Annotation Guidelines for Implicit and Subtle Hate Speech Detection

1 Introduction

Social media have faced mounting pressure from civil rights groups to enforce their anti-hate speech policies. However, the increasing availability of online user-generated content and the growth of social media platforms make the detection of this content quite challenging. This issue has led to automatic systems that depend on computational techniques to identify inappropriate messages. Most approaches developed to detect hate speech (HS) messages automatically rely on supervised machine learning systems that predict if a message has hateful content according to certain message features. For example, that is the case of (Agarwal and Sureka 2015), (Corazza et al. 2020), (Jurgens, Hemphill, and Chandrasekharan 2019), (Wulczyn, Thain, and Dixon 2017). These techniques depend on human-annotated samples that indicate the expected result of supervised algorithms. However, the lack of a common framework and definitions of HS among scholars from various disciplines leaves room for subjective interpretations. For example, the same linguistic phenomenon can have different names or labels for multiple phenomena (Poletto et al. 2021). The lack of agreement might lead to a poorly annotated corpus and undesirable results after the classification process (Leonardelli et al. 2021). Therefore one aim of these guidelines is to settle standard definitions that we will be using to build our corpus.

While explicit HS is more easily identifiable by recognizing hateful words or phrases, implicit HS employs metaphor, irony, rhetorical questions, or peculiar syntax that can only be captured by understanding its overall complex meaning (Waseem et al. 2017). Unfortunately, this phenomenon is invisible to automatic classifiers (Corazza et al. 2020), with no previous work that provides any quantitative measurement of the classifier performance against implicitness in this context. That is why our second objective is to include annotations in our corpus that differentiate explicit and implicit hate speech messages among consistent definitions.

2 Definitions

2.1 HS and related concepts

On the one hand, we dealt with the fuzzy boundaries between HS and broader concepts such as abusive/toxic language or overlapping concepts such as offensive and aggressive language. On the other hand, we considered HS and more specific focus-driven notions such as racism, antisemitism, sexism, misogyny, and homophobia. Several attempts to understand these overlapping concepts are in the literature (Poletto et al. 2021), (Waseem et al. 2017). Since we consider that acknowledging these concepts is crucial for annotators, based on previous work, we specified and depicted how these broader concepts overlap to introduce our annotation criteria in Section 3.

To begin with, we start providing the definitions of the following main notions:

Abusiveness/toxicity. Hurtful language, including hate speech, derogatory language, and also profanity (Fortuna, Rocha da Silva, et al. 2019)

Aggressiveness. Intention to be aggressive, harmful, or even to incite, in various forms, to violent acts against a given target (Zampieri et al. 2019)

Offensiveness. Profanity, strongly impolite, rude, or vulgar language expressed with fighting or hurtful words in order to insult a targeted individual or group (Fortuna and Nunes 2018).

Hate speech (HS) according to Facebook Community Standards. We define hate speech as a direct attack against people – rather than concepts or institutions – based on what we call protected characteristics (PC): race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and severe disease. We define attacks as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing, and calls for exclusion or segregation. We also prohibit the use of harmful stereotypes, which we define as dehumanizing comparisons that have historically been used to attack, intimidate or exclude specific groups that are often linked with offline violence. We consider age a protected characteristic when referenced along with another protected characteristic. We also protect refugees, migrants, immigrants, and asylum seekers from the most severe attacks, though we do allow commentary and criticism of immigration policies. Similarly, we provide some protections for characteristics such as occupation when they are referenced along with a protected characteristic. Sometimes, we consider certain words or phrases as code words for PC groups based on local nuance (*Facebook: HS policies* n.d.).

The distinction between explicitness and implicitness is analogous to the linguistic perception of denotation and connotation. Denotation represents the literal meaning of a term or a symbol, while connotation refers to the vast array of positive and negative associations that most words naturally carry with them (Waseem et al. 2017). Thus, we may postulate that the explicit meaning refers to the literal definition of a word found in a dictionary. However, the implicit meaning represents the emotional weight of a word or the meaning that the message's author implies and wants to convey. From this perspective, the notions of explicit and implicit hate speech have the following definitions:

Explicit HS. 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 (Waseem et al. 2017).

Implicit HS. It does not immediately imply or denote abuse/hate. Here, ambiguous terms, sarcasm, and other means obscure a message's true nature, generally making it more difficult to detect by both annotators and machine learning approaches. (Waseem et al. 2017).

The other notion that appeals to our attention is subtlety. Some implicit hate messages convey hateful meaning in a peculiar indirect way that needs extra knowledge and syntactic and grammatical proficiency. As well as that, explicit hate may be expressed delicately. As a result, the true hateful meaning is hard to comprehend, although the message represents an explicit hate instance. On the other hand, some hateful messages are implicit and need particular experience and erudition on top to comprehend the real meaning and target. This characteristic is expressed in the term subtlety.

Subtle. An implicit or explicit message that is so delicate or elusive as to be difficult to analyze or describe; ingenious; indirect method to deliver the meaning.

Those messages that straightforwardly convey the meaning are addressed as non-subtle.

Considering HS as an instance of abusive language, we have to mention that hate messages can have different targets. Therefore, we will describe different types of HS in the following part (Chiril 2021).

2.1.1 Ethnicity-based HS

Racism. A belief that race is a fundamental determinant of human traits and capacities and that racial differences produce an inherent superiority of a particular race (*Merriam Webster* n.d.).

Xenophobia. Fear and hatred of strangers or foreigners or of anything that is strange or foreign (*Merriam Webster* n.d.).

Islamophobia. A fear, prejudice, and hatred of Muslims or non-Muslim individuals that leads to provocation, hostility, and intolerance by means of threatening, harassment, abuse, incitement, and intimidation of Muslims and non-Muslims, both in the online and offline world. Motivated by institutional, ideological, political, and religious hostility that transcends into structural and cultural racism that targets a Muslim's symbols and markers (Awan and Zempi 2020).

Anti-Semitism. A certain perception of Jews, which may be expressed as hatred toward Jews. Rhetorical and physical manifestations of antisemitism are directed toward Jewish or non-Jewish individuals and/or their property, toward Jewish community institutions and religious facilities (*International Holocaust Remembrance Alliance* n.d.).

2.1.2 Gender-based HS

Sexism. Sexist HS aims to humiliate or objectify women, undervalue their skills and opinions, destroy their reputation, make them feel vulnerable and fearful, and control and punish them for not following a certain behavior (*Combating Sexist Hate Speech* n.d.).

Misogyny. Hatred of, aversion to, or prejudice against women (*Merriam Webster* n.d.).

Despite the fact that sexism and misogyny express hatred toward women highlighting their inequality with men, the ideologies of these two vary. While the ideology of sexism tends to discriminate against women to uphold men's superiority, the misogyny ideology separates good women from bad women implying hateful content toward bad women who may attempt to take over from men (Chiril 2021).

Homophobia. Irrational fear of, aversion to, or discrimination against homosexuality or gay people (*Merriam Webster* n.d.).

Transphobia. Irrational fear of, aversion to, or discrimination against transgender people (*Merriam Webster* n.d.).

2.2 Overlapping analysis

Even though we focus on HS, annotators have to be capable of recognizing related concepts. Figure 1 shows a Venn diagram that attempts to depict and clarify how the previous definitions overlap. In the first place, the set of abusiveness/toxicity is the umbrella term that contains HS, aggressiveness, offensiveness, and subtlety. This umbrella is divided into explicit and implicit counterparts, separating other concepts. Offensiveness remains in the explicit set since it depends on the language used. It is necessary to note that if we project HS, it contains messages related to racism, xenophobia, islamophobia, among others. However, this does not mean that they are subsets of HS. We only wanted to emphasize that HS is always related to protected characteristics (according to our definition).

