

UTILIZING AI AGAINST HATE SPEECH

A guide to annotation,
classification, and detection



HDV
PUBLICATIONS

HRANT DINK FOUNDATION

Hrant Dink Foundation was established after the assassination of Hrant Dink in front of his newspaper Agos on January 19, 2007, in order to avoid similar pains and to continue Hrant Dink's legacy, his language and heart and his dream of a world that is more free and just. Democracy and human rights for everyone regardless of their ethnic, religious or cultural origin or gender is the Foundation's main principle.

The Foundation works for a Turkey and a world where freedom of expression is limitless and all differences are allowed, lived, appreciated, multiplied and conscience outweighs the way we look at today and the past. As the Hrant Dink Foundation 'our cause worth living' is a future where a culture of dialogue, peace and empathy prevails.

UTILIZING AI AGAINST HATE SPEECH: A GUIDE TO ANNOTATION, CLASSIFICATION, AND DETECTION

ISBN 978-605-71835-9-0



editors

Tirşe Erbaysal Filibeli, Tunga Güngör

project team

İnanç Arın, Didar Akar, Başak Can, Somaiyeh Dehghan, Elif Erol, Burak Işık, Sıla Kartal, Buket Kapısız, Yasemin Korkmaz, Arzucan Özgür, Nural Özel, Pelin Önal, Gökçe Uludoğan, Ayşecan Terzioğlu, Murat Tercan, İrem Topçu, Tuğba Özsoy, Berrin Yanıkoğlu, Elif Yararbaş, Umut Şen

publication coordinators

Başak Can, Elif Erol, Buket Kapısız, Yasemin Korkmaz, Pelin Önal, Tuğba Özsoy, Elif Yararbaş

translators

Burcu Becermen, Simon Charles Popay

proofreader

Neil Patrick Doherty

design and data visualisation

Yasemen Cemre Gürbüz

graphic application

Selin Uluer

printed by

Sena Ofset Ambalaj Sanayi ve Ticaret. Ltd. Şti.
Yakuplu Mh. 194. Sk. 3. Matbaacılar Sit. N:1 D:465
Beylikdüzü İstanbul/Türkiye
T: (212) 613 38 46

Istanbul, March 2025



© Hrant Dink Foundation Publications

Anarad Hıçutyun Binası Papa Roncalli Sokak No: 128
Harbiye, 34373 Şişli, İstanbul
T: 0212 240 33 61
info@hrantdink.org
www.hrantdink.org

Utilizing Digital Technology for Social Cohesion, Positive Messaging and Peace by Boosting Collaboration, Exchange and Solidarity project is supported by the European Union and the Friedrich Naumann Foundation. The Hrant Dink Foundation is solely responsible for the content in the publication, which does not reflect the views of the supporters.



FRIEDRICH NAUMANN
FOUNDATION For Freedom.
Türkiye



HRANT DINK VAKFI
HRANT DINK FOUNDATION
ՀՐԱՆԹ ԴԻՆԿ ԶՆՏԱՎԱԿ



Sabancı
Universitesi



UTILIZING AI AGAINST HATE SPEECH

A guide to annotation,
classification, and detection



HDV
PUBLICATIONS

CONTENTS

| | |
|---|---|
| Introduction | 7 |
| The “Utilizing Digital Technology for Social Cohesion, Positive Messaging and Peace by Boosting Collaboration, Exchange and Solidarity” project | 7 |

HATE SPEECH LABELING GUIDE

10

| | |
|--|-----------|
| 1. How do we decide whether discourse is hate speech? | 11 |
| 2. Labeling interface | 12 |
| 2.1. Determining the target group | 12 |
| 2.2. Determining the type of speech | 16 |
| 2.3. Determining the category of hate speech | 17 |
| 2.4. Assessing challenging examples | 34 |
| 2.4.1. Tweets containing hashtags and emojis | 34 |
| 2.4.2. Tweets that include/quote the speech of others | 35 |
| 2.4.3. Sarcastic content | 37 |
| 2.4.4. Covert hate speech | 39 |
| 2.5. Additional labeling headings | 40 |
| 2.5.1. Language of tweets | 40 |
| 2.5.2. Identifying the hate speech span | 40 |
| 2.5.3. Hate speech strength | 41 |
| 2.5.4. Offensive language | 42 |

DEVELOPING THE AI MODEL

46

| | |
|--|-----------|
| 1. Data collection & annotation | 47 |
| 1.1. Data collection | 47 |
| 1.2. Annotation | 48 |

| | |
|---|-----------|
| 2. Developed AI tool for detecting and measuring hate speech | 52 |
| 2.1. Data preprocessing and paralinguistic features | 52 |
| 2.2. Hate speech detection and classification | 53 |
| 2.3. Hate speech strength prediction | 53 |
| 2.4. Target identification | 54 |
| 2.5. Specific group identification | 55 |
| 2.6. Span detection | 56 |
| 2.7. Hate speech detection in Turkish print media | 57 |
| 2.8. Media tracking & analysis | 58 |
| 3. Error analysis | 59 |
| 4. Limitations | 65 |
| CONCLUSION | 68 |
| Appendix A: Labeling interface | 70 |

Introduction

Increasingly widespread in both traditional and social media, hate speech is reaching a worrying level for both social cohesion and peace, especially in times of crisis. Although limited in traditional media by professional codes, legal regulations, and in-house rules, hate speech circulates and spreads very quickly on social media platforms due to the rapid flow of information and high levels of interaction, despite attempts to control, monitor, and restrict it. While social media platforms have established some internal regulations, their use of algorithms developed to prolong user engagement in the system means that they cannot develop effective control mechanisms that serve both profit and social benefit. For this reason, civil society organizations and academics from different disciplines all over the world have come together to start working on mechanisms aimed at combating hate speech. In Turkey too, the Hrant Dink Foundation (HDF), which has undertaken numerous activities against the production, circulation, and dissemination of hate speech in digital media, has partnered with Boğaziçi and Sabancı Universities to develop a **hate speech detection and classification tool** for use on social media.

This report was prepared by 13 researchers from the HDV ASULIS Discourse, Dialogue, Democracy Laboratory and departments of computer engineering, linguistics, and cultural studies at two universities. It introduces the hate speech detection tool *pari* that has recently been developed, as well as other activities carried out within the scope of the project **Utilizing Digital Technology for Social Cohesion, Positive Messaging and Peace by Boosting Collaboration, Exchange and Solidarity**. The report outlines the accepted definition of hate speech, as well as the target groups and keywords selected for the collection of data used to train the detection and classification tool. It explains each hate speech category with examples, and addresses questions that may arise in the labeling procedure. In addition, it presents the model developed for the detection, classification, and rating of hate speech using the collected data.

The “Utilizing Digital Technology for Social Cohesion, Positive Messaging and Peace by Boosting Collaboration, Exchange and Solidarity” project

The **Utilizing Digital Technology for Social Cohesion, Positive Messaging and Peace by Boosting Collaboration, Exchange and Solidarity** project began

in 2022 and aims to combat hate speech, discrimination, and disinformation in the digital sphere by establishing cooperation between different fields such as linguistics, computer science, social sciences, the information sector, and civil society. The main output of the project is a hate speech detection and classification tool that will detect online hate speech using artificial intelligence technology.

Detecting hate speech using traditional methods is a process that largely relies on human labor. Rapid increases in the use of and number of users on digital media platforms and the unsustainability of labor-based media monitoring efforts have played an important role in the emergence of this project. Although social media companies have begun combating hate speech and disinformation by means of policy changes, detailed information in user agreements, and various other projects, these changes and actions only aim to protect their platforms' own interests. As such, hate speech needs to be independently and objectively detected and examined by rights-based civil society organizations and academic institutions, in order to move beyond the limits of hate speech as determined by companies, to better understand the roots of hate speech, and to pave the way for effective and scientific measures to combat it. Within the scope of the **Utilizing Digital Technology for Social Cohesion, Positive Messaging and Peace by Boosting Collaboration, Exchange and Solidarity** project, the digital tool *pari* has been developed that uses new technologies and artificial intelligence to detect and combat **ethnic, religious, and gender-based discrimination** and to detect hate speech targeting such groups. By making it an open source, it is hoped that the project's automatic hate speech detection tool will contribute to hate speech monitoring efforts and provide a more effective and sustainable solution in the fight against hate speech and discrimination.

Many groups in Turkey are targeted by hate speech in different forms. Academics, researchers, civil society organizations, policy makers, professional organizations, and similar institutions working on **regular and irregular migrants, refugees, asylum seekers, minorities (legally recognized and not), ethnic and religious groups, women, LGBTI+ individuals, the disabled**, and similar vulnerable groups are the ultimate beneficiaries of this project.

Posts from the social media platform now called X¹ were collected during the project in order to develop the tool. The platform is suitable for this study because it is a user-generated content platform (i.e., it is based on users sharing instant written content on current events), also because text-based social media studies is a suitable subject for academic research, and finally because the platform is frequently preferred as a site for political discourse.

¹ In June 2023, the social media platform known as Twitter changed its name to X. The platform is referred to as X in this report.

In order to collect data, content targeted against Jews, Greeks, Arabs, Alevis, Armenians, Kurds, LGBTI+ individuals, and refugees was extracted using determined hashtags and keywords through X's academic API and scraping. These hashtags and keywords were selected through regular monitoring of current events and included groups frequently subjected to hate speech in Turkey. A total of **16,254 tweets were labeled and each of them was labeled by three different people**. Before labeling, each annotator underwent the same training to reduce potential differences and also to establish a consistent labeling process. The data used to develop the tool are the result of this collective effort.

1 HATE SPEECH LABELING GUIDE

Big data is important for artificial intelligence and deep learning algorithms. In order to train the artificial intelligence tool developed within the scope of the project, Turkish social media content was gathered and labeled according to whether it contained hate speech, as well as the category and severity of hate speech. In this guide real tweets are given as examples. Usernames and other identifying information have been removed. In addition to the tweets, 10 years worth of data from the Hrant Dink Foundation's Media Watch on Hate Speech project², aimed at hate speech in the written press, was used to create a data repository of examples for the development of the tool. Since the Middle East and North Africa (MENA) region was targeted within the scope of the project, the same actions were carried out on a smaller scale for Arabic³.

This guide was prepared within the scope of a study conducted on content shared on X, with the purpose of explaining how the collected data was labeled and to serve as an example for future research. If data from other platforms are used, relevant changes must be made.

2 <https://hrantdink.org/en/asulis/activities/projects/media-watch-on-hate-speech/420-media-watch-on-hate-speech>

3 Within the scope of this project, both Turkish and Arabic tweets were labeled to train the tool. However, due to space constraints, only English translations are included in the report.

1. HOW DO WE DECIDE WHETHER DISCOURSE IS HATE SPEECH?

There is no universally accepted, unchanging definition of hate speech. Therefore, studies on hate speech are based on different definitions of hate. In the context of this project, the definition of hate speech set forth in the Council of Europe’s Recommendation of the Committee of Ministers on Hate Speech (1997)⁴ is taken as the basis:

“...the term “hate speech” shall be understood as covering all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin.”

In addition to this definition, many different factors should be considered when deciding whether something is hate speech, such as the characteristics of the society and language in which the speech is made, the context, as well as current events. Three basic questions are important in deciding whether speech is hate speech:

- What group or identity is mentioned in the utterance?
- What approach does the utterance take toward group or identity?
- What are the possible effects and consequences of the utterance? (Could it lead to human rights violations?)

In line with the adopted definition of hate speech, the following guidelines have been established:

- If the tweet contains speech that directly targets a national, ethnic, religious or gender identity that is hostile, discriminatory, or incites polarization, it should be labeled as **hate speech**. Bearing freedom of expression in mind, speech that does not directly target a national, ethnic, religious, or gender identity should be labeled as **‘There is no hate speech’**.
- If there is *covert* hate speech in the tweet, that is, if the tweet itself does not appear to have hate speech, but the annotator understands it as such from the context, it should be labeled as **hate speech**.

⁴ Council of Europe. 1997. *Recommendation No. R (97) 20 of the Committee of Ministers to Member States on “Hate Speech”*. In *Recommendations and Declarations of the Committee of Ministers in the Field of Media and Information Society*, 106–108. Strasbourg: Council of Europe.

- For data accuracy, where it cannot be decided whether the speech is hate speech or not, it should be marked as **“Not sure”**. (Tweets marked as such are collected and re-assessed.) If marking ‘There is no hate speech’ or ‘Not sure’, the other parts of the labeling form **should not be left blank**.

Images, emojis, hashtags, tags, inverted sentences, abbreviations, sarcastic comments, etc. in content shared on social media can make it difficult to decide whether the content is hate speech. In such cases, the context of the post becomes important. Detailed guidance and the decisions we made in labeling challenging examples (such as covert hate speech) are discussed in more depth in the following sections. Our labeling interface was created based on the accepted definition of hate speech and the three basic questions above and is explained step by step below.

2. LABELING INTERFACE

2.1. Determining the target group

The tweets captured to create the dataset were selected based on certain keywords and hashtags used for identities frequently targeted by hate speech. The determination of the target group is located in the interface under the heading **“Overall attitude and stance”**. As there is more than one target group, the identity description in the “Overall attitude and stance” section in the interface changes according to the identity in the data set. For example, keywords and hashtags such as “#birgeceansızıngelebiliriz” (‘we might come suddenly one night’), “#Yunankaşınıyor” (‘Greece is asking for trouble’) and “denize dökmek” (‘to cast into the sea’) are used to detect hate speech against the Greeks. Accordingly, the label “anti-Greek” appears in the “Overall attitude and stance” section of the dataset of captured tweets. Other labels include “anti-Armenian”, “anti-LGBTI+”, “anti-Alevi”, and similar expressions. In order to proceed with a single example in this guide, the explanation is provided using ‘anti-Semitic’ examples.

The options available in this section are:

- Not sure
- Anti-Semitic
- Freedom of expression (neutral)⁵
- Irrelevant

⁵ The labeling interface offers a “Neutral” option for expressions that fall within the scope of freedom of expression, this is also stated in the report.

If a decision cannot be made about whether a tweet is anti-Semitic or not, or if there is uncertainty about the overall attitude of a statement directed at another identity, it should be labeled **“Not sure.”**

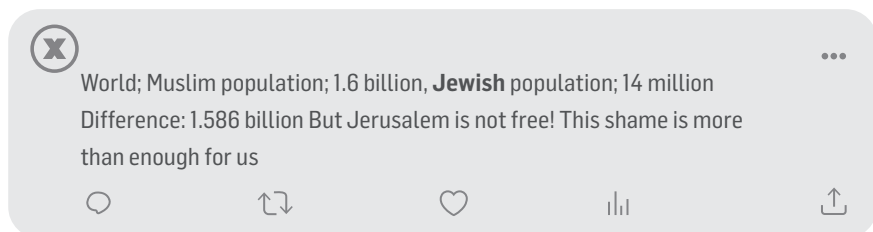
If the tweet is deemed to contain discriminatory discourse or hate speech against Jews, the **“Anti-Semitic”** option must be selected. For tweets targeting other target groups, there are also options such as “Anti-Refugee”, “Anti-Greek”, “Anti-LGBTI+” in this section of the labeling.

If it is deemed that the wording of the tweet aimed at the target group in question is neutral, does not contain hate speech, or does not target the identity in question as a whole, it should be considered as **“Neutral.”**

If the tweet is considered to not contain hate speech towards the Jewish identity and its content is irrelevant, the **“Irrelevant”** option should be selected. In addition, if the tweet does not contain hate speech towards Jews but is considered to contain hate speech towards a different identity, it should be labeled as “Irrelevant”. For example, when labeling a dataset prepared in relation to Jews, a tweet that does not contain hate speech towards this group but contains hate speech towards LGBTI+ people should first be labeled as “Irrelevant” because it is outside the relevant topic, and then the ‘Sexual Orientation’ option should be selected in the ‘target group’ section of the labeling.

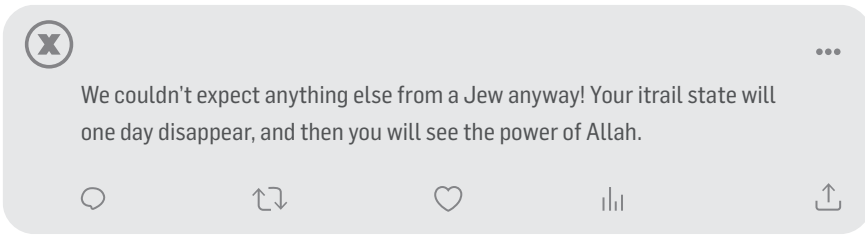
Below is an example of each option in “Overall attitude and stance”:

Not sure



In the example above, figures are given to compare the Muslim and Jewish populations, and the difference in them is related to the situation in Jerusalem. However, it cannot be fully understood whether this tweet is Anti-Semitic. In this case, the option “Not sure” should be marked.

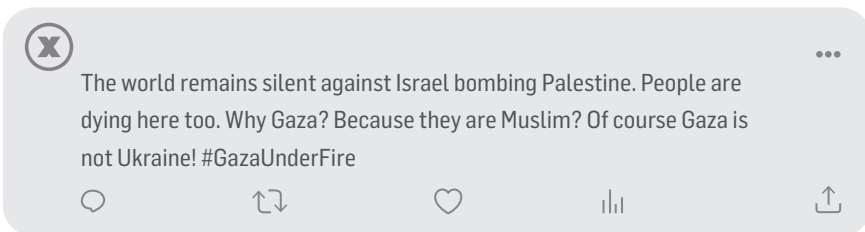
Anti-Semitic



In the Israel–Palestine dataset, the phrase “itrail,” as a slang reference to Israel that refers to the state as a dog, is one of the most frequently encountered phrases. However, how we assess this phrase depends on the content of the tweets. In this tweet there are two related points. First, in this example, the Jewish people and Israel are seen as an undivided whole. This attitude identifies the actions of Israel with all Jews who are targeted through insults, swearing, defamation, and dehumanization, and this constitutes hate speech. Second, there are references to negative characteristics associated with the Jewish identity. In the eyes of the writer of the above, ruthlessness (or similar characteristics) is shown to be an inherent and innate characteristic of Jews.

Freedom of expression (neutral)

Apart from these, statements that do not express discrimination, prejudice, enmity or hatred towards Jews collectively, but simply report on, criticize or express sadness towards a certain event or practice, should not be considered “anti-Semitic.” Since they do not contain hate speech, they should be marked as “Neutral,” because they are outside the main purpose of the study.

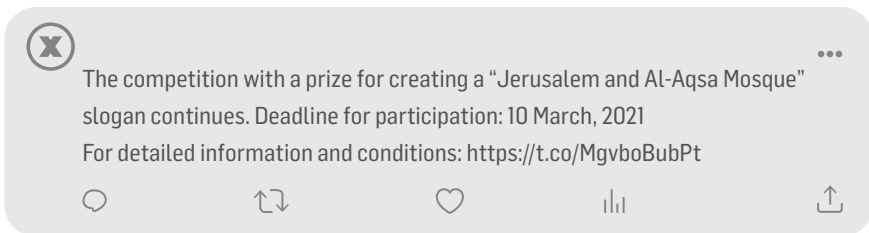


In the above tweet, there is no hate speech towards Jews collectively. The tweet emphasizes Israel’s attack on Gaza following Hamas’ attack on Israel on October 7, 2023, and the ongoing war in which civilians are currently being killed. In referring to the bombing, the disproportionate force used by Israel is criticized, as is the

silence of other states in the world in relation to this. In addition, by referring to the war between Ukraine and Russia, it is emphasized that other states do not act as they did in opposition to the occupation of Ukraine and the ongoing war there, and thereby discrimination is implied. Furthermore, the comment that states that the inhabitants of Gaza are Muslim is an indication of the reason for this discrimination. Since only a specific incident and inequality in the response to it are criticized, and since it does not amount to hate speech directed at the Jewish community collectively, this tweet should be marked as freedom of thought and expression.

Irrelevant

We cannot predict whether this contest will contain anti-Israeli or anti-Semitic statements, and the tweet itself does not contain any trace of anti-Semitism. Therefore, the option “Irrelevant” must be marked.



There may be a single target group or multiple target groups in tweets, and in some tweets the target group may not be clearly identifiable. In such cases, when labeling tweets:

- In tweets with a **single target group**, the **relevant group** must be selected. If **more than one group** is targeted, all targeted groups must be selected.
- **If the tweet does not contain hate speech**, but it still expresses a view on the relevant identity group, the target group should be marked **freedom of expression (neutral)**. Only in tweets that contain hate speech should the relevant identity be marked as the target group.
- **Even if there is no clearly stated target group**, in order to both protect freedom of expression and reduce the risk of “false positives,” it should be marked that **a target group is present**. (For example, even if terms such as refugee, Syrian, or Afghan are not directly used in the speech, there may actually be a covert/hidden target group in the collected tweets about refugees. Therefore, a target group should be selected.)

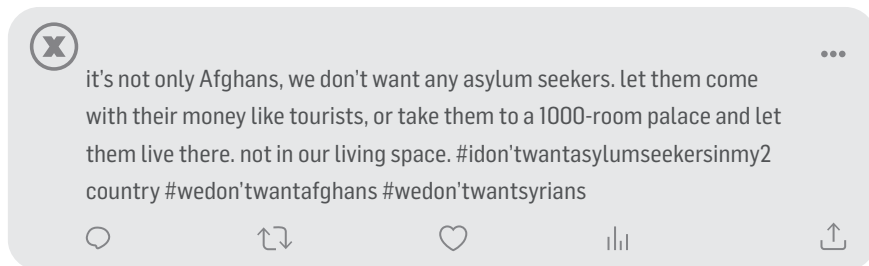
- **Demographic/Socioeconomic group:** Hate speech under this category should be marked according to whether it targets **race/ethnicity, country/nationality, religion, gender, or sexual orientation**. This group refers to instances where a whole group is targeted based on these characteristics or where a person or group is targeted based on the identity characteristics listed here. In speech where more than one identity characteristic is targeted, all relevant characteristics should be marked.
- In some utterances, it is not clear whether the speech is **directed at a target group or an opinion group**. In such cases, it may be difficult to select the hate speech category. For example, in some cases, it may be unclear whether the writer of the tweet is anti-Zionist or anti-Israeli and/or anti-Jewish. In these tweets, **'Target group is unclear or absent'** should be selected.
- If more than one target group is labeled, two different options in the demographic/socioeconomic section can be marked, such as country/nationality and religion.

2. 2. Determining the type of speech

After the target group has been identified, information about how these groups are targeted is detailed. Here, it must first be stated in the interface whether the tweet contains discriminatory discourse or hate speech.

Exclusionary / discriminatory discourse

This is discourse in which an entire group or some members of a group are seen negatively as different from the dominant group due to their identity, with respect to matters such as inclusion in society and benefiting from rights and freedoms.



This discourse, which opposes all people who come to the country seeking asylum, should be labeled as exclusionary/discriminatory discourse, because it does not recognize rights such as asylum and shelter for this group.

Hate speech

If a tweet is hate speech, hate speech categories are used to analyze how the speech targets a group. A tweet should be labeled according to the hate speech categories determined within the scope of the project and specified below. If a tweet fits more than one hate speech category, **more than one category should be selected**. The person labeling should try to choose one category whenever possible, but cross-labeling is still possible if there are examples of hate speech that fall into different categories when there is more than one target group.

2.3. Determining the category of hate speech

By taking advantage of international scientific studies on this subject and taking into account country-specific language and cultural differences, categories based on how the speech targets a group play a functional role as a unit of analysis in understanding and explaining how the content of the text in question constitutes hate speech. In the HDV Media Watch on Hate Speech project, hate speech is divided into four categories:

- 1) Exaggeration/attribution/distortion/generalization:** Speech that includes negative generalization, distortion, exaggeration, or negative references towards a group of people on the basis of a person or event.

In this category, hate speech is most often produced through generalization.

(e.g. ‘Suriyeliler gına getirdi’, ‘I’m fed up with Syrians’; ‘Yunan ölüme terk etti’, ‘The Greek left them to die’; ‘Yahudi havadan saldırdı’, ‘The Jew attacked from the air’; ‘Eşcinsel sapkınlar dehşet saçıyor’, ‘Homosexual perverts are spreading terror’; ‘Ermenilerin tazminat ve toprak hayalleri suya düştü’, ‘The Armenians’ dreams of compensation and land have been smashed’; ‘Hristiyan terörünü İslam’a maletti’, ‘They attributed Christian terror to Islam’; ‘Akdeniz’de Rum Gerilimi’, ‘Orthodox Greeks Pushing the Limits in the Mediterranean’)

- 2) Swearing/insult/defamation:** Speech containing direct swearing, defamation, or insults about a community (e.g.: ‘Küstah Rum’a Gözdağı’, ‘Intimidation of the impudent Greek’; ‘Barbar Yunan’, ‘Barbarian Greek’;

'Hadsiz Yahudi', 'Impertinent Jew'; 'Danimarkalı itler iftar basıp, Kur'an yaktı', 'Danish dogs stomped on iftar and burned the Quran'; 'Barbar ve ahlaksız Fransızlar', 'Barbarian and immoral French').

3) Enmity/war discourse: Discourse that contains hostile, war-like expressions about a community (e.g. 'Rum vahşeti', 'Orthodox Greek savagery'; 'Haydut Rumlar Ateşle Oynuyor', 'Thuggish Orthodox Greeks are playing with fire'; 'Rumlar yine tahrik ediyor', 'Orthodox Greeks are inciting again'; 'Mültecilere Yunan zulmü', 'Greek cruelty to refugees'; 'Hans iyice kudurdu', 'The German has gone completely rabid').

4) Symbolization: Discourse in which an element of the identity itself is used and symbolized as an element of hatred and defamation (e.g. 'Bizi Eurovision'da Yahudi mi temsil edecek?', 'Will the Jews represent us in Eurovision?'; 'Ermeni gibi konuştular', 'They spoke like Armenians'; 'Yunan aynı Yunan', 'Greeks are still Greeks'; 'Yunan artığına Atatürk cevabı', 'Ataturk's response to Greek leftovers'; 'Rum ağzıyla rapor', 'Report with Orthodox Greek accents'; 'Cenk Tosun'a gavur eziyeti', 'Heathen persecution of Cenk Tosun'; 'Yunan askerinden mültecilere bir gavurluk daha', 'Another heathen act of Greek soldiers to refugees'; 'İçimizdeki İsraililer', 'The Israelis among us').

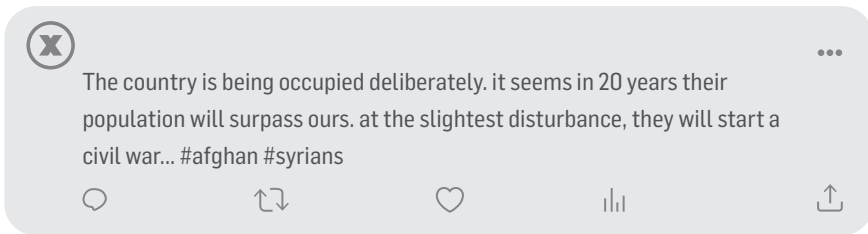
Within the scope of the Utilizing Digital Technology for Social Cohesion, Positive Messaging and Peace by Boosting Collaboration, Exchange and Solidarity project these categories were expanded and detailed as follows:

- Not sure
- There is no hate speech
- Exaggeration, generalization, attribution, distortion
- Swearing, insult, defamation, dehumanization
- Threat of enmity, war, attack, murder, or harm
- Symbolization

Below are brief descriptions of these categories used in the labeling process and selected examples for each. Three examples are given for each category and briefly examined. While the first example clearly fits the category in question, the second example is less clear but still contains hate speech from that category. The third example, which at first glance appears like it could belong to the category in question, does not actually contain hate speech or contain speech from other categories.

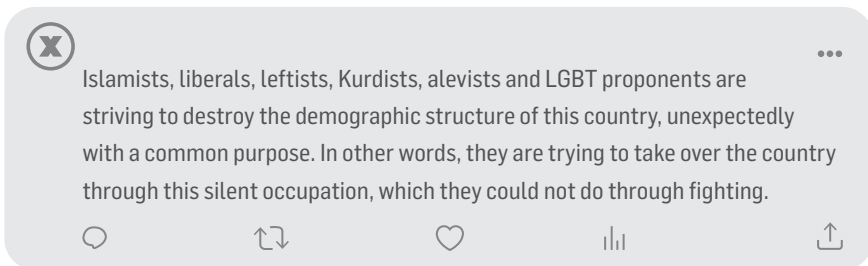
Exaggeration: Making an event, situation or action seem more significant than it is or reaching conclusions and inferences that are not supported by reality.

Clear example:



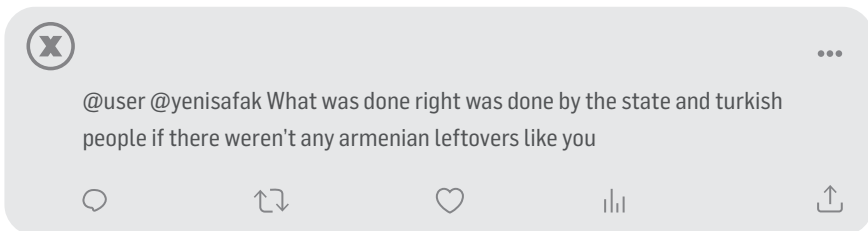
The prediction in this example can be labeled as hate speech by exaggeration because it implies a threat of civil war without presenting a plausible scenario.

Unclear example:



In the example above, it is emphasized that accepting refugees into the country in the name of human rights actually serves a secret agenda and it is stated that the aim is to disrupt the demographic structure of Turkey through a “silent occupation.” This example should be classified as hate speech by exaggeration.

Misleading example:

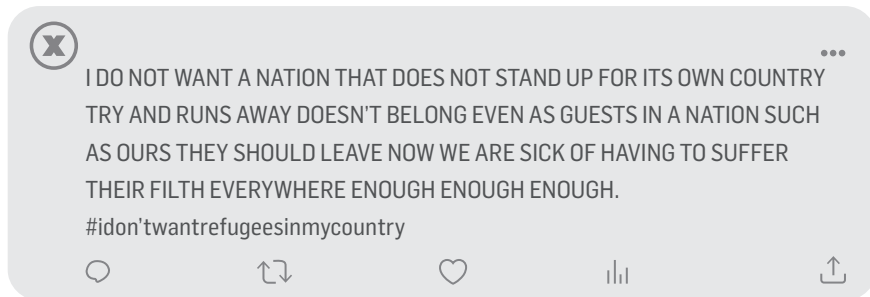


In the example above, while the “right” actions carried out within the country are attributed to a group, the responsibility for the events that are considered to contradict this situation is placed on the Armenians by saying “if there weren’t any people like you”. The lack of punctuation marks may make a difference in

the interpretation and categorization of the tweet. However, it does not contain direct exaggeration.

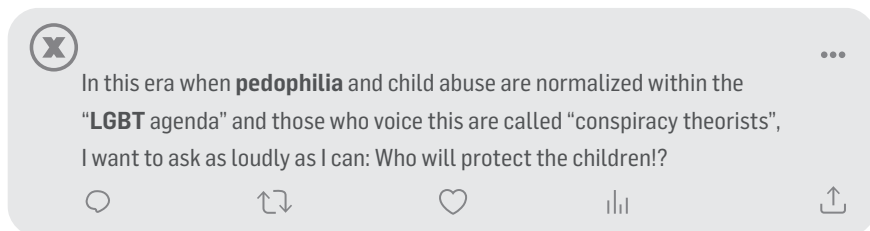
Distortion: Conveying an event, situation or action by deviating from the real data in a way that is incorrect, incomplete, or causes misunderstanding, such that the reader's perception and inferences are manipulated.

Clear example:



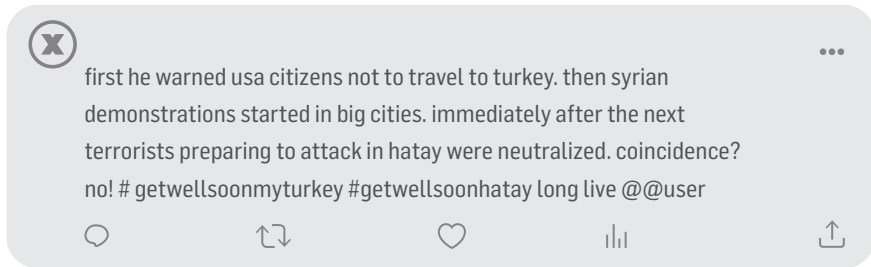
Portraying Syrians, who were forced to migrate due to the civil war, as running away from their country both distorts the truth and creates a perception that refugees are of no benefit to their own country and cannot be of any benefit to Turkey either. Together with the hashtag, this constitutes hate speech and should be categorized as “distortion.”

Unclear example:



The tweet above, although it is against pedophilia, actually indirectly alleges that individuals and groups with LGBTQ+ identities support rape and pedophilia. For this reason, it can be included in the category of hate speech by distortion.

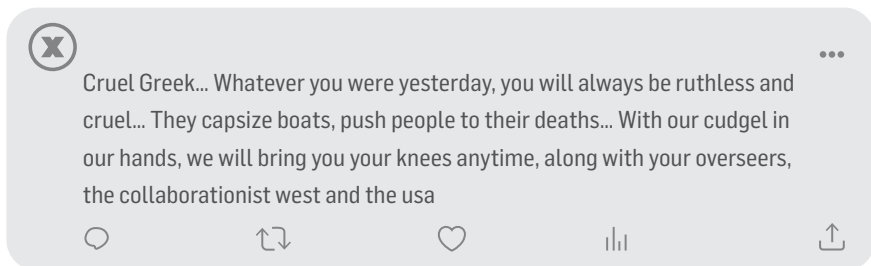
Misleading example:



In this example, the events are listed chronologically. As the tweet’s author claimed that there is a connection between the events, it is not a distortion of the truth but rather an individual’s inference. As such, this tweet should not be categorized as hate speech by distortion.

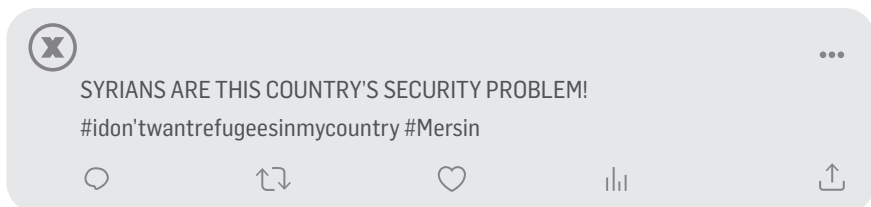
Attribution: Baselessly asserting that a group or identity is the cause of an event or situation.

Clear example:



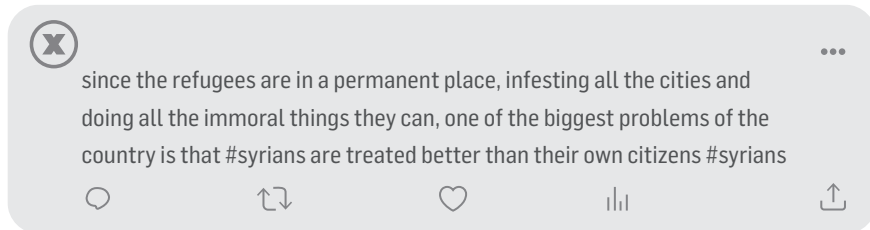
This tweet, which quotes a news report saying that the Greek Coast Guard illegally pushed refugee boats back to Turkish territorial waters, can be shown as an example of hate speech by attribution. It attributes the actions of the coast guard to Greeks collectively and accuses them of cruelty.

Unclear example:



In the tweet above, no clear incident is referred to in terms of why Syrian refugees are a security problem. However, it is seen that refugees are held responsible for security problems, and hatred is incited by the use of hashtags. Therefore, it can be labeled as hate speech by attribution.

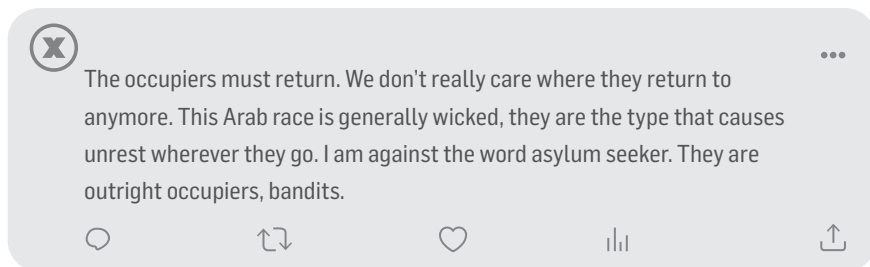
Misleading example:



This example contains hate speech against Syrian refugees by exaggeration and generalization. It should not be labeled as hate speech by attribution, as it does not place responsibility for a particular event or situation on refugees. The correct categories are exaggeration and generalization.

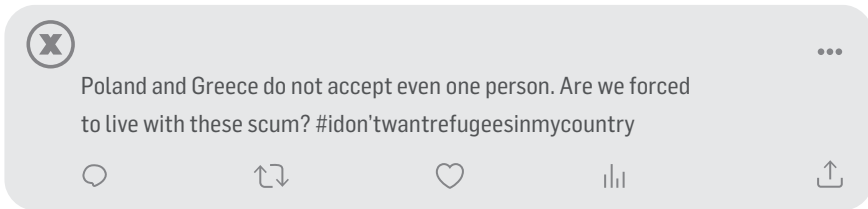
Generalization: Attributing an event, situation, characteristic, or action or its results to an entire identity.

Clear example:



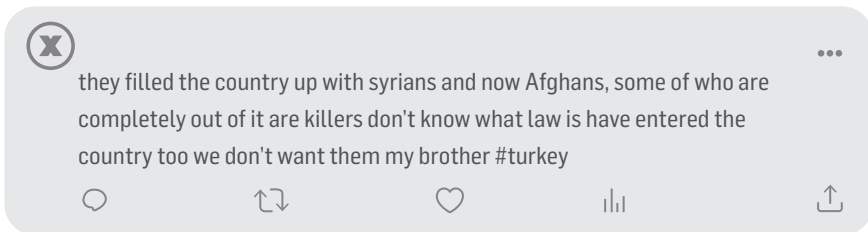
This example can be labeled as hate speech by generalization, as the accusations made target the entire Arab race.

Unclear example:



In this example, when the hashtag at the end of the tweet is taken into account, the word *scum* appears as an insult used against all refugees. In addition to the label of swearing/insult/defamation, it should also be labeled as hate speech by generalization.

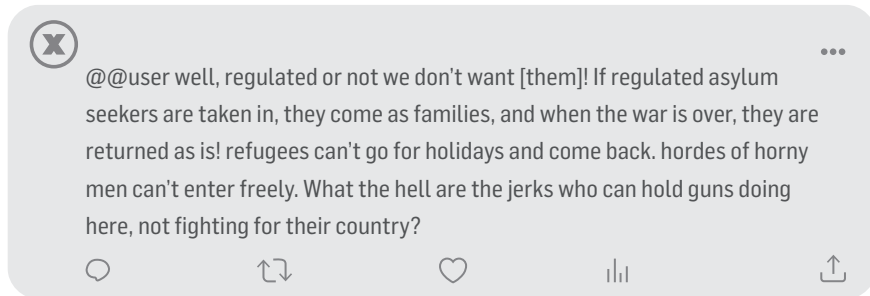
Misleading example:



In the example above, the invective against the Afghan identity group cannot be considered as an attack on the group as a whole due to the phrase “some of,” so this tweet should not be included in the category of hate speech by generalization.

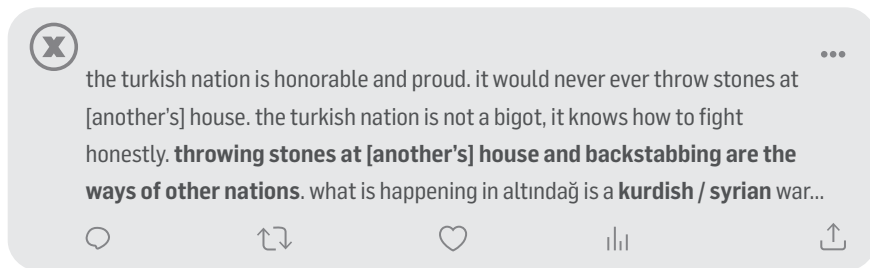
Insult: Attributing a physical act or fact to a race or community in a way that may offend their honor, dignity, and respect. Insult is an imputation of a characteristic. For example, legally, saying “you are a thief” to someone is considered an insult. In such cases, the truth of the event can be examined. If it is true, it is considered a mitigating circumstance.

Clear example:



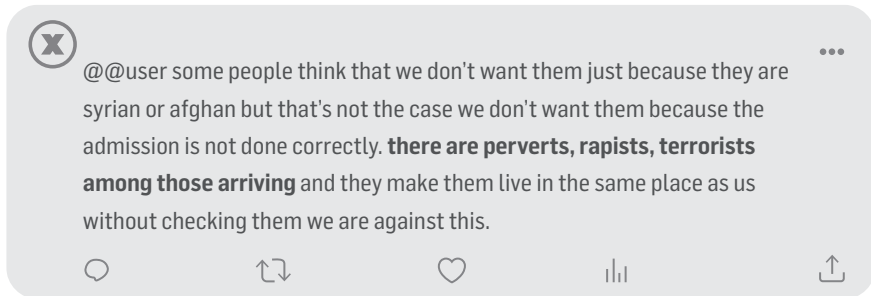
In the example, insulting words such as “hordes of horny men” and “jerks” were used. It is understood from the generality of the example that these negative adjectives were attributed to the mentioned asylum seekers collectively. All asylum seekers were targeted and insulted, and this constitutes hate speech.

Unclear example:



The example begins by listing positive descriptors of the “Turkish nation.” As the example continues, it is implied, albeit indirectly, that these positive descriptors of the “Turkish nation” do not apply to the other stated races or communities. The words “Kurd” and “Syrian” were clearly used and targeted, constituting hate speech against the stated races or communities.

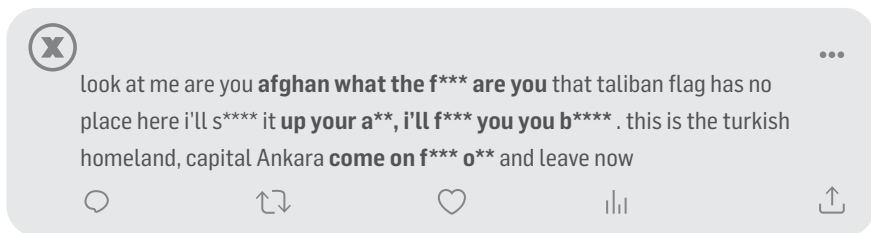
Misleading example:



In the example, insulting terms such as “pervert”, “rapist”, and “terrorist” were used. However, when we look at the example as a whole, it is understood that these terms are not directed at “Syrians” and “Afghans” collectively. It is explained that there may be people in the groups arriving who are inclined to commit crimes and therefore their entry into the country should be accepted in a controlled way. A criticism of state policy is being made. Therefore, it does not constitute hate speech.

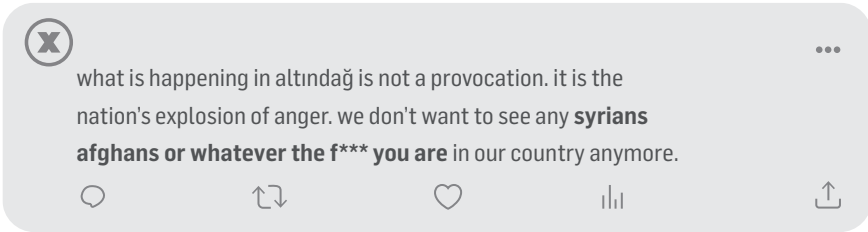
Swearing: Expressing or adopting a desire for a derogatory action towards a race or community in terms of cultural and traditional relations. It does not ascribe a characteristic, but refers to an action. Legally, the truth or reality of swearing is not questioned.

Clear example:



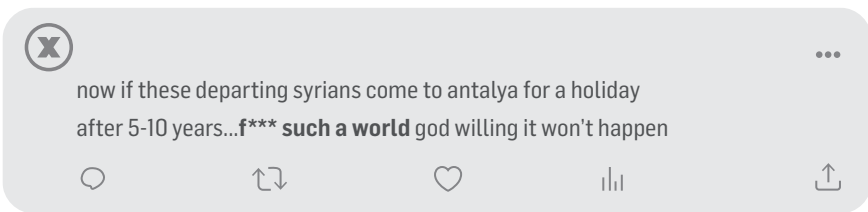
As seen in the censored parts of the example, swearing is used to target “Afghans” collectively.

Unclear example:



Although not as direct as in the previous example above, the censored section uses swear-laden language targeting a whole race.

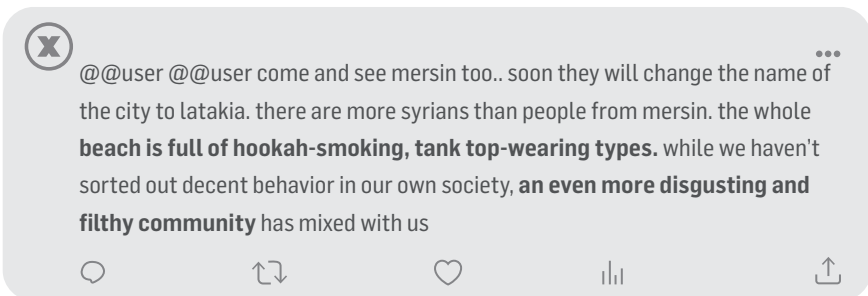
Misleading example:



In the example, although swear-laden language was used in the censored part, it is understood that this language does not target any race or community, but was used indiscriminately. It does not contain any hate speech.

Defamation: Creating the impression that a person, race, or group has inferior values compared to those which are generally accepted, or regarding these values with disdain.

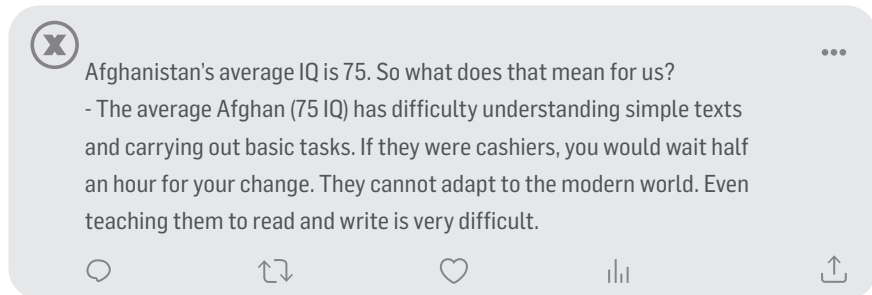
Clear example:



In the example, targeting "Syrians" collectively, it is claimed that they are "more disgusting and filthy." In this way, the community referred to was defamed by

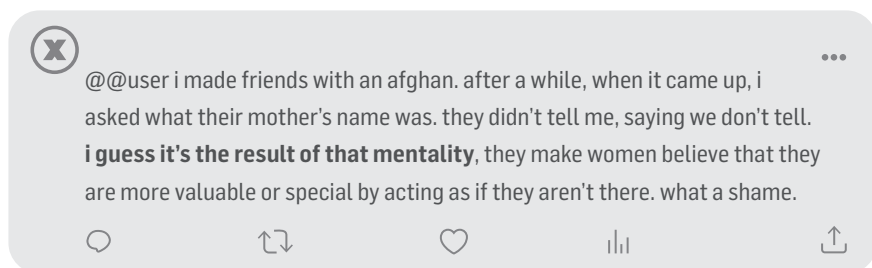
stating that it has inferior values. In addition, with the expression “hookah-smoking, tank top-wearing types” in the example, a certain outward appearance was generalized to the community as a whole. As far as we can understand from the general nature of the example, the community was collectively defamed through this appearance.

Unclear example:



The tweet above considers the results of a previously mentioned study. Even if the quote from the study in question is correct, and ignoring the reliability of the study and that the results given represent an average, it is stated that Afghans could not possess the cognitive development to adapt to the modern world. Looking at the tweet as a whole, while it appears to be a scientific commentary dealing with the results of a study, the inferences made constitute hate speech through defamation.

Misleading example:



In this example, the values of a certain mentality are seen as inferior. Although the tweet begins by referring to someone belonging to a certain race and then arrives at this supposedly inferior mentality, it is not understood whether “that mentality” is a value of the entire race. This mentality could be read as the views imposed by certain groups and organizations. As such, this example is a criticism against the views imposed by certain groups.

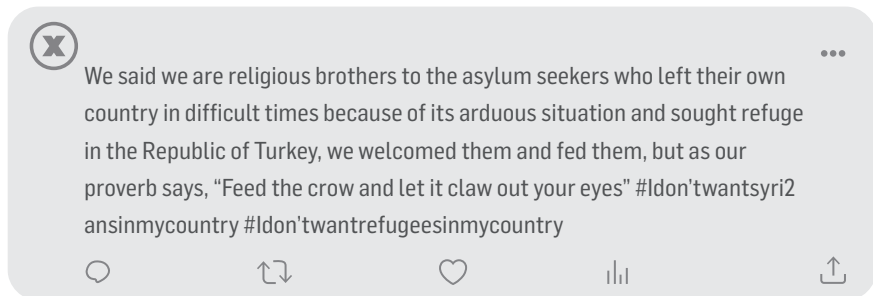
Dehumanization: Defaming a race or group by likening them to something non-human (such as an animal) or ascribing actions and descriptions specific to non-humans to them. All such uses are hate speech. Positive comparisons are excluded from this scope (for example, “as easy-going as a cat,” “faithful as a dog”). Frequently used words include verbs such as “to feed,” “to breed.”

Clear example:



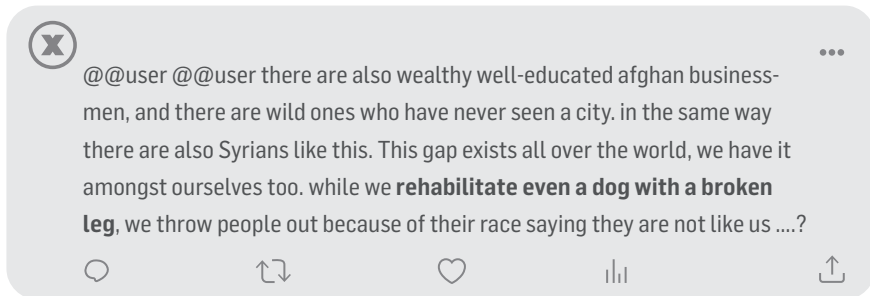
As clearly seen in the example, it is said that the “Syrians” had “become like dogs” and that the country had become a “dog shelter” because of their presence in the country. The “Syrians” are defamed by being likened to the non-human, in this case, dogs.

Unclear example:



Although the word “beslemek [to feed]” in the example is sometimes used for people, it is a usage more commonly associated with animals. By attributing an action specific to this kind of being to a certain race, this race has been defamed, with support from the hashtags, the tweet constitutes hate speech.

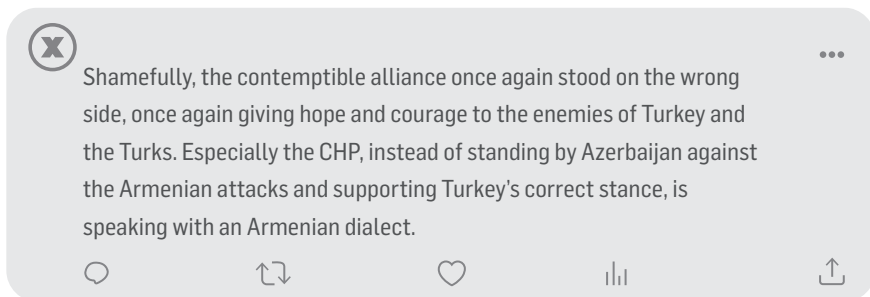
Misleading example:



In the example, the expression **“rehabilitate even a dog with a broken leg”** makes a comparison. Although the expression seems to establish a link with non-humans, no attribution is made to the race in question. Furthermore, it is not used with the intent to defame.

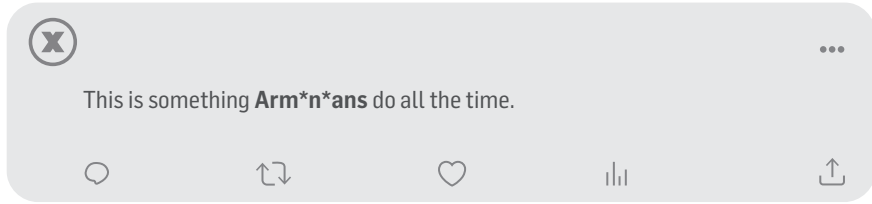
Symbolization: Speech in which an aspect of an identity itself is used and symbolized as an insult or an element of hatred and defamation. While some aspects of identity are targeted in the other categories, in this category the aspect of identity itself is used to create an element of insult, defamation, or hatred.

Clear example:

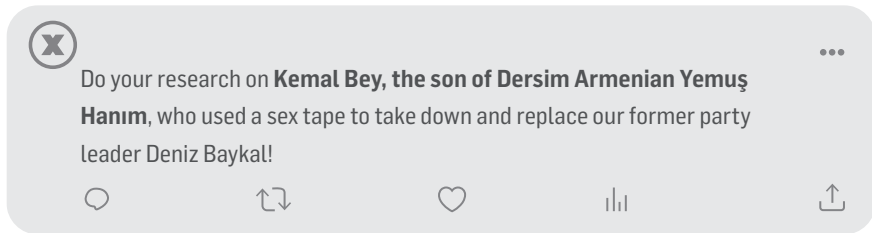


In this example, the expression “Armenian dialect” was used as if it had a very negative meaning. Although the example was not written to target Armenians, the expression “Armenian dialect” was used to denote an element of inferiority, while at the same time making a political criticism. In this way, it constitutes hate speech.

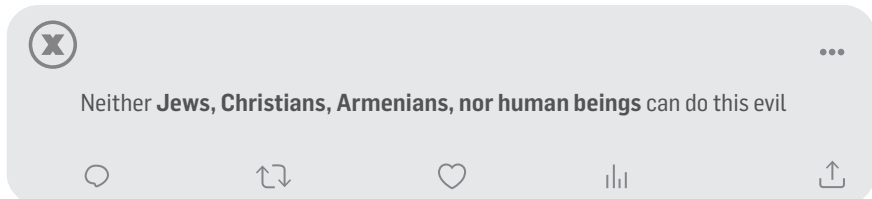
Unclear example:



In this example, the lack of context means it is not immediately clear what kind of action is associated with the group targeted under the term ‘Armenians’. As such, it is not possible for us to know whether the action in question is good or bad. However, the asterisks were intentionally placed and the word “Armenian” was censored as if it were profanity. The asterisk is an expression frequently used and associated with insults on social media. As such, the same symbol was used here as an aspect of identity, symbolizing the word “Armenian” and this therefore constitutes hate speech. This example was categorized as being unclear in order to emphasize the importance of paying attention to how identity is expressed in hate speech produced through symbolization. While hate speech in this example tweet is not constituted by being directed to an identity, it is constituted by symbolizing the identity itself.

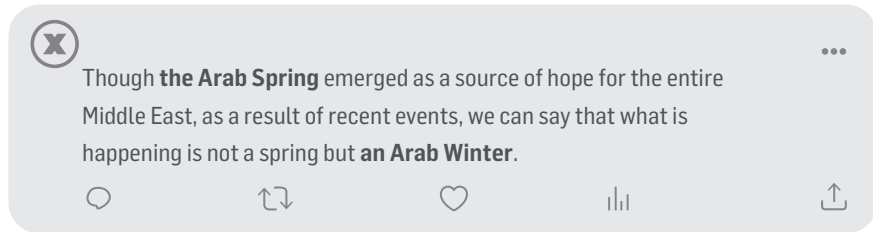


In this example, although it seems that the ethnic origin of the person mentioned is merely being reported, it is implied that being “Armenian” is an extremely negative thing with expressions such as “do your research”. In addition to the symbolization of ethnic identity, Kemal Kılıçdaroğlu, a political figure, is targeted by the claim that he is of a distinct ethnic origin, and this constitutes hate speech.



Looking at this example, it seems that a positive expression is being used towards the races and communities mentioned; however, more generally, it is implied that the races and communities mentioned have the most inferior of human values, expressing that 'even they wouldn't do it'. By being symbolized in this way, these races and communities are defamed.

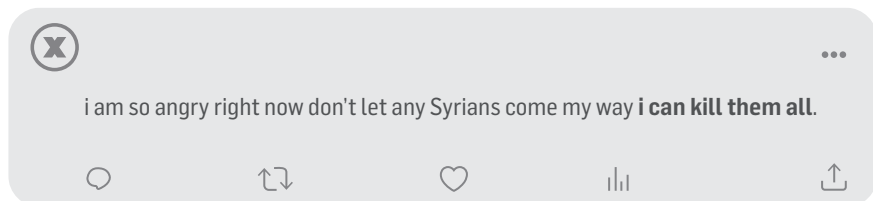
Misleading example:










The expressions Arab Spring and Arab Winter in the example are used to refer to social events, not to a nation. Therefore, they do not contain hate speech.



Enmity and threats of attack: A category of hate speech that includes statements that legitimize acts of physical and psychological aggression that may be carried out against a community, as well as those who commit such acts. It also includes statements that express a desire to carry out such acts or that such acts be carried out. Additionally, statements that express enmity towards a community or incite it should also be assessed under this category. Statements that contain insults, swearing, and dehumanization of the other categories are also acts of enmity. However, this category only covers tweet examples that legitimize or express a desire for attacks.






Clear examples:



 Don't let any such Syrian appear in front of me i swear **i'll break their bones** so badly i'm dying of anger right now. 



    






 Will the Russians, our enemy of a thousand years, now become our friends? 

In the first two examples above, there is a clear threat of doing harm. In the last example, enmity towards a community is clearly expressed. Therefore, these examples should be assessed in the “Enmity” category.



Unclear example:






 We want a purge. #Wedon'twantSyrians 

The term “purge” in the tweet above might not be understood by someone unfamiliar with the idea. As such, the hate speech constituted may go undetected. However, “The Purge” is the name of a movie in which it is legal for people to kill or torture each other for one night. When the tweet is assessed in the light of the hashtag at its end, it is understood that the person who posted the tweet is calling for various attacks against Syrians to be considered legitimate, as in the film.

Misleading example:



 the war is over, the joy of being a guest is gone. they should return to their country, our country is not a holiday resort. hurry up back to your country. 






As there is no clear expression of a desire to cause any harm in the example above, it does not fall into this category. However, the phrase ‘they should return to their own country’ should be considered hate speech because it points to a potential violation of the legal status of Syrians under temporary protection of the Turkish state.



War discourse: A category of hate speech which includes expressions that call for war against a community, are used to incite war, attempt to legitimize military interventions that will lead to war, or attempt to legitimize an existing war.

Clear examples:






 

We can't take anymore. What are the soldiers waiting for? Let's launch two missiles and they'll come to their senses.



 

If they fired on our ship, let's sink their ships too. Eye for an eye, tooth for a tooth.






    

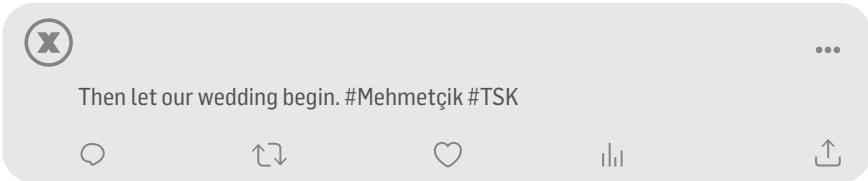
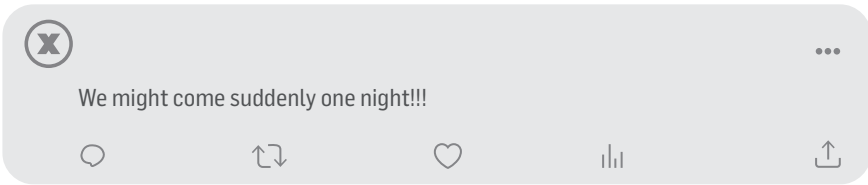
In the examples above, there is a clear call for war as can be seen from such expressions as “let’s launch missiles” and “let’s sink their ships”. Additionally, attempts are made to legitimize the desire for war with expressions such as “we can’t take anymore” and “eye for an eye, tooth for a tooth”.

Unclear examples:

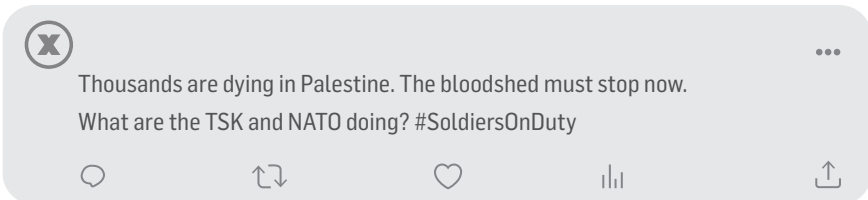
82 London 83 New York!!!



In the examples above, there is no explicit desire for war. However, they include threats of war with the phrases “82 London 83 New York” and “we may come suddenly”. In addition, the phrase “let our wedding begin” in the last tweet, when assessed together with the hashtags at the end of the tweet, is understood to be a call for war.

Misleading example:



The hashtag “#SoldiersOnDuty” used in the tweet can be perceived as a call to war. However, the phrase “the bloodshed must stop” indicates that the author of the tweet has an attitude favoring the end of conflict. Therefore, this tweet should not be considered as discourse favoring war.

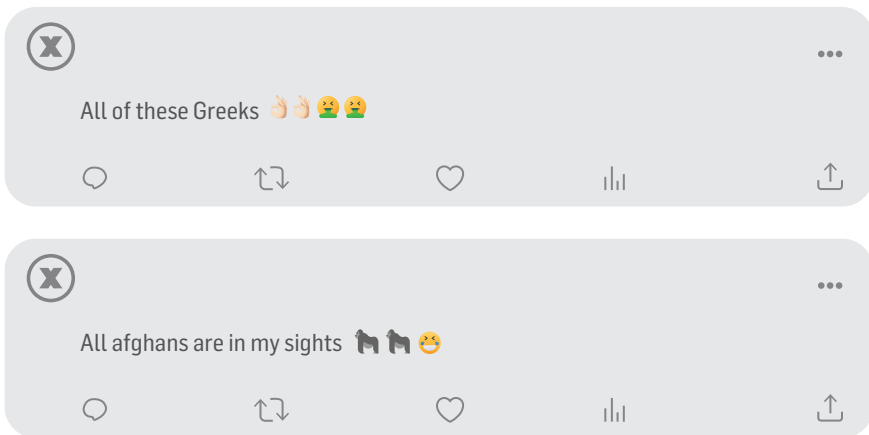
2. 4. Assessing challenging examples

2.4.1. Tweets containing hashtags and emojis

As various hashtags based on identified topics were used in the tweet gathering phase, the tweets that come up also contain hashtags. Some **hashtags** that artificial intelligence models cannot directly understand are also present in tweets. While in some examples the hashtags used align with the ideas in the text of the tweet, in others they are independent and the hashtag is used only to move the

tweet to the top of the feed. Tweets should be assessed together with hashtags. In cases where the tweet content itself does not contain hate speech but the hashtag does so, the expression should be considered as hate speech.

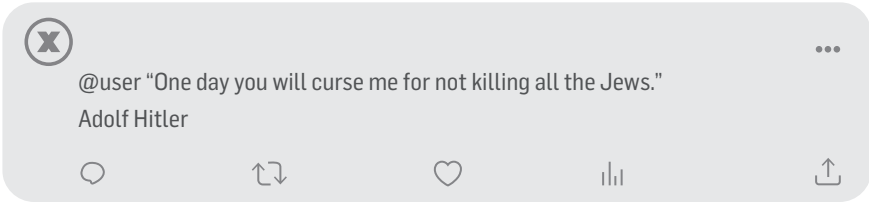
As seen in the tweets below, various emojis are present in some instances. Tweets containing emojis are also included in the labeling. The emojis used may be unrelated to the text or they may support its ideas. For this reason, tweets are assessed in relation to the meaning added by emojis.



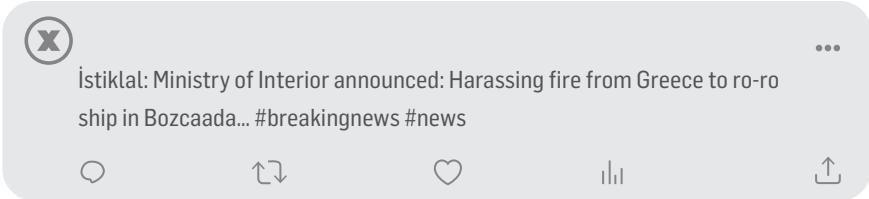
In the tweet examples above, hate speech is formed by using emojis; when assessed without paying any heed to the emojis, the hate speech cannot be detected. As in the first example above, various hand sign emojis, which are considered socially insulting, are widely used to create hate speech. In addition, as seen in the second example, various animal emojis can be used in tweets for the purpose of dehumanization. In addition, various vegetable and fruit emojis are used in a way that evokes sexual organs and can be insulting as such. Therefore, when labeling, the content of the tweet and the emojis should be assessed as a whole.

2.4.2. Tweets that include/quote the speech of others

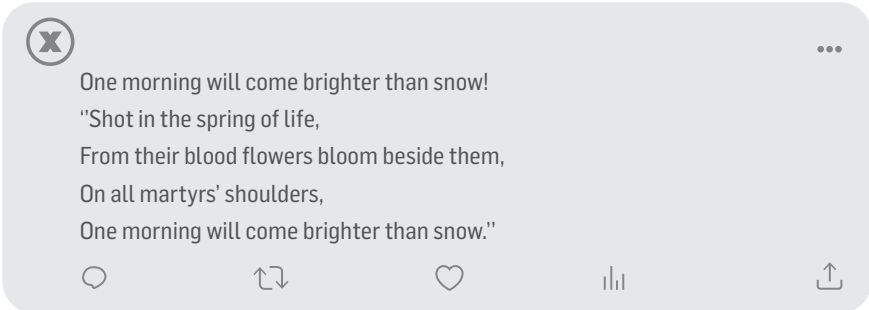
Tweets that quote/convey another’s hate speech-containing statement recirculate that statement and cause it to spread. If such tweets are shared without a critical comment, as in the example below, that tweet should be marked as **containing hate speech**.



In the example above, a quote legitimizing the Holocaust is included. Looking at the entire tweet, it can be seen that the words in the quote are not criticized. The person who posted the tweet has re-circulated the quote and associated it with an unrelated event. Since this quote, which includes enmity and threats of attack against Jews, has been re-circulated without criticism, it should be labeled as hate speech. In contrast, the tweet in the example below, which also includes a quote, does not directly target an identity group and there is no hate speech in the quoted content:



As in this example, while labeling the content of tweets which do not contain hate speech, and convey their intent by means of quotation and are also considered to be newsworthy, the neutral or irrelevant option should be selected in the "Overall attitude and stance" section. Another example is shown below:



The irrelevant option should be selected because the tweet content is not directly related to the topic.

2.4.3. Sarcastic content

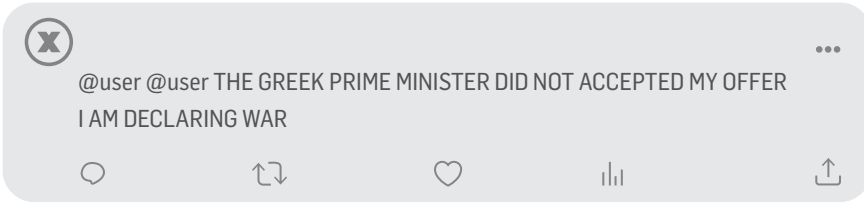
When example tweets considered to contain sarcasm were examined, it became apparent that, in fact, not all of them did so. Accordingly, when deciding whether a tweet contains sarcastic content or not, attention should be paid to the definition of sarcasm and what kind of expressions can be considered to fall within this definition.

Sarcastic expressions generally aim to convey the opposite of what is said. In written discourse an important clue that can help us in determining is emojis. Emojis allow the writer of any tweet to clearly express intention. The use of an emoji that contradicts written expression may indicate that the expression in question is actually sarcastic. For example, the expression “Beş milyon mülteci kardeşimizle mutlu bir yıl dilerim” (“I wish us a happy new year with our five million refugee brothers”) is a sentence that appears to wish the refugees well. However, the use of an emoji expressing anger after this phrase hints that the person who wrote it feels disturbed by the presence of refugees. Therefore, it can be said that this expression is sarcastic. Similarly, the use of happy or funny emojis after a sentence describing a sad event shows that the person using this expression does not care about the event or minimizes it. Therefore, the use of emojis is a guide to detecting sarcastic expressions.

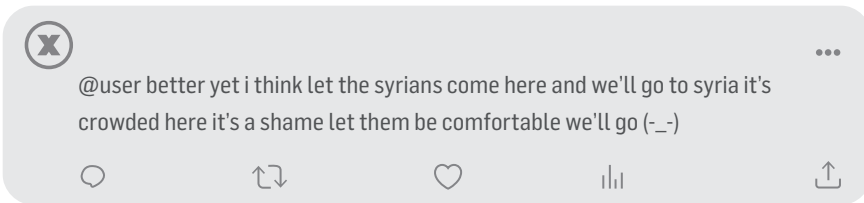
In everyday conversations, emphasis or intonation can be used to imply that an expression is sarcastic, but naturally it is not possible to detect these in written texts. However, different ways of spelling and other signs used can be helpful in this regard. For example, in “YENİ YILINIZI BEŞ MİLYON MÜLTECİ KARDEŞİMİZ İLE KUTLARIM!” (“I CELEBRATE YOUR NEW YEAR WITH MY FIVE MILLION REFUGEE BROTHERS!”), the fact that all the words are written in capital letters shows that the expression is overemphasized. In this way, it is understood that the person who wrote it is disturbed by both the number of refugees in the country and the expression “brother”. In addition to the use of capital letters, some words can be placed in quotation marks or parentheses, and some words can be followed by exclamation marks. These choices can indicate that, by emphasizing these words, the writer wishes to express the exact opposite of the phrase used.

However, we can also determine whether a sentence is sarcastic or not without recourse to the clues mentioned above. Continuing with the example above: “I wish us a happy new year with our five million refugee brothers”. As the statement contrasts with the general public opinion, someone with a grasp of this could very well perceive it to be sarcastic. However, it should be noted that there will always be a minority group that holds opinions contrary to those of the majority.

Examples:



In the example tweet above, the first person use of the verb ‘to declare war’ indicates that the expression is sarcastic. Even if the context of the tweet is unknown, it is seen that the tweeter is making fun of the act of ‘declaring war’ and expressing this in a sarcastic way. Due to the expression of war discourse in the tweet, it should be assessed as ‘anti-Greek’ and be tagged as ‘enmity/war discourse’ in the categories.



In this example, the situation expressed by the phrase “it’s crowded here” and the contrast evoked between them and us is unexpected and unnatural, therefore it can be said to contain sarcastic content. At the same time, the phrase “it’s a shame let them be comfortable we’ll go” at the end of the tweet indicates that the user is disturbed by the fact that Syrians have access to rights in their daily lives in Turkey. Therefore, this tweet should be assessed as ‘anti-refugee’ and an example of ‘discriminatory discourse’.

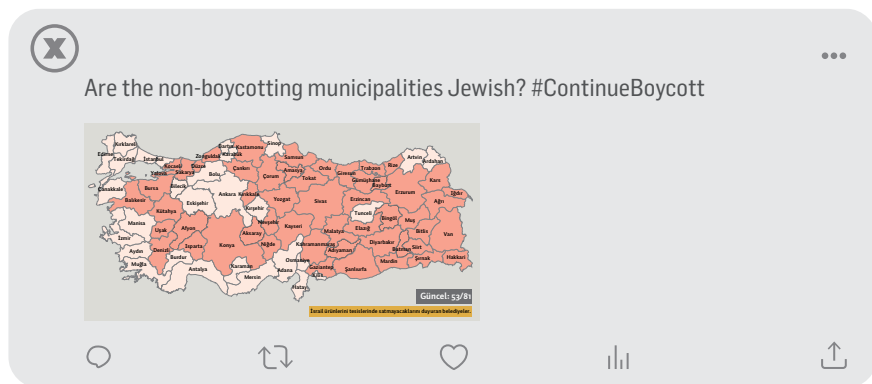
Another case which we can include in the category of sarcastic expression is the use of words typically used in hate speech contexts that are then adopted by minority groups themselves, be it on banners, slogans, or in intra-group communications. For example, the word “dönme”, roughly meaning transsexual, which is used to discriminate against and insult LGBTI+ individuals, was embraced by this group and began to be used in a way that was contrary to its original purpose, such as the slogan “Velev Ki Dönmeyiz [so what if we are dönme]” which was used in Pride Month marches and events. Since it is normal to encounter situations of this nature on social media, namely situations in which minority groups “reclaim” words that are usually used in the context of hate speech, it is important to include

them in this methodology guide. Similarly, LGBTI+ individuals may interact with one another on social media using words and expressions that would normally be considered hate speech, in order to joke, be ironic or to simply to convey meaning. When encountering tweets containing such expressions during labeling, if it is understood from the context of the tweet that the expressions contain insults or defamation, the tweets should be assessed as hate speech. However, in cases where the content is not understood from the context or the expressions are used for “reclaiming” purposes, the expressions themselves should be selected as triggering words or swear words/insults for the correct training of the tool and should not be considered as hate speech.

In any case, in order to develop a broad and inclusive understanding of hate speech, it is important to be aware that certain expressions on social media do not always directly correspond to their standard meaning.

2.4.4. Covert hate speech

Expressions used in tweets may not contain explicit hate speech. Expressions that appear neutral at first glance, but which may constitute hate speech when combined with information regarding the reader’s cultural context, are categorized as covert hate speech. This kind of speech is as important and damaging as explicit hate speech. Therefore, it is important to label such tweets as hate speech. An example of covert hate speech from X is given below:



In the example given above, the tweet is supported by a photo, and the photo in question indicates, by means of color, those municipalities that boycott Israeli products and those that do not. At first glance, the sentence above may not seem to contain hate speech, as it does not use any overt insults or established

offensive/discriminatory expressions. However, the contextual information and the ethnic word used indicate hate speech. In addition to equating the “Jewish” identity with the actions of the Israeli state, that identity is symbolized and used as an element of defamation. Therefore, the tweet covertly contains hate speech and should be flagged accordingly.⁶

2. 5. Additional labeling headings

In the last section, other labeling headings used in the interface in the project are presented together. These headings, which were formed after a long planning process, were developed to make labeling more detailed and to train the tool more effectively. It is intended that these headings will also be a starting point for future work.

2.5.1. Language of tweets

There are “Turkish” and “Not Turkish” options in this section. For tweets in a language other than Turkish, the “Not Turkish” option should be selected and the tweet labeling should be completed **without labeling other sections**. This should be done even if the content of tweets in other languages is understood.

In cases where a person uses words from different languages in otherwise Turkish-language tweets, the option “Not Turkish” should be selected. In this context, tweets containing frequently used internet English-language abbreviations such as “LOL (laughing out loud)” or “OMG (oh my god)” should also be selected as “Not Turkish”.

In addition, tweets with hashtags written in languages other than Turkish should also be labeled as “Not Turkish” even if the tweet is in Turkish. This was preferred for technical reasons so that the digital tool being developed could interpret the data more accurately. If the tweet text is written in Turkish but the visual content contains non-Turkish text, the ‘Not Turkish’ option must be selected.

2.5.2. Identifying the hate speech span

It is also expected that relevant words will be marked in messages containing hate speech, both for the guidance of those labeling them and for use in different artificial intelligence methods and hate speech detection algorithms. In this regard, the following should be taken into consideration:

6 Content with images is not labeled within the scope of this project. The tweet above is given as an example of covert hate speech.

- When assessing the speech in a tweet, words or phrases containing hate speech or discriminatory discourse that is triggering for the reader should be selected. For example, in a long tweet, care should be taken in marking a maximum of three prominent expressions considered to be triggering. While some expressions considered to be triggering may fall into the category of enmity/war discourse, others may fall into the category of swearing/insulting. In tweets containing hate speech that belong to both categories, all relevant expressions should be selected by clicking in order on the category headings seen in the image.

Tweet

@HDPgenelmerkezi a real Kurd would be a Muslim they would not let an Armenian traitor into their party. Armenians' purpose is to cause a war between Kurdish and Turkish and to found Western Armenia in Kurdish lands. They are all secret agents of Armenians, Jews. Do not send your kids to PKK, let them send theirs since they are Kurdish....

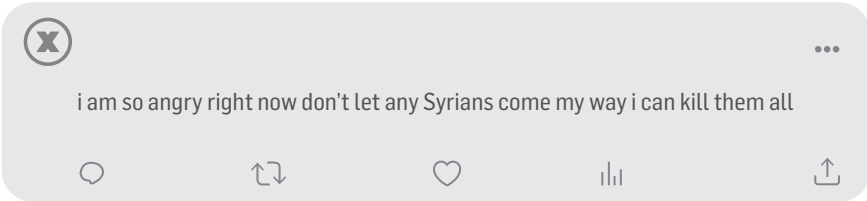
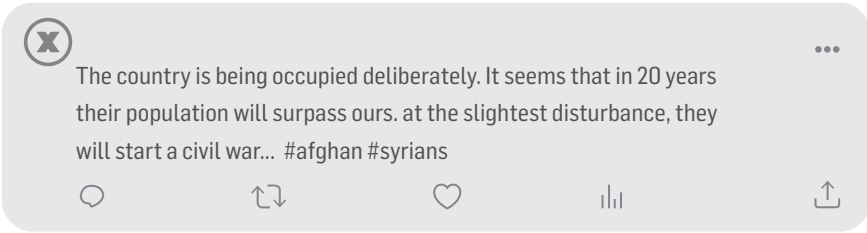
After clicking one of the following boxes, you can select words or phrases that cause hate speech. (Maximum 3 words or phrases can be selected)

Triggering Word 1 Swearing/Insult 2 Enmity Discourse 3

2.5.3. Hate speech strength

When labeling tweets, the degree of hate speech should be selected. Marking should be done on a scale of **1–10**, with 1 being the lowest and 10 being the highest. Tweets that do not contain hate speech should be marked as 0. When labeling, the degree and category markings should be considered independently of each other in order not to establish a relationship between the degree of hate speech and categories based on violence. When rating, it is the content of the language used that should be taken into consideration, rather than making a choice based on the stated category. The degree should be directly proportional to the intensity and frequency of words containing hate speech.

Examples:

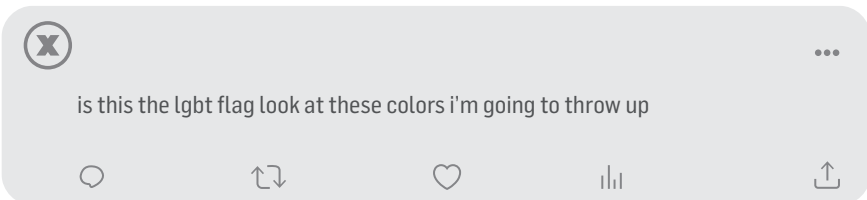


Both examples contain hate speech. However, since their content is not at the same level of intensity, different ratings should be given. The explicit physical threat used in the second tweet has a higher intensity of hate speech when compared to the first. Therefore, the ratings of the two tweets will be different.

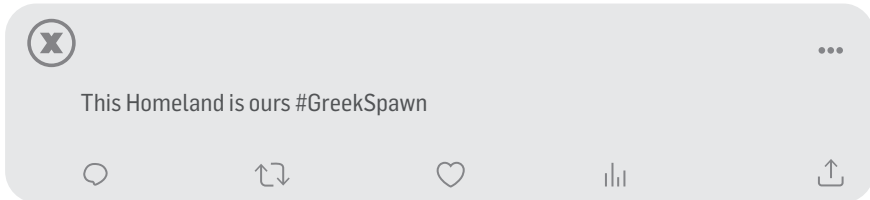
Finally, if “Not Sure” was checked in the option for hate speech/not hate speech, the option “Not Sure” should be checked here as well.

2.5.4. Offensive language

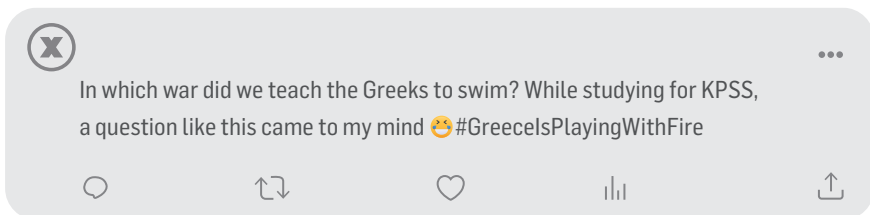
At this stage, we need to assess the tweets we are examining in terms of the threat of and/or desire for an attack. If there is no offensive expression in the content of the tweet, we should select the “None” option. When deciding whether the threat of or desire for attack in the content is “Low” or “High”, we can consider the magnitude and degree of the impact of the threat in question. Therefore, it is important to assess this not in relation to the hate speech category and severity sections, but to assess it specifically for this category, for the sake of consistency of detection.



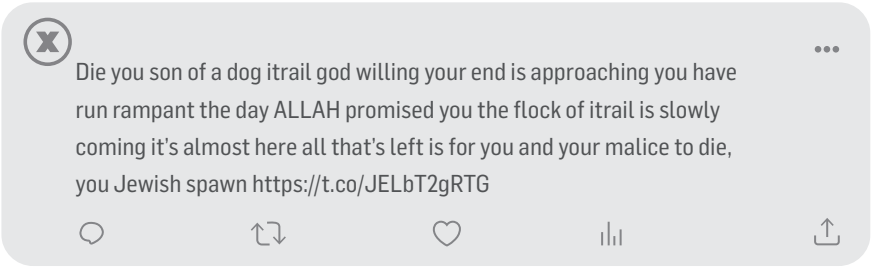
In this example, we can establish that LGBTI+ people are targeted by the negative expressions employed. However, when we consider whether it contains any actual aggression towards LGBTI+ people, we can say that there is no threat of or desire for an attack on the target group, and therefore there is no offensive language in the tweet. When labeling, we should select the “None” option for this and similar examples.



In this example, the Greek identity is targeted by means of symbolization. We cannot see to whom the tweet was written as a response, but we can understand that the phrase “This Homeland is ours” is directed at a certain identity group, and this is made clear in the phrase in the hashtag. In this phrase, the Greeks are mentioned for the sole purpose of defamation. When we assess the offensive language in the tweet, we should label it with ‘None’ since there is no threat of or wish for attack.



In this example, we need to assess the rest of the tweet in relation to the hashtag used. When we consider the hashtag and the main text together, we can consider that historical events are being referred to and hate speech is being supported. As the expression “playing with fire” used in this hashtag expresses a threat to the group being targeted, we can say that this tweet contains offensive language. However, we should select the “Low” option for the offensive language in this tweet. Although we detect a threat of an attack, as far as the content is concerned, we cannot say that this threat is significant in terms of it actually being planned or having a potentially damaging effect. Therefore, we should rule out the “High” option.



In this example, we can say that more than one group is being targeted, Jews are targeted by means of symbolization, while Israelis are the targets of swearing and insults. Because of the first word in this tweet and the subsequent invocation of violence, we should also mark the offensive language in this tweet as “High.”

2 DEVELOPING THE AI MODEL

To address the challenges in combating hate speech, we developed an AI-powered tool to detect hate speech in Turkish tweets by using the guidelines outlined in the previous section. This tool not only identifies hate speech but also evaluates its intensity, categorizes it, and pinpoints its location within the text. Our tool is further enhanced with models designed to identify hate speech in Arabic tweets and also in Turkish print media⁷. Another important feature of the tool is its ability to monitor the X platform periodically in real time, collecting relevant tweets and automatically labeling them using our AI models.

7 Hrant Dink Foundation's 10-year data from the Media Watch on Hate Speech Project, focused on print media, was used to create a sample data pool.

1. DATA COLLECTION & ANNOTATION

The hate speech detection tool was developed using a machine learning approach. The development of machine learning tools requires labeled (annotated) data for both training and testing the models. In our case, this involves a collection of tweets containing different levels and categories of hate speech, as well as tweets that contain no hate speech. Tweets are defined as short texts with limited contextual information. They often contain non-standard language such as abbreviations, as well as typos and grammatical errors.

In order to respond to the needs of the CSOs and researchers in the MENA region, data collection and processes of tweets in the Arabic language was determined and developed through consultations with native speakers of Arabic and our collaborators from the region. These consultations involved experts from our network platform on hate speech, Arabic annotators who participated in project activities, and kick-off participants with whom we initiated collaboration at the beginning of the project.

The following sections provide detailed information on the data collection and annotation processes.

1.1. Data collection

Tweets were downloaded using X's academic API and scraping methods from the specific time periods in which they were written and this was based on certain keywords and hashtags. These hashtags and keywords were selected by regularly monitoring current events and also by including groups that are frequently exposed to hate speech in Turkey. Based on these, Turkish content that targeted nine different target groups for hate speech (Alevi, Arab, Armenian, Greek, Jewish, Kurdish, LGBTI+, refugees (in [Arabic], and refugees [in Turkish])) was retrieved.

It should be mentioned that the hate speech monitoring activities of the Hrant Dink Foundation, which continued during the duration of this project, do not have limitations regarding target groups. Accordingly, more than 100 groups and identities have been found to be the target of hate speech in print media over the years. However, for the purposes of this project, we have narrowed our data pool to include tweets concerning the nine groups mentioned above, all of whom have been frequently targeted. This has allowed us to collect data more efficiently by using certain hashtags and keywords and by facilitating the annotation processes.

Nevertheless, it was our aim that our algorithm could be generalized in a manner that allowed it to also detect hate speech targeting other groups.

Based on the criteria laid down above, a total of 16,254 tweet labels were collected. The total number of tweet labels and the tweets per hate speech topic is shown in Table 1. All the topics except for “Refugees (Arabic)” consist of Turkish tweets, while the “Refugees (Arabic)” topic contains Arabic tweets exclusively. Some tweets contain hate speech relevant to multiple targets, leading to the same tweet being included under different target groups.

Table 1. Total number of labels and downloaded tweets per hate speech topic

| Topic Name | Number of Tweets Annotated by at Least 3 Annotators | Number of Retrieved Tweets |
|--------------------|---|----------------------------|
| Jewish | 3720 | 8200 |
| Greek | 2418 | 19500 |
| Refugees (Turkish) | 2289 | 4350 |
| Refugees (Arabic) | 2999 | 5750 |
| Alevi | 1000 | 5650 |
| Armenian | 979 | 3300 |
| Arab | 1005 | 7550 |
| Kurdish | 947 | 18500 |
| LGBTI+ | 897 | 1350 |
| Total | 16254 | 74150 |

1.2. Annotation

The manual annotation of tweets, particularly hate speech annotation, is challenging due to the subjectivity in the task, the nature of the tweets themselves, and dependence on context. While annotators generally agree on labeling discourse that contains obscene language such as swear words or threats towards a target group as hate speech, they often disagree on how to classify more subtle discriminatory speech (e.g. “Refugees should not get government assistance”). To manage these discrepancies, researchers often address the issue by discarding samples that contain annotator disagreements, resulting in data loss and overly optimistic model results. Another approach to dealing with annotator disagreements and to

improve data quality is to have a second annotation phase where annotators are expected to reach a consensus.

To obtain a high-quality dataset, our team of computer scientists, linguists, social scientists, and civil society experts worked closely together to develop a set of guidelines for annotating tweets for hate speech (see the Guidelines section). We adopted an iterative approach in order to develop the annotator guidelines and also refined the guidelines to resolve ambiguities and conflicts in the annotation process. For instance, we incorporated more examples for different hate speech categories which aimed to eliminate confusion and clarify the guidelines about what to do in ambiguous situations such as when a tweet contains hate speech towards multiple groups or when it contains covert hate speech (i.e. someone detects hate only if she/he knows the context). In order to ensure that different perspectives on hate speech and its target groups are represented during the data annotation processes, another recommendation was made to work with annotators from diverse backgrounds.

The resulting guidelines are quite comprehensive and include labels for hate speech categories, target groups, and the perceived degree of hate speech viewed independently of its category in order to evaluate the effect it has on the AI model.

For hate speech category labeling, we conducted detailed discussions on inclusivity and coverage of the categories. As a result, in addition to the “no hate speech” category, we identified four specific types of hate speech: i) Symbolization, ii) Exaggeration/Generalization/Attribution/Distortion, iii) Swearing/Insult/Defamation/Dehumanization, and iv) Threat of Enmity/War/Attack/Murder/Harm. For the degree of hate speech, we defined different levels of hate speech, ranging from 0 to 10, where 0 represents no hate speech and the severity increases at higher levels.

Table 2 shows the explanations of the hate speech categories. These categories were built upon the existing ones used by the Hrant Dink Foundation for monitoring hate speech in Turkish print media⁸. During the course of the project, the researchers from all three institutions provided their insights in order to refine the approach and ensure that it remains current and suitable for social media settings.

8 see Hate Speech and Discriminatory Discourse in Media 2019 Report: <https://hrantdink.org/en/asulis/publications>

A team of annotators was formed, consisting primarily of university students from diverse fields such as media studies and sociology. The selection was based on their expressed interest in the topic and a review of their resumes. Before the annotation process started, the annotators received thorough training in the methodology of the project by the HDF project team. This training included an introduction to our annotation guidelines and a review of several tweet examples for each category and level of hate speech. The tweets were divided into batches, each containing 50 tweets, for annotation. They were subsequently uploaded to the labeling server, where annotators used the Label Studio interface for the labeling process. To ensure the quality of the labels, each tweet was labeled by three different annotators, meaning that each batch was labeled by three individuals across three separate ports. In cases where the number of labels was insufficient, such as for the “Refugees (Arabic)” topic, tweets labeled by only one or two annotators were also used in model training. Multiple selections were allowed for the “Target Group” and “Hate Speech Category” labels, as multiple groups can be targeted by a single tweet. In such cases, we split the annotator vote among the selected groups or categories.

Table 2. Explanations for hate speech categories

| Hate Speech Category | Explanation |
|--|--|
| Symbolization | Discourses in which an element of identity itself is used as an element of insult, hatred, or humiliation and the identity is symbolized in such manners. |
| Exaggeration/ Generalization/ Attribution/ Distortion | Discourses that draw larger conclusions and inferences from an event, situation, or action, manipulate real data by distorting it, or attribute isolated incidents to the entirety of an identity. |
| Swearing/Insult/ Defamation/ Dehumanization | Discourses that include direct insults, slurs, or demeaning remarks towards a community, or describe them with actions or attributes typically associated with non-human entities. |
| Threat of Enmity/ War/Attack/ Murder/Harm | Discourses that contain hostile statements, invoke war-like language, or express a desire to harm the specific identity in question. |

To assess the consistency of the labels of the annotators for tweets labeled by more than one annotator, Krippendorff’s alpha coefficient method was used. This coefficient ranges between -1 and 1, where 1 represents perfect agreement, -1 represents complete disagreement, and 0 indicates random correlation among annotators’ selections. Values between 0.33 and 0.67 are considered moderately

reliable (moderate agreement), while those above 0.67 are regarded as highly reliable (high agreement). Krippendorff’s alpha coefficient values per hate speech topic are shown in Table 3. For instance, the agreement about whether offensive language was used when the target group was Jewish (top row), the coefficient was 0.329, indicating moderate agreement. However hate speech strength and category coefficients were smaller. This is due to challenges caused by the subjective nature of hate speech and the volunteer nature of the annotation task. Factors such as high annotator turnover, a large number of annotators (resulting in fewer tweets being annotated by each), and varying interpretations of hate speech contribute to lower agreement rates. This is explored more in detail in the “Limitations” section of this report.

Table 3. Krippendorff’s alpha coefficient values per hate speech topic

| Topic Name | Overall Attitude and Stance | Hate Speech Strength | Offensive Language | Target Group | Hate Speech Category |
|--------------------|------------------------------------|-----------------------------|---------------------------|---------------------|-----------------------------|
| Jewish | 0.31 | 0.176 | 0.326 | 0.268 | 0.197 |
| Greek | 0.294 | 0.176 | 0.341 | 0.283 | 0.296 |
| Refugees (Turkish) | 0.502 | 0.064 | 0.296 | 0.416 | 0.281 |
| Refugees (Arabic) | 0.522 | 0.213 | 0.186 | 0.158 | 0.198 |
| Alevi | 0.013 | 0.055 | 0.386 | 0.237 | 0.075 |
| Armenian | 0.284 | 0.068 | 0.343 | 0.179 | 0.143 |
| Arab | 0.306 | 0.158 | 0.327 | 0.377 | 0.211 |
| Kurdish | 0.214 | 0.147 | 0.368 | 0.329 | 0.348 |
| LGBTI+ | 0.191 | 0.158 | 0.229 | 0.252 | 0.346 |

It should be noted that despite all the effort devoted to collecting and annotating this data, the annotations can be further refined in the future: i) as stated above, while the majority of the tweets were annotated by three annotators, the rest were annotated by one or two persons, so these could be rounded up to three annotators, and ii) when multiple annotators are present, annotator disagreement is not explicitly handled, but is instead left to the discretion of those in question. Nonetheless, we do provide different strategies aimed at combining multiple annotations (i.e. mean or majority or weighted majority voting) and evaluating these approaches for their effect on the overall performance of the AI system.

2. DEVELOPED AI TOOL FOR DETECTING AND MEASURING HATE SPEECH

To develop robust AI models for hate speech processing, we utilized a BERT model which is a state-of-the-art language model known for its ability to understand and process text⁹. BERT models are machine learning models that have been trained on large datasets and that can be adapted (fine-tuned) to specific tