823	.409	.398	.403
GM+Revised	.905	.892	.899	.811	.839	.825	.451	.419	.435
ALL	.902	.894	.898	.816	.817	.816	.488	.543	.514
DeBERTaV3	P	R	F1	P	R	F1	P	R	F1
RSA	.912	.893	.902	.803	.877	.838	.441	.242	.312
AAV	.916	.904	.910	.832	.858	.845	.431	.403	.417
RNE	.922	.883	.902	.807	.880	.842	.430	.382	.405
RI	.922	.894	.908	.821	.878	.849	.460	.398	.427
RA	.909	.914	**.911**	.841	.859	**.850**	.482	.360	.412
EDA	.907	.899	.903	.813	.859	.835	.460	.312	.372
BT	.919	.885	.902	.830	.857	.844	.428	.543	**.479**
GM	.913	.899	.906	.839	.843	.841	.399	.468	.431
GM+Revised	.918	.893	.905	.819	.873	.845	.425	.366	.393
ALL	.924	.887	.905	.814	.867	.840	.456	.478	.467
HateBERT	P	R	F1	P	R	F1	P	R	F1
RSA	.895	.895	.895	.814	.830	.822	.452	.382	.414
AAV	.899	.900	.899	.819	.825	.822	.428	.398	.412
RNE	.904	.891	.897	.815	.850	**.832**	.415	.355	.383
RI	.902	.894	.898	.808	.845	.826	.408	.312	.354
RA	.895	.904	**.900**	.830	.823	.826	.459	.419	.438
EDA	.890	.901	.895	.808	.820	.814	.454	.317	.373
BT	.910	.880	.895	.820	.823	.822	.378	.543	.446
GM	.901	.898	.899	.824	.827	.825	.414	.425	.419
GM+Revised	.905	.891	.898	.816	.835	.826	.408	.419	.414
ALL	.903	.896	.899	.827	.827	.827	.502	.559	**.529**

(a) Results of SOTA models using data augmentation on task A.

	Non-HS			Non-Subtle HS			Subtle HS		
USE + SVM	P	R	F1	P	R	F1	P	R	F1
RSA	.891	.871	**.881**	.787	.832	**.809**	.800	.103	.182
AAV	.891	.871	.881	.786	.831	.808	.571	.103	.174
RNE	.891	.868	.879	.783	.831	.806	.571	.103	.174
RI	.891	.868	.879	.782	.832	.806	.750	.077	.140
RA	.891	.868	.879	.783	.832	.806	.800	.103	.182
EDA	.892	.870	.881	.788	.828	.807	.263	.128	.172
BT	.892	.868	.880	.789	.831	.809	.739	.436	**.548**
GM	.891	.867	.879	.786	.827	.806	.269	.179	.215
GM+Revised	.892	.866	.879	.785	.826	.805	.286	.205	.239
ALL	.888	.874	.881	.797	.818	.807	.263	.256	.260
BERT	P	R	F1	P	R	F1	P	R	F1
RSA	.894	.911	**.902**	.840	.828	.834	.200	.051	.082
AAV	.898	.896	.897	.824	.836	.830	.300	.154	.203
RNE	.899	.895	.897	.826	.839	.833	.400	.256	.312
RI	.899	.889	.894	.819	.840	.830	.240	.154	.188
RA	.906	.893	.900	.823	.850	.836	.190	.103	.133
EDA	.902	.885	.893	.813	.845	.829	.143	.077	.100
BT	.898	.900	.899	.839	.832	.835	.304	.359	.329
GM	.899	.899	.899	.836	.839	**.837**	.194	.154	.171
GM+Revised	.903	.893	.898	.826	.843	.835	.206	.179	.192
ALL	.904	.883	.893	.813	.845	.829	.385	.385	**.385**
DeBERTaV3	P	R	F1	P	R	F1	P	R	F1
RSA	.922	.894	.908	.826	.879	.852	.333	.103	.157
AAV	.910	.907	.908	.841	.858	.849	.267	.103	.148
RNE	.923	.893	.907	.829	.881	.854	.261	.154	.194
RI	.910	.894	.902	.828	.860	.843	.364	.205	.262
RA	.923	.884	.903	.815	.883	.848	.188	.077	.109
EDA	.924	.888	.905	.819	.882	.850	.188	.077	.109
BT	.920	.897	.908	.835	.876	.855	.385	.256	.308
GM	.911	.910	**.911**	.847	.860	**.854**	.316	.154	.207
GM+Revised	.911	.902	.907	.837	.856	.846	.267	.205	.232
ALL	.926	.881	.903	.817	.877	.846	.306	.385	**.341**
HateBERT	P	R	F1	P	R	F1	P	R	F1
RSA	.900	.893	.896	.823	.841	.832	.273	.154	.197
AAV	.901	.894	.897	.823	.842	.832	.292	.179	.222
RNE	.897	.894	.896	.823	.836	.829	.167	.103	.127
RI	.906	.886	.896	.812	.852	.832	.176	.077	.107
RA	.897	.892	.895	.816	.836	.826	.077	.026	.038
EDA	.902	.890	.896	.819	.845	.832	.217	.128	.161
BT	.909	.883	.896	.820	.848	.834	.207	.308	.247
GM	.899	.898	**.899**	.831	.834	**.832**	.250	.231	.240
GM+Revised	.894	.898	.896	.826	.826	.826	.192	.128	.154
ALL	.903	.881	.892	.816	.844	.830	.391	.462	**.424**

(b) Results of SOTA models using data augmentation on task B.

Table 6: Obtained results on tasks A and B by all models and different types of augmented data.

B Implicit Properties

In the following part, we provide a list of implicit properties with their definitions. All the examples illustrating implicit properties are used in implicit hateful messages and their descriptions are presented in the annotation guidelines.

Antithesis – the rhetorical contrast of ideas through parallel arrangements of words, clauses, or sentences (as in "action, not words" or "they promised freedom and provided slavery") (Merriam-Webster, 2022)

Black humor – humor marked by the use of usually morbid, ironic, grotesquely comic episodes; humor treating sinister subjects like death, disease, deformity, handicap or warfare with bitter amusement (Willinger et al., 2017)

Circumlocution – the use of an unnecessarily large number of words to express an idea (Merriam-Webster, 2022)

Context – the parts of a discourse that surround a word or passage and can throw light on its meaning (Dadvar et al., 2013)

Euphemism – the substitution of an agreeable or inoffensive expression for one that may suggest something unpleasant (Casas Gómez, 2009)

Exaggeration (hyperbole) – an act or instance of exaggerating something: overstatement of the truth (Troiano et al., 2018)

Extralinguistic knowledge – any knowledge that exists outside knowledge of the language. In other words, it refers to knowledge that an author or a recipient of a message may possess about the message itself or about the world, but which is not expressed by any linguistic means.

Fallacy – a false or mistaken idea; an often plausible argument using false or invalid inference (Merriam-Webster, 2022)

Humiliation – the embarrassment and shame a person feels when someone makes them appear stupid or when they make a mistake in public (Dictionary, 2022)

Inference – something that is inferred. The premises and conclusion of a process of inferring (Merriam-Webster, 2022)

Irony – the use of words to express something other than and especially the opposite of the literal meaning; incongruity between the actual result of a sequence of events and the normal or expected result (Potamias et al., 2020)

Metaphor – a figure of speech in which a word or phrase literally denoting one kind of object or idea is used in place of another to suggest a likeness or analogy between them (Choi et al., 2021; Gao et al., 2018)

Metonymy – a figure of speech consisting of the use of the name of one thing for that of another of which it is an attribute or with which it is associated (such as "crown" in "lands belonging to the crown") (Merriam-Webster, 2022)

Rhetorical question – a question not intended to require an answer, used mainly for dramatic effect (Frank, 1990)

Sarcasm – a mode of satirical wit depending on its effect on bitter, caustic, and often ironic language usually directed against an individual. Sarcasm differs from irony with one distinct characteristic: negativity. Sarcasm is mostly witty mockery having a negative connotation whereas irony does not represent negativity (Potamias et al., 2020)

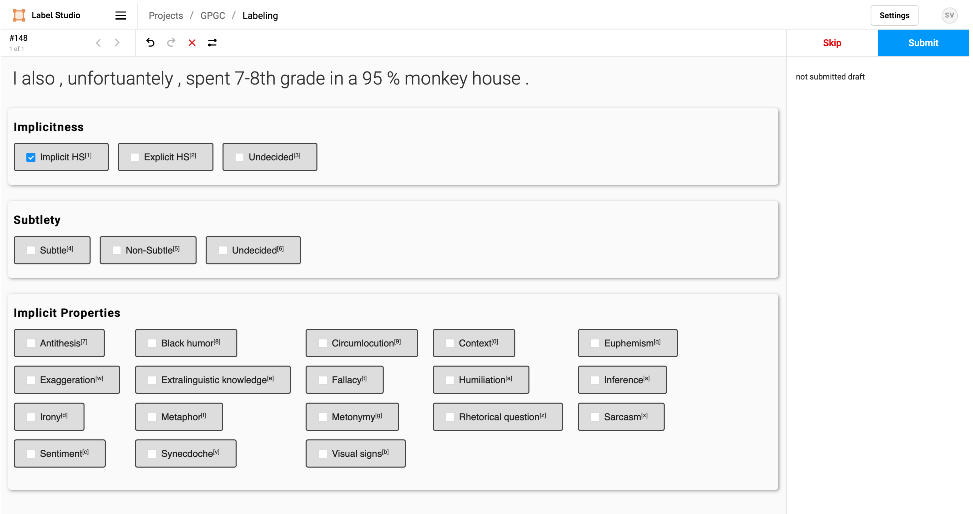
Sentiment – an attitude, thought, or judgment prompted by feeling; the emotional significance of a passage or expression as distinguished from its verbal context (Li et al., 2021)

Synecdoche – a figure of speech by which a part is put for the whole, the whole for a part, the species for the genus, the genus for the species, or the name of the material for the thing made (Merriam-Webster, 2022)

Visual signs – punctuation marks, quotes, and use of uppercase that play a role of support in hate messages.

C Annotation Tool Interface

Figure 3a demonstrates a screenshot of the annotation interface of the Label Studio tool used for the labeling process. According to the annotation scheme represented by three annotation layers (discussed in Section 3 and Subsection 4.2) Label Studio has three consecutive annotation steps. The first step consists in implicitness with three choices: Implicit HS, Explicit HS, Undecided, keeping in mind that the tool allows to filter Non-Hate out before starting the labeling process. The first choice of Implicit HS or Explicit HS brings in the appearance of the second step of subtlety with three choices: Subtle, Non-Subtle, Undecided. This step does not appear with an Undecided choice at the previous step. As well as that, the choice of Implicit HS triggers the appearance of the third step which consists of implicit properties being characteristic of only implicit messages. Figure 3b shows the shape of the resultant dataset after annotation.



I also , unfortuantely , spent 7-8th grade in a 95 % monkey house .

not submitted draft

Implicitness

☑ Implicit HS[1] □ Explicit HS[2] □ Undecided[3]

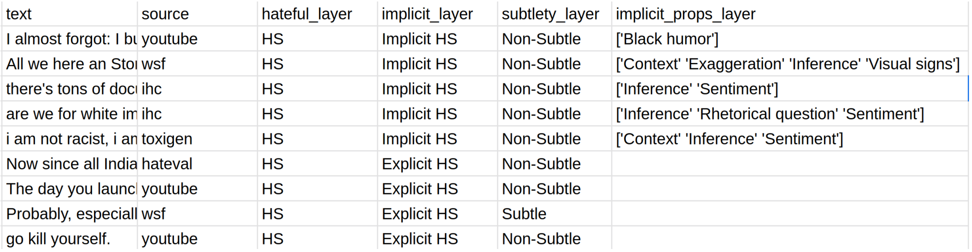
Subtlety

□ Subtle[4] □ Non-Subtle[5] □ Undecided[6]

Implicit Properties

□ Antithesis[7] □ Black humor[8] □ Circumlocution[9] □ Context[0] □ Euphemism[q]
□ Exaggeration[w] □ Extralinguistic knowledge[e] □ Fallacy[t] □ Humiliation[a] □ Inference[s]
□ Irony[d] □ Metaphor[f] □ Metonymy[g] □ Rhetorical question[z] □ Sarcasm[x]
□ Sentiment[c] □ Synecdoche[v] □ Visual signs[b]

(a) Annotation tool interface.



text	source	hateful_layer	implicit_layer	subtlety_layer	implicit_props_layer
I almost forgot: I bu	youtube	HS	Implicit HS	Non-Subtle	['Black humor']
All we here an Stoi	wsf	HS	Implicit HS	Non-Subtle	['Context' 'Exaggeration' 'Inference' 'Visual signs']
there's tons of doci	ihc	HS	Implicit HS	Non-Subtle	['Inference' 'Sentiment']
are we for white im	ihc	HS	Implicit HS	Non-Subtle	['Inference' 'Rhetorical question' 'Sentiment']
i am not racist, i an	toxigen	HS	Implicit HS	Non-Subtle	['Context' 'Inference' 'Sentiment']
Now since all India	hateval	HS	Explicit HS	Non-Subtle	
The day you launci	youtube	HS	Explicit HS	Non-Subtle	
Probably, especiall	wsf	HS	Explicit HS	Subtle	
go kill yourself.	youtube	HS	Explicit HS	Non-Subtle	

(b) Sample of the ISHate dataset after the annotation process.

Figure 3: Label Studio interface to enhance the 7 HS datasets described in Section 4 with three new additional annotation layers: *implicit_layer* (Explicit HS/Implicit HS), *subtlety_layer* (Non-Subtle HS/Subtle HS), and *implicit_props_layer* (Antithesis/Black humor/Context/etc.). The annotation layer *hateful_layer* (Non-HS/HS) consists of the already provided labels of each HS corpus, with the exception of the Youtube dataset where we re-annotated it.