We enumerated each possible overlapping in the Venn diagram providing examples of this distribution. Each example was extracted from available corpora built in previous work. We also referenced the dataset from where messages were obtained or the article in which they were cited. The analyzed corpora and related articles can be found in Section 5. In the cases where a reference is not provided, the example was crafted by ourselves based on existing messages.

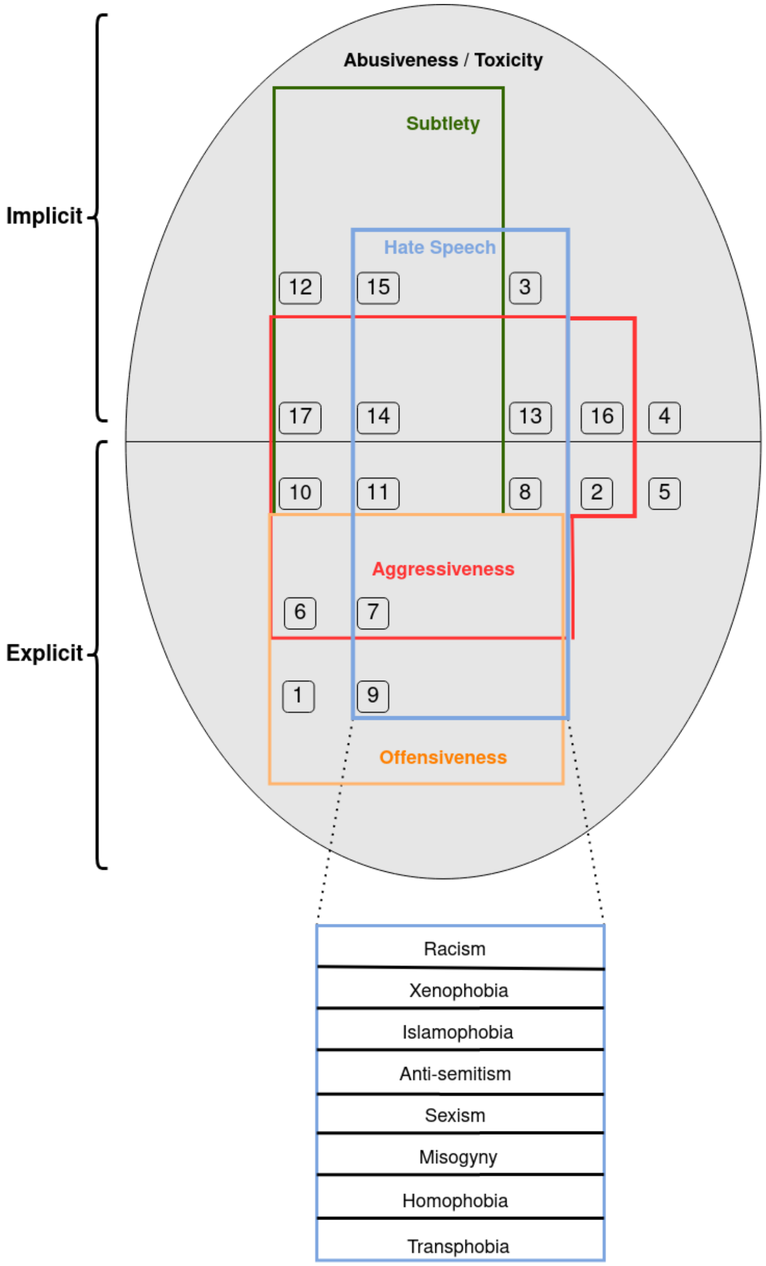

Figure 1: Relations Between HS and Related concepts

Segment 1. It only overlaps the offensive and explicit sets. Note that every message of this set is offensive and explicit since this type of message depends on the use of unambiguous hateful words directed to a target. An offensive message strongly depends on its syntactic meaning and the use of certain words. Two examples of these are:

1. _Why you worried bout that other h*e? Cuz that other h*e aint worried bout another h*e and I knew Kendrick Lamar was onto something when he said “I call a b*tch a b*tch, a h*e a h*e, a woman a woman_ (CKTG)

2. _Youre one of the ugliest b*tches Ive ever fucking seen_ (Waseem et al. 2017)

Expressions as _h*e_, _ugliest_, _b*tches_, and _fucking_, are the explicit and offensive factors. In addition, the use of personal pronouns such as _You_ indicates that the target is an individual. Moreover, even though this message has sexist words, it cannot be considered HS without further context.

Segment 2. Not all explicit messages are offensive. That is the case of messages that suggest unambiguous aggression without offensive words. The aggressiveness set contains messages that intend to be aggressive or promote an action that might harm a specific target.

3. _String ’ em up !_ (GPGC)

According to our definition, the example 3 does not contain vulgar language, but it is still clearly aggressive since it indicates a violent act against a target. The verb _string up_ combined with a personal pronoun is the indicator of this example, making it quite explicit. However, it does not address any protected characteristic to be considered HS.

Segment 3. HS may be confusing combined with aggressiveness and offensiveness. The most significant difference of HS w.r.t. the rest is the message’s target. In this case, it has to be a person or a group of people belonging to the category of protected characteristics.

4. _most of them come north and are good at just mowing lawns_ (DWMW)

Example 4 contains neither offensive nor aggressive meaning, but is still an attack on immigrants as the authors suggest that immigrants are good at doing very basic tasks only. There is no use of offensive language and it does not suggest any violent act. 4 is our first example of implicitness since it uses properties such as exaggeration and sentiment to express a harmful stereotype. One possible indicator of the previous properties is the use of *just* to quantify the idea of the message.

Segment 4. The most extensive set in Figure 1 is abusiveness/toxicity. It contains every message that can be considered inappropriate. However, this overlapping only considers those implicit abusive/toxic messages that do not belong to any of the other sets.

5. *Hey Brendan, you look gorgeous today. What beauty salon did you visit?* (DWMW)

6. *you're intelligence is so breathtaking!!!!!!* (DWMW)

Messages 5 and 6 are target-directed against individuals. Both messages are considered hurtful and entirely implicit. 5 takes an advantage of irony and rhetorical questions, whereas 6 uses irony, exaggeration, and sentiment. No protected characteristics are addressed in both messages. There is also neither sign of offensive nor aggressive language.

Segment 5. This set is the explicit counterpart of abusive/toxic explicit messages that do not belong to the other sets.

7. *More petulant commentary from yet another, um, 'academic chair-human', umm, ' intellectual'....*

Example 7 is a toxic message for a disrespectful bad-intentional complaint about commentaries related to a certain context. We considered it explicit as the expression *More petulant commentary from yet another* clarified the message's intention, followed by filler phrases *um*, *umm* for mocking the situation, and quoted examples like *academic chair-human* and *intellectual*. We do not consider it offensive or aggressive since it is not directed to a person or group but toxically addresses a situation.

Segment 6. In this case, there is an interlacement between offensiveness and aggressiveness. Here a message can contain rude language but also incite an action.

8. *Go kill yourself”, “You’re a sad little f*ck* (DWMW)

kill yourself is the explicit aggressive factor against a person, whereas *You’re* and *f*ck* are the explicit offensive sign.

Segment 7. It is similar to the previous segment, but the target of the message is a person or a group of people addressing protected characteristics.

9. *@JuanYeez shut yo beaner ass up sp*c and hop your f*ggot ass back across the border little n*gga* CONAN

10. *@MDreyfus @NatFascist88 Sh*t your ass your moms p*ssy u Jew b*stard. Ur times coming. Heil Hitler!* (CONAN)

Note how *hop yo ass* or *shut your ass up* are both aggressive expressions. *sp*c* and *f*ggot* are offensive expressions. The only difference with example 9 is the use of hurtful language against black people utilizing the word *n*gga* and *across the border*, indicating racism and xenophobia. In the case of example 10, *Sh*t your ass* makes this sentence aggressive, whereas *p*ssy* and *b*stard* are notions of offensiveness. It is also HS since the targets are Jews. Another thing to point out in this example is the expression, *Heil Hitler!* (German for ”hail”) used to express that someone is a fascist/nazi, combined with the traditional Nazi-era salute. Even though the latest may be considered extralinguistic knowledge, the message is unambiguous and explicit.

Segment 8. This set is similar to Segment 2, where we have aggressive explicit messages without signs of offensive language. The only difference is that it also has HS language.

11. *On a short holiday , Go back and do something about your own country and stop putting crap on our doorstep !* (GPGC)

Note that 11 is against immigrants where the aggressive language is *Go back* and *do something about your own country*. The clues to identify HS are *your own country* and the posterior use of personal pronouns like *our* emphasizing hateful language.

Segment 9. The combination of offensiveness and HS without any aggression is the essence of this set. An example of this segment is the following:

12. *Traitors are pathetic ignorant fools , and negroes are fools because they are starting to literally bite the hands that feed them .* (GPGC)

Here, the target is against people of black African heritage and white American people against racism. Note the use of the offensive terms *negroes* and *traitors* to refer to these groups, respectively. There are other offensive languages, such as *pathetic ignorant fools*. It is also possible to perceive exaggeration using adverbs of manner like *literally* before a justification.

Segment 10. Here there is an overlap of subtlety and aggressiveness. In this case, we couldn't find an example among our revised messages, but still, we proposed a possible example that belongs to this set.

13. *As a polite Englishman I was just about to show you the way out of my home.* (crafted)

We intended to add the violent act *kick out* but with euphemism to deviate the message's intention without eliminating explicit signs. In this case, *As a polite Englishman* and *show you the way out of my home* ingeniously deviate the focus of the message, making it subtle.

Segment 11. This set is similar to the previous one, but it adds the HS overlapping.

14. *Gas the skypes* (DWMW)

Here, the word *skypes* is used as a euphemism for slurs against jews. This expression can not be found in dictionaries. This subtle factor influences the message's meaning and makes it difficult to understand at first glance. Note that aggressiveness is apparent once we recognize the role of *skypes* in the example.

Segment 12. This segment is similar to 4 because it contains toxic, implicit messages that are not HS, aggressive, or offensive. However, the difference here is that meaning is delicate and difficult to analyze, making them subtle.

15. *I'm shocked to learn about that human traffickers had guns some probably illegal too but CA needs more gun control. But don't worry about the actual crime coz the pic of black guns is worse.* (crafted, based on HatEval)

This example is a complaint about gun control in California that at first glance may be HS against African heritage people because of the mention of *black guns*. However, this is just a synonym that refers to assault automatic guns like AKs, AR's, mac 10's, GSG's. One may think that the targets are human traffickers, but it is just a way to emphasize and exaggerate the message. Another important property is the use of sarcasm in the phrase *But don't worry about the actual crime coz the pic of black guns is worse* which makes it subtle. As well as that, the use of coordinate and subordinate connectors and circumlocution at the beginning make the delivery of the meaning quite delicate, thus, this message may be considered subtle.

Segment 13. This set is quite peculiar since it overlaps HS and aggressiveness in the implicit counterpart of our Venn diagram. Here the violent act has to be implicitly injected into the message.

16. *Separation is the only answer !* (CONAN)

17. *If it 's brown , flush it down .* (CONAN)

16 indicates the separation of black African heritage people from white people. However, it is implicit since it needs extralinguistic knowledge of

where this message was posted, in this case, a white supremacy forum. It also shows exaggeration and sentiment. 17 has signs of irony and metonymy. Indicators of the properties of 17 are the expressions *flush* for waste removal and how *brown* replaces black people. This hidden meaning suggests the violent act of ostracizing and getting rid of this community.

Segment 14. It is similar to 13 but with subtlety, so a message has to contain HS, be aggressive, subtle and implicit.

18. *Sort of a back handed compliment that the anti-Whites are so keen to come on here and tell us that all is lost - in their heart of hearts even they must know that what they have constructed is only temporary and all will be ripped down some day .*

The subtle characteristic of this message is its complex structure that contains coordinate and subordinate clauses which make it complicated to trace the relations between parts of a phrase. Moreover, this message contains circumlocution which adds some extra weight to it. The part *all will be ripped down some day* illustrates aggressiveness of the message, however, the meaning is delivered in such an intricate way as we still consider this message implicit. The necessity to have extralinguistic knowledge and context to understand the true meaning proves the implicitness of this example.

Segment 15. The segment is for those messages that are only HS, implicit, and subtle. That is, no sign of implicit violent acts. One example of this is the following:

19. *I 'm moving to a bigger house pretty soon , and that map is perfect for doing a groidal analysis .* (GPGC)

Example 19 means to avoid moving to the same zones where black people live. Note that 19 does not contain any implicit aggressive act, but it contains black humor and metaphor which prove its implicitness.

Segment 16. This particular set contains messages that are abusive/toxic with implicit aggression. There's no overlapping with the other sets.

20. *Why you think this party has your chance to leave ur cave? (crafted)*

21. *Have your teachers ever showed you how the doors look like? (crafted)*

With message 20, we intended to create an implicit message based on rhetorical question, metaphor, and aggressive notions. The expression *has your chance to leave your cave* is the aggressive part of this sentence, discriminating for participating in a situation. Furthermore, the cave concept refers to someone's home, comparing it to wildlife, whereas *chance* represents that the situation is not an excuse or reason to be a part of the *party*. The last expression refers to the context of this message. 21 is another case of this implicit aggression but with neither HS nor offensiveness signs. We again based it on a rhetorical question and metaphor. Here *teachers always showed* along with rhetorical signs represent having no common sense, whereas *how the doors look like* refers to the understanding that someone is not welcome or well suited to participate in the situation.

Segment 17. The last case contains implicit, aggressive, and subtle messages. Here we also created the example using an existing one.

22. *oh dear isnt that terrible well i suppose you will just have to leave then and go somewhere you are more valued* (crafted, based on a real message from GPGC)

Example 22 uses a rhetorical question, sarcasm, and exaggeration expressed lexically in the adjective *terrible* and semantically in the phrase. It is implicit and suggests aggression. The original message was written with the personal pronouns *they* instead of *you* changing the target to immigrants and therefore HS. It is also subtle since the message distorts the meaning leaving it under the reader's interpretation.

3 Annotation of Messages

3.1 Annotation Scheme

The annotation scheme is depicted in Figure 2. The annotation consists in assigning a message a 3-layer label. The 1st layer has three possible cases, Implicit HS, Explicit HS and Undecided. Note that we are taking into account neither offensiveness nor aggressiveness. We will only annotate following the definition of HS. However, it was important for us to explain related concepts, so there is a clear understanding when recognizing them. In particular, If a message belongs to one of the sets described in Subsection 2.2 it will be considered HS. Otherwise, it will be Non-Hate.

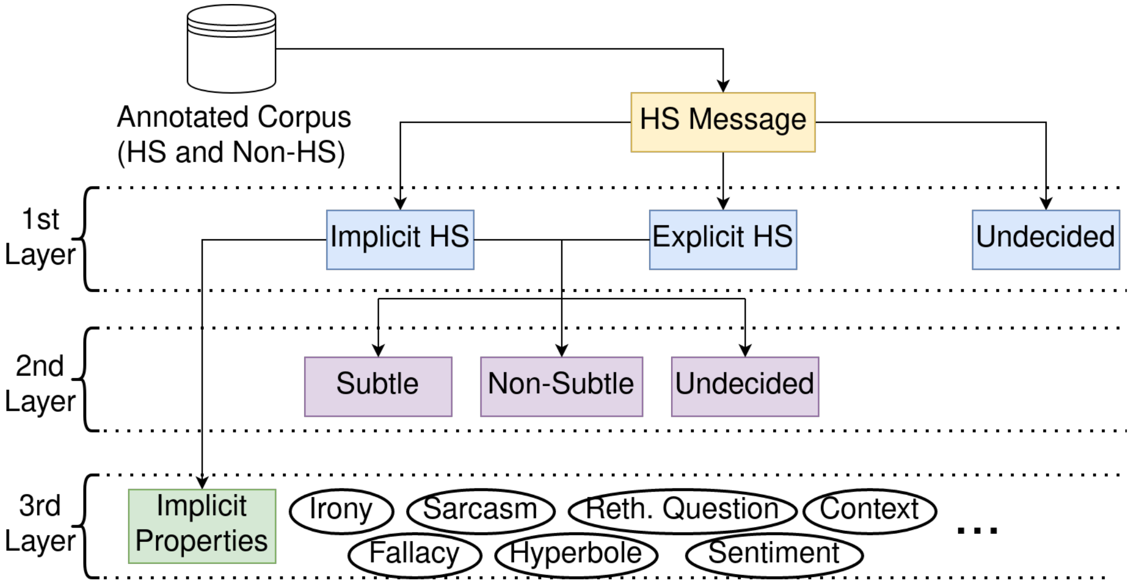

Figure 2: Annotation Scheme

The 2nd layer indicates if an HS message is subtle. This applies to both implicit and explicit messages. The 3rd layer focuses on adding a list of properties to HS implicit messages such as sarcasm, black humor, irony, rhetorical question, among others. The whole description of these properties will be discussed in Subsection 3.2.

At the end of the annotation process, we would like to have a corpus of the following structure:

message	implicit_layer	subtlety_layer	implicit_properties_layer
String ' em up !	Non-Hate		
On a short holiday , Go back and do something about your own country and stop putting crap on our doorstep !	Explicit HS	Non-Subtle	
I 'm moving to a bigger house pretty soon , and that map is perfect for doing a groidal analysis .	Implicit HS	Subtle	[Black humor]
They're coming out of the woodwork to oppose white unity and White History Month .	Implicit HS	Non-Subtle	[metaphor, metonymy]

Table 1: Annotated Corpus Format

Note that the first row only has a "Non-Hate" label but blank spaces for the columns "subtlety" and "implicit_properties". For the second row, we only left as blank the "implicit_properties" column since it does not apply for "HS explicit" cases. The last two rows are "HS implicit" examples with implicit properties. Note how commas separate them so they later can be easily preprocessed.

3.2 Implicit Properties

According to explicit and implicit HS definitions, we have to mention that instances of explicit HS are rather evident, thus, easy to recognize. On the other hand, implicit hate messages are difficult to detect as they do not directly convey hateful content. Nevertheless, having studied HS corpora, we identified typical features of implicit HS messages that may be used in implicit hate messages detection and annotation. In the following part, we will describe the features of implicit hate messages and provide their examples.

Rhetorical question. A question not intended to require an answer, used mainly for dramatic effect (*Merriam Webster* n.d.), (*Collins Dictionary* n.d.). We noticed that this stylistic device appears in line with other linguistic tricks aiming to veil the hateful meaning of a message. For illustration purposes, there are some examples explained in detail below.

23. *Do you really want to redefine marriage?* (CONAN)

24. *Why the mainstream media don't use more masculine voices in the news?* (CONAN)

Examples 23 and 24 show an interlacement of rhetorical questions and irony (underlined).

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 (*Merriam Webster* n.d.).

In example 24, irony overlaps with metonymy: the phrase *masculine voices* signifies masculine broadcasters on the TV or radio.

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* n.d.).

In both cases, the interlacement is so intricate as it is difficult to define the frontiers between different stylistic devices. As well as that, it is complicated to identify which words precisely convey ironical meaning: irony seems to be hardly palpable, being somewhere ”in the air” of a phrase.

25. *What have Muslims ever done for us?* (CONAN)

26. *Do you have any proof that the Holocaust ever happened?* (CONAN)

27. *Is it really ok to date someone of the same sex? I find it highly questionable.* (CONAN)

28. *Just visited the swimming baths with my little lad, 'Muslims Only' Session, couldn't get in. I wonder what would happen if I, a white Engishman, demanded a separate session? Its just Islamic appeasement!* (CONAN)

Examples 25, 26, 27 and 28 illustrate an interlacement of rhetorical questions and exaggeration, which is expressed by the adverbs *ever*, *really*, *highly*, *only*, *just*. They add extra weight to the phrase emphasizing the importance of notions in question.

Exaggeration (hyperbole). An act or instance of exaggerating something: overstatement of the truth (*Merriam Webster* n.d.).

Examples 27 and 28 imply irony in line with rhetorical questions and exaggeration. Again, it is complicated to identify which part of the two phrases contains ironic meaning. Apart from the stylistic devices described above, we have to mention a purely linguistic feature: lexical component. Use of words also provides us with the information on the topic and target – *Muslims*, *Holocaust*, *same sex*, *white* in examples 25, 26, 27 and 28 respectively do not possess any negative denotation. However, they shed light on the topic, and together with stylistic devices exploited, they let us place the messages in the implicit HS category.

Rhetorical question and exaggeration together may appear not only with irony but with sarcasm.

Sarcasm. A mode of satirical wit depending on its effect on bitter, caustic, and often ironic language usually directed against an individual (*Merriam Webster* n.d.).

29. *oh dear isnt that terrible well i suppose they will just have to leave then and go somewhere they are more valued* (GPGC)

Example 29 shows how a rhetorical question overlaps with sarcasm and exaggeration expressed lexically in the adjective *terrible* and semantically in the phrase.

We will notice various combinatorial tendencies if we switch over to stylistic devices other than rhetorical questions. Turning to cases of **exaggeration**, we may notice that this figure of speech can be used alone:

30. *I now only go into the city when i have to and i had to yesterday and I could not belive what has happend in there .I think i must have seen every peice of dirt from every corner of the globe in there it was horrible .* (GPGC)

Nevertheless, in many cases, this stylistic device is supported by others. To illustrate this fact, we will provide some examples below.

31. *The white man has no future in Canada , that is for sure , because the country is being flooded with non-white immigrants* . (GPGC)

32. *I agree, but how would you do so? Perhaps at the expense of our veterans or old people like me? If we give them everything we'll be left with nothing.* (CONAN)

Examples 31 and 32 show an interlacement of exaggeration with antithesis.

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* n.d.).

In example 31, *white men* is contrasted with *non-white immigrants*, while *that is for sure* and *flooded* express exaggeration of the immigration problem. In example 32, *everything* and *nothing* show both: antithesis on the one hand and exaggeration on the other hand. As well as that, 32 needs some extralinguistic knowledge to understand the phrase's meaning. Therefore, we can consider extra knowledge as one more feature to define implicit HS as there are many other cases of extra knowledge necessity which will be presented further.

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. For example:

33. *A Muslim's only real allegiance is to Allah (as a comment in relation to election for public office etc.)* (CONAN)

Example 33 illustrates an overlap of exaggeration (expressed by the phrase *only real*) and extralinguistic knowledge necessary to understand why this message may be considered HS. If we consider the information in parenthesis, we get the idea of the author: any kind of election does not mean anything for Muslims as their only belief is Allah and Koran.

A reference to some extralinguistic knowledge is always used in line with other stylistic devices. For example, frequent cases of metaphor used in implicit HS often need extralinguistic knowledge:

34. *My wife had to spend time at the UP Hospital in Philadelphia - thank God I was licensed to carry in Pa. Downtown Philadelphia is becoming a cesspool .* (GPGC)

Example 34 shows a metaphor *cesspool* that refers to the community of Philadelphia together with the necessity of extralinguistic knowledge to understand the target of the message.

The other linguistic tool which is close to the extralinguistic knowledge by its effect and which is also used in implicit HS is context. The lack of context makes a message ambiguous and difficult to comprehend.

Context. The parts of a discourse that surround a word or passage and can throw light on its meaning (*Merriam Webster* n.d.).

35. *It 's a symbol demonstrating a lack of oxygen to the head of all its members .* (GPGC)

Example 35 illustrates the lack of context which influences the understanding of the meaning. If we consider this message in isolation with no indication of the source, it is impossible to grasp the target and the hateful part. On the other hand, if we have some previous messages in a thread, they may specify some absent information. In the case of this message, the whole corpus emphasizes the target, the corpus contains only racist messages. As well as extralinguistic knowledge, context is always used in line with other stylistic devices. In 35 a metaphor is implied by the phrase *a lack of oxygen*.

Now we would like to shed light on three associated stylistic devices: **metaphor**, **metonymy**, and **synecdoche** used in implicit hate messages. These linguistic tools are based on the replacement of an element by a related or close concept. Some consider synecdoche to be a type of metonymy. Others say that both metonymy and synecdoche are subconcepts of metaphor. In order to differentiate these notions, we will provide their definitions some examples and compare them.

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 (*Merriam Webster* n.d.).

36. *I'd say electrify the water but that would kill wildlife.* (HatEval)

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

The whole phrase 36 is a metaphor: the concept of water represents a country that suffers from immigrants where wildlife represents a native population of this country which will also be punished if they take drastic measures against immigration. However, this hateful message had an explicit hashtag which is usually kept after a preprocessing step (hashtag separation). As for 37, the part *monkey house* represents a metaphor used to denote the class full of blacks. As well as that, this example contains exaggeration of percentage.

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* n.d.).

38. *I retract - always .... Always because an animal lacks the skills to properly raise a (half) human child .* (GPGC)

39. *They're coming out of the woodwork to oppose white unity and White History Month.* (GPGC)

Example 38 illustrates an instance of metonymy: *an animal* replaces blacks, who, according to the author of a message, cannot upbring children. As for 39, we would postulate that this example shows both stylistic devices – metaphor and metonymy. *Woodwork* and *White History Month* represent two metonymies – a) work for lower classes and b) a period of white supremacy, respectively, while *white unity* seems to be a metaphor for the white society.

Synecdoche. A figure of speech by which a part is put for the whole (such as fifty sail for fifty ships), the whole for a part (such as society for high society), the species for the genus (such as cutthroat for assassin), the genus for the species (such as a creature for a man), or the name of the material for the thing made (such as boards for stage) (*Merriam Webster* n.d.).

40. *Sorry, but I have seen enough popes.* (CONAN)

Example 40 illustrates an instance of synecdoche expressed by the word *popes*, representing all Catholic priests and not only popes according to the context of religious HS.

Now we would like to turn to the use of humor to disguise hatefulness of messages. As **irony** and **sarcasm** may stand alone, they do not always need the support of other stylistic devices. This is also the case with **black humor**. All three are pretty strong to express hatefulness on the one hand and to make it implicit on the other.

Black humor. Humor marked by the use of usually morbid, ironic, grotesquely comic episodes (*Merriam Webster* n.d.).

41. *Their gods often had pale skin and blue eyes but I suppose literacy is n't valued highly by their cultural intelligencia .* (GPGC)

42. *If it 's brown , flush it down . If it 's black , send it back . It ai n't right unless it 's White .* (GPGC)

43. *They probably have a 150 I.Q. with all three of them combined .* (GPGC)

44. *He walks upright , and has human arms and legs .* (GPGC)

Examples 41, 42, 43 and 44 are instances of black humor. They are not subtle, as the context is clear and rough, but these messages are still implicit because there are no transparent indications of targets and hatefulness.

The following examples 45 and 46 are clear sarcastic cases. Hatefulness of the messages is hardly understood even with slight indications of context as *Whites* and *Jews*. As for 47, this sarcastic message contains hate expressed peculiarly: negating the Second World war facts, including Holocaust, it turns hatred towards Jews.

45. *“ Those evil <u>Whites</u> .* (GPGC)

46. *<u>Jews</u> are welthy because they are naturally good with money.* (CONAN)

47. *Nazi were not that bad as we depict them. EU should stop blaming them.* (CONAN)

The cases of irony are extremely complicated considering HS detection. 48 and 49 show neither targets nor lexicon, according to which hatefulness may be detected. As for the instance of HS in 50, irony is supported by exaggeration expressed in the phrase *may very will be living* which adds extra weight to the humorous context.

48. *We have become too cosmopolitan and tolerant.* (CONAN)

49. *They do what they always do take something of someone else 's and make it their own .* (GPGC)

50. *If not kept in check , we may very will be living in an actual Planet of the Apes soon .* (GPGC)

The other illustrative stylistic device used in implicit HS is euphemism.

Euphemism. The substitution of an agreeable or inoffensive expression for one that may offend or suggest something unpleasant (*Merriam Webster* n.d.).

51. *The goods that England offers should belong to us. Not to people who do not contribute to the state* (CONAN)

52. *UK Education Secretary: 'All children from the age of 4 will have mandatory lessons on same sex relationships.' – This is not about Education, this is about willingly confusing and indoctrinating kids and preventing them from developing their personalities.* (CONAN)

Example 51 illustrates an instance of euphemism expressed by the phrase *people who do not contribute to the state* which replaces rough address *migrants* or other offensive variants. As for 52, it is a mix of euphemisms *confusing*, *indoctrinating*, *preventing them from developing their personalities*, and circumlocution used in the second part of a message, where a shorter phrase could have conveyed the meaning.

Circumlocution. The use of an unnecessarily large number of words to express an idea (*Merriam Webster* n.d.).

The use of fallacy characterizes implicit HS as well. The example 53 is an instance of fallacy supported by exaggeration.

Fallacy. A false or mistaken idea; an often plausible argument using false or invalid inference (*Merriam Webster* n.d.).

53. *99% of Muslims in the UK are unemployable and refuse to work.* (CONAN)

It is not a very common case in the corpora we have analyzed, but it seems to be productive as there are a lot of facts that may be falsified and used against vulnerable groups of people.

The other important stylistic device is humiliation. In implicit hate messages, instances of humiliation are quite intricate and may be indirect, but the whole message allows us to understand the meaning and the target. The example 54 illustrates a clear and direct case of humiliation. However, it may be detected as non-hate by an automatic system as there are no insults or aggressive words. Thus, we consider it implicit HS. In the given example, humiliation is supported by exaggeration and subtle irony. The example 55 shows an instance of indirect humiliation. If we do not go deep into the meaning of the message, it seems to be quite positive on the surface: there are no negative words, and the allusion of children implies some positive effect. However, it has an ironic connotation that humiliates women. The example 56 illustrates an instance of humiliation that overlaps with sarcasm.

54. *A male need a woman. Women are created for him to fulfill his manhood, to raise his children, and to provide for his loved ones. And not to mention we can have sex to procreate and heal the world.* (CONAN)

55. *Women have it so easy. They stay home and have babies.* (CONAN)

56. *A woman is more equal when she is wearing a dress. This is what is fair: women can be good women when they respect men.* (CONAN)

Humiliation. The embarrassment and shame a person feels when someone makes them appear stupid or when they make a mistake in public (*Collins Dictionary* n.d.)

Now we would like to turn to pragmatics. Pragmatics is a term that covers the relationship of sentences to the environment in which they occur on the one hand and the relationship between signs or linguistic expressions and their users on the other (*Merriam Webster* n.d.). In other words, we would like to discuss those meanings of messages which may be implied by an author or depend on a contextual situation. We have already tackled the problem of extralinguistic knowledge, but two other features implicitly express hate.

Sentiment. An attitude, thought, or judgment prompted by feeling; the emotional significance of a passage or expression as distinguished from its verbal context (*Merriam Webster* n.d.).

57. *If only women behaved like men wanted and didn't fight back, 10% of all sexual assaults would be solved.* (CONAN)

58. *I like to listen to this preacher , but why do black prreachers always have to scream .* (GPGC)

59. *How much longer are we going to wait for them to take care of themselves ?* (GPGC)

60. *It is n't enough that we support mesico now we have to support africa .* (GPGC)

The example 57 becomes an instance of HS having women as a target if an author of a message adds some emotional weight that we call sentiment. Otherwise, it looks more like a toxic fact highlighting the superiority of men. 58 also depends on sentiment: if it is not very tense, we might consider this message toxic. To understand the example 59, we need some extralinguistic knowledge as we are unfamiliar with the context. As well as that, emotional weight influences the label, which may be either toxic or HS in case the sentiment is tense. As for 60, some indignation is inferred, but there are no clear indications of negative feelings. That is why understanding emotional weight is essential in this message.

The following examples illustrate instances of inference.

61. *I think it is a bit late to think to look after the safety and the future of white people in South africa .*

62. *France has been gone for a long time , we will have to retake it later because there will be no French people in not too long .*

63. *No wonder we are in such bad shape with this tribe in charge of the White House and Congress .*

Inference. Something that is inferred. The premises and conclusion of a process of inferring (*Merriam Webster* n.d.).

Examples 61 and 62 do not provide readers with the information that the author infers, so this information may be understood by having some contextual knowledge. As for 63, there is an instance of time inference. This message seems to be hateful in the case of that exact composition of the government.

Visual signs. The last significant group of properties characterizing implicit HS that has to be discussed unites visual signs: punctuation marks, quotes, and use of uppercase. These features cannot mark HS in isolation, but they play a role of support in many hate messages.

64. *The problems start when you start to accommonadate these people in the first place ! , give an inch and they 'll take a mile ......* (GPGC)

65. *Separation is the only answer !* (GPGC)

Examples 64 and 65 use exclamation marks to emphasize the tension of the topic discussed, while full stops in 64 express some suspension. It is evident that even without punctuation tools, the messages are hateful, as the first one uses an idiom in a negative sense and the second one sounds aggressive. However, punctuation appeals to the reader's attention more.

The following examples 66 and 67 illustrate the use of quotes. The first is an instance of black humor supported by quotation marks and an obvious sign in brackets, while the second is a mix of humiliation with black humor and suspension. Here the quotes express doubts about equality.

66. *If you want to hear what a black " intellectual " (lol) sounds like , check out the woman in this video trying to debate Dr. William Shockley .* (GPGC)

67. *Black "women"on London transport . . .* (GPGC)

The following two examples illustrate the use of uppercase letters. In 68, the word *White* starts with a capital letter to emphasize the superiority of white people, whereas 69 uses irony and metaphor to express hate supporting it with the capitalization of the word *MORE*.

68. *Lack of compassion seems to go hand and hand with lack the of White blood .* (GPGC)

69. *I suppose the concept of a male cow and a female cow producing MORE cows , and the repeating process is too much for them to understand .*

3.3 Common Properties of Implicit HS

This part represents a list of possible combinations of implicit hate speech properties. This compilation is based on the corpora analysis and the implicit properties described above. It may be helpful to have such a list during the annotation process to quickly find instances of HS and avoid memorizing all interlacements of implicit properties.

RHET + IRONY
RHET + METON
RHET + IRONY + METON

RHET + EXAGG

RHET + IRONY + EXAGG

RHET + SARC

RHET + METAPH

EXAGG

EXAGG + ANT

EXAGG + EXTRA

IRONY

IRONY + EXAGG

SARC

BH

BH + EXAGG

METAPH

METAPH + EXTRA

METON

METON + EXTRA

SYN

SYN + EXTRA

EUPH

EUPH + CIRC

FALL

FALL + EXAGG

HUM

HUM + EXAGG

HUM + IRONY

HUM + IRONY + EXAGG

HUM + SARC

SENT

SENT + EXTRA

SENT + EXTRA + RHET

SENT + IRONY

SENT + SARC
INF
INF + IRONY
INF + SARC
RHET = rhetorical question
METAPH = metaphor
METON = metonymy
SYN = synecdoche
EXAGG = exaggeration
SARC = sarcasm
BH = black humor
ANT = antithesis
EXTRA = extralinguistic knowledge
EUPH = euphemism
CIRC = circumlocution
FALL = fallacy
HUM = humiliation
INF = inference
SENT = sentiment

3.4 Subtle Properties

According to the definition of subtlety, it represents an indirect method to convey the meaning and understanding of that meaning requires extra knowledge and syntactic and grammatical proficiency. As well as that, both implicit and explicit hateful messages may be subtle since the meaning of explicit ones may be delivered in a delicate and elusive way, while implicit subtle ones are very hard to detect. The most complicated factor of subtlety is that it is rather based on the peculiarity of a message, thus, on the speaker's perception of a message. Subtlety does not treat literal meanings expressed directly by words, but rather the meaning which goes beyond the literal meaning.

Here we can talk about semantic ambiguity and pragmatics which are hard to study even by human beings.

Regardless of the fact that human perception is impossible to characterize and describe in a schematic way, we found certain criteria which influence message subtlety. The majority of them lie in the domain of syntax since a peculiar syntactic structure creates difficulties to get the true meaning. In the following part, we will describe the criteria of subtle messages, compare them to non-subtle ones and provide their examples. It has to be noted that in most cases these criteria overlap, creating more peculiar phrases.

Subject + State Verb + Subject + Verb + Object... This structure implies using state verbs such as ”think”, ”consider”, ”appear” (with the meaning ”to be”) etc. which do not add important semantic value, but add extra weight to a phrase.

Subject + Complex Predicate + Object... Under Complex Predicate we understand a predicate structure Verb+Verb which may be illustrated by a Verb + Gerund in English, a predicate structure Verb + Noun and other complex predicates containing more than one semantic verb (not considering auxiliary verbs).

Complex Subject + ... Complex Subject works in combination with other subtle properties since it cannot make a whole phrase intricate standing alone.

Complex structure. In this case we mean any complex syntactic structure that may prevent from clear understanding of the meaning.

II/III conditional + Complex structure. During our analyses of subtle phrases we found out that I conditional does not appear in subtle messages, while II and III conditionals are quite characteristic for subtlety. Regardless of this fact, they are used in combination with complex sentence structure, otherwise the phrase is not considered subtle.

NEG + POS + Complex structure. In this case a phrase is characterized by the negation at the beginning followed by a positive clause or vice versa and enhanced by a complex phrase structure mentioned above.

Coordinate + Subordinate Connectors. A multiple clause phrase that contains coordinate and subordinate clauses which make a phrase long and hardly comprehensible from the first glance.

Introductory constructions + . . . In this case a message has a long introductory part which does not convey important semantics, but rather makes a phrase heavy. Introductory constructions are always enhanced with other subtle properties.

Word order. The word order may prevent a reader from understanding the right relations between actors of a message from the first reading, thus, the true meaning may appear to be distorted.

Circumlocution. The use of an unnecessarily large number of words to express an idea (*Merriam Webster* n.d.). This criteria was also introduced among implicit properties since clear distinction between implicitness and subtlety is hardly amenable even to human analysis.

Here we provide some examples to illustrate subtle properties and their interelacements.

70. *It appears I cannot criticize islam for being a cult which encourages wars and takes over host nations by any means justified by their scriptures.*

The message 70 illustrates an overlap of the Subject + State Verb + Subject + Verb + Object. . . structure with POS + NEG at the beginning of the phrase and a Complex structure that makes it complicated to trace the relations between clauses. Regardless of the fact that this message is explicit, we consider it subtle due to its delicate manner to introduce the meaning.

71. *Is that couple in the second picture with the dark yellow background an Asian female with a White male or is it something that looks like designed to send the same message ?*

71 is an example of a Complex Subject at the beginning of a message and a Subject + Complex Predicate + Object. . . in the second part (*that looks like designed to send. . .*). As well as that, we could consider is it something as

the beginning of the Subject + State Verb + Subject + Verb + Object… construction which also adds extra weight to the message in question. Since there are no explicit indications of hate speech, the message is ironic and we need context to understand its hateful side, we consider this expression implicit and subtle.

72. *I suppose the concept of a male cow and a female cow producing MORE cows , and the repeating process is too much for them to understand .*

The message 72 illustrates an example of two Complex Subjects underlined and the whole sentence follows the structure Subject + State Verb + Subject + Verb + Object… where the second verb may be considered Complex Predicate (*is too much … to understand*). This message belongs to the implicit subtle category.

73. *I think it is a bit late to think to look after the safety and the future of white people in South africa .*

73 represents a case of a phrase with an Introductory construction, which may be removed without meaning loss, followed by circumlocution. Together with implicit properties (inference, irony etc.) these subtle criteria make this message implicit and subtle.

74. *If you had your way their would be no whites to pay taxes so the blacks could collect welfare and as they have proved in Africa they are incapable of feeding themselves* (GPGC)

75. *If Muslims counteracted child rapists in their communities instead of engaging in public discussions about Islam, one day Islam indeed would become a peaceful religion.*

Messages 74 and 75 illustrate II conditional. In 74 II conditional overlaps with Coordinate + Subordinate Connectors criteria since we have several different clauses (coordinate and subordinate) communicating with connectors _so_, _and_, _as_ and without connection in the last clause. 75 may be considered a case of circumlocution with II conditional. Both phrases are explicit and subtle.

76. _The issue is not whether Muslims should be *allowed* to go to those tribunals, but whether they should be *denied* a right which the rest of us have, viz._

As for the message 76, it represents the case of NEG + POS + Complex structure. The negation at the beginning of the phrase is opposed to the positive part _should be *allowed*_ which follows and its parallelism _should be *denied*_. Together with that, several clauses make the syntactic structure of the message complex.

77. _Someone Please Tell me what can be done or which people to tell that the Mexicans and or other hispanics here are undercutting price quotes on Good working People ._

77 represents a phrase with Coordinate + Subordinate Connectors which make it difficult to trace the relations between clauses. Thus, this message is considered subtle and implicit if we take its implicit properties (extralinguistic knowledge, irony etc.) into account.

78. _I'm 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 . "_ (GPGC)

The message 78 illustrates how the Word order may influence our understanding. From the first glance the part _how stupid the Jews_ seems to have a meaning different from what the phrase actually conveys if we read it completely. As well as that, we may notice circumlocution in the second part of the message.

If we compare subtle messages with non-subtle ones, we realize that non-subtle expressions possess less intricate and more simple syntactic structures.

For instance, SVO, VO and concatenation of both together with simple active verbs are characteristic of non-subtlety. As well as that, many non-subtle messages start with There is/There are or Demonstratives, while such property is not detected among subtle cases.

4 Border Cases

This part is devoted to border cases of implicit HS. Analyzing HS and other corpora that contain abusive messages, we found many peculiar examples which merit our attention. Among them, some messages do not seem hateful, but rather toxic, regardless of the fact that they are annotated as HS in studied corpora. Moreover, there are cases which are difficult to place in one group: on the one hand, they seem to be implicit as they do not have insults or cursing, on the other hand, they do not possess any properties described in the previous part, so the hateful meaning is rather clear. As well as that, there are some messages annotated as HS, but the targets of those cases do not belong to protected groups. Thus, it is questionable whether we have to consider them in the framework of our project.

The following examples illustrate border cases of HS.

79. *Most of these videos belong on the Animal Planet tv channel .* (GPGC)

80. *Hispanics , Blacks and Asians are all doing it .* (GPGC)

81. *Well I wonder if its started messing with the daughter yet .* (GPGC)

82. *Jews love to gamble* (CONAN)

83. *We cannot allow people to come in from certain countries.* (CONAN)

84. *Jews are a different race.* (CONAN)

85. *”“When there is an empty space, people look to fill it ...”The modus operandi of the Catholic Church in a nutshell. If you don’t fill your own soul, they have an app for that!*

The example 79 illustrates a case of HS which is complicated to detect: first, it needs some extralinguistic knowledge on the videos mentioned, second, having this complementary information, especially on the target of the message (blacks), is hardly possible to decide if the message is hateful or toxic. We placed it in the group of implicit HS as the ironic and sentimental weight seems to be hard to withstand. As well as that, we should note that the level of hatefulness of this message depends on the sentiment implied by the author and on the perception of the reader. As for 80, this message does not seem to be hateful at first glance as there is no context given and no violence, cursing, or aggressive action expressed. However, if we know what the author is talking about, this message starts talking itself, showing hatred towards non-white races. We can say the same about 81: if we know the context of a message, we understand that this sarcastic phrase is full of hatred towards non-white immigrants. If we analyze the example 82, we notice that this message appeals to our attention by the absence of any hateful or even toxic signal. Although, semantic and pragmatic analysis of this message allows us to understand that the connotation of the verb *to gamble* does not mean to play games of chance but to trade on countries and people's lives. This "fact" emphasized by such a short phrase with the verb *to gamble* expresses hatred towards Jews for their actions. On the one hand, if we analyze 83 and 84, we observe that there is no expression of hate or any indignation in those messages. One sounds like a rule, while the other looks like a fact. On the other hand, the targets of these messages are clear, and the sense that their authors wanted to convey is comprehensible by readers, so they are considered implicit HS. The example 85 illustrates a peculiar case of irony with the lack of extralinguistic knowledge that may seem to be rather toxic than hateful. However, the acidity of this message lets us consider it implicit HS, and the level of hatefulness perceived depends on the reader's view.

As we have mentioned in the description of implicit properties, the lexicon used in messages may help to detect instances of HS, but it is not an exclusive criterium that we can depend on. We have to take into account representative words as *Hispanics*, *Jews*, *Catholic*, and other underlined ones in the examples studied. However, it is incorrect to overlook implicit properties.

The following examples illustrate border cases of implicitness/explicitness.

86. *Then we will see who is afraid to leave thier homes after dark or venture into the wrong ’ hood .* (GPGC)

87. *And this is why SA went from First World under Botha ’s rule (deKlerk does n’t count) to a Third World begger horde .* (GPGC)

88. *the entire system of classes are a marxist invention. the judeo-bolsheviks are using this to divide our society. in other words ’ Divide et impera ’ .* (GPGC)

89. *We spent too much time and effort worrying about Blacks while the Asians and Hispanics , were the ones carrying the ball .* (GPGC)

90. *Blacks and mexicans were represented thoroughly. do yourselves a favor , shoot your t.v.* (GPGC)

91. *”The UK Government recently published a paper on ‘Online Harms’ – in other words they are going to be working with Jewish and Muslim anti-hate organisations to censor the Internet. (Referring to Tell Mama, CST and Stop Hate UK)”* (CONAN)

92. *The Koran says that men are superior to women.* (CONAN)

The example 86 is a clear menace that may be considered an instance of explicit HS, but the way an author conveys it is implicit, especially for a machine learning system. As for 87, it contains explicit toxic indications such as *begger horde*, which let us consider this message explicit on the one hand, but the target of the message is hidden since it is not clear if the message is aimed at the government or the citizens on the other hand. That is why we consider this message implicit HS with a question mark. Looking at the example 88, we may suppose that it is an instance of explicit HS according to the words *divide* and the phrase *Divide et impera*. However, if we delve into the phrase, it seems that an instance of hatred towards Jews is intricately hidden under the cover of disgrace of marxism, so this example is rather implicit. Now we would like to study two messages 89 and 90. These messages are also border cases as they have explicit rough phrases *were the ones carrying the ball* and *do yourselves a favor , shoot your t.v.* , but on the other hand, these parts are not insults or humiliation and we cannot assure that the machine learning system will be able to indicate them as hateful. If we look at the message 91, it seems rather explicit because of the lexicon. However, it is not as simple because the use of words does not precisely indicate the target of disgrace. Moreover, the phrase is a double-edged sword since it has two levels of meaning: hatred towards the government on the one hand and ethnicity-based hate on the other. Taking these considerations into account, we mark this message as implicit HS but with a question mark. The message 92 is an interesting example of implicit HS where the target may be perceived differently. If we look at this case without delving into the meaning, the surface target is women. However, this message represents an instance of implicit HS towards Muslims as the Koran is mentioned negatively, humiliating women.

The following examples illustrate cases of hate towards groups that are not considered "protected" according to the HS definition. These messages are toxic, offensive, or aggressive and may not be studied in the framework of this project, but they also merit our attention.

93. *well, the liberals would go ape crazy .* (GPGC)

94. *#Conservatives Govt have run up debt in spite of austerity cuts <u>while the rich have doubled their wealth</u>. #inequality URL via @USER*

The example 93 may be studied in two different ways: from the political point of view, this message represents a toxic phrase against the political movement of liberalism, but from the human position, it may imply certain characters of the liberal movement, thus, may even involve a race characteristic. As we do not have enough information about the author's inference, we do not consider this message as implicit HS. As for 94, this message targets the government, but we also notice an instance of indignation towards people with high income: *while the rich have doubled their wealth*. So if we take into account this part which is aimed at people and not at a political organization, we may consider this message implicit HS.

5 Corpora

Throughout our analysis, we revised messages of different corpora in order to gain knowledge and build these guidelines. Many of these messages were used as examples of Section 2, Section 3 and Section 4. We provide a table with the corpora used below.

Note that the Notation column is used to reference corpora and messages. This notation is based on Poletto et al. (2021), a survey of resources and corpora for HS detection. It indicates the initial letter of the authors, last names, or the name of the resource. These datasets have an associated article with its description indicated by the column *Article*. We provided the *URL* to websites where these datasets can be downloaded. We also revised other corpora guided by the survey in Poletto et al. (2021), but they were not included in this table due to their lack of implicit messages or unavailability of the dataset.

Notation	Article	URL
CONAN	Chung et al. (2019)	`https://github.com/marcoguerini/CONAN`
BVSAS	Bohra et al. (2018)	`https://github.com/deepanshu1995/HateSpeech-Hindi-English-Code-Mixed-Social-Media-Text`
DWMW	Davidson et al. (2017)	`https://github.com/t-davidson/hate-speech-and-offensive-language`
GPGC	Gibert et al. (2018)	`https://github.com/Vicomtech/hate-speech-dataset`
OffensEval	Zampieri et al. (2019)	`https://sites.google.com/site/offensevalsharedtask/olid`
IRE-Project-hatEval-2019	*HatEval* (n.d.)	`https://github.com/ash0904/IRE-Project-hatEval-2019`

Table 2: List of Corpora

References

Agarwal, Swati and Ashish Sureka (2015). "Using KNN and SVM Based One-Class Classifier for Detecting Online Radicalization on Twitter". In: *Distributed Computing and Internet Technology*. Ed. by Raja Natarajan, Gautam Barua, and Manas Ranjan Patra. Cham: Springer International Publishing, pp. 431–442. ISBN: 978-3-319-14977-6.

Awan, Imran and Irene Zempi (2020). "A Working definition of Islamophobia". In: *Preparation for the report to the 46th Session of Human Rights Council*. URL: `https://www.ohchr.org/sites/default/files/`

Documents / Issues / Religion / Islamophobia - AntiMuslim / Civil%20Society%20or%20Individuals/ProfAwan-2.pdf.

Bohra, Aditya et al. (June 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*. New Orleans, Louisiana, USA: Association for Computational Linguistics, pp. 36–41. DOI: 10.18653/v1/W18-1105. URL: https://aclanthology.org/W18-1105.

Chiril, Patricia (Nov. 2021). "Automatic Hate Speech Detection on Social Media". Theses. Université Paul Sabatier - Toulouse III. URL: https://tel.archives-ouvertes.fr/tel-03599458.

Chung, Yi-Ling et al. (July 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*. Florence, Italy: Association for Computational Linguistics, pp. 2819–2829. DOI: 10.18653/v1/P19-1271. URL: https://www.aclweb.org/anthology/P19-1271.

Collins Dictionary (n.d.). URL: https://www.collinsdictionary.com.

Combating Sexist Hate Speech (n.d.). URL: https://rm.coe.int/1680651592.

Corazza, Michele et al. (May 2020). "A Multilingual Evaluation for Online Hate Speech Detection". In: *ACM Transactions on Internet Technology* 20.2, pp. 1–22. DOI: 10.1145/3377323. URL: https://hal.archives-ouvertes.fr/hal-02972184.

Davidson, Thomas et al. (2017). *Automated Hate Speech Detection and the Problem of Offensive Language*. arXiv: 1703.04009 [cs.CL].

Facebook: HS policies (n.d.). URL: https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/.

Fortuna, Paula and Sérgio Nunes (July 2018). "A Survey on Automatic Detection of Hate Speech in Text". In: *ACM Comput. Surv.* 51.4. ISSN: 0360-0300. DOI: 10.1145/3232676. URL: https://doi.org/10.1145/3232676.

Fortuna, Paula, João Rocha da Silva, et al. (Aug. 2019). “A Hierarchically-Labeled Portuguese Hate Speech Dataset”. In: *Proceedings of the Third Workshop on Abusive Language Online*. Florence, Italy: Association for Computational Linguistics, pp. 94–104. DOI: 10.18653/v1/W19-3510. URL: https://aclanthology.org/W19-3510.

Gibert, Ona de et al. (Oct. 2018). “Hate Speech Dataset from a White Supremacy Forum”. In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Brussels, Belgium: Association for Computational Linguistics, pp. 11–20. DOI: 10.18653/v1/W18-5102. URL: https://aclanthology.org/W18-5102.

HatEval (n.d.). URL: https://docs.google.com/document/d/11BiV8JceusuAu4oxQS3d6KFdXlQg-w-jbZJiMngHo/edit.

International Holocaust Remembrance Alliance (n.d.). URL: https://www.holocaustremembrance.com/resources/working-definitions-charters/working-definition-antisemitism?focus=antisemitismandholocaustdenial.

Jurgens, David, Libby Hemphill, and Eshwar Chandrasekharan (July 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*. Florence, Italy: Association for Computational Linguistics, pp. 3658–3666. DOI: 10.18653/v1/P19-1357. URL: https://aclanthology.org/P19-1357.

Leonardelli, Elisa et al. (2021). *Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators’ Disagreement*. DOI: 10.48550/ARXIV.2109.13563. URL: https://arxiv.org/abs/2109.13563.

Merriam Webster (n.d.). URL: https://www.merriam-webster.com.

Poletto, Fabio et al. (2021). “Resources and benchmark corpora for hate speech detection: a systematic review”. In: *Lang. Resour. Evaluation* 55, pp. 477–523.

Waseem, Zeerak et al. (Aug. 2017). “Understanding Abuse: A Typology of Abusive Language Detection Subtasks”. In: *Proceedings of the First Workshop on Abusive Language Online*. Vancouver, BC, Canada: Association for Computational Linguistics, pp. 78–84. DOI: 10.18653/v1/W17-3012. URL: https://aclanthology.org/W17-3012.

Wulczyn, Ellery, Nithum Thain, and Lucas Dixon (2017). *Ex Machina: Personal Attacks Seen at Scale*. arXiv: 1610.08914 [cs.CL].

Zampieri, Marcos et al. (June 2019). “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)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 1415–1420. DOI: 10.18653/v1/N19-1144. URL: https://aclanthology.org/N19-1144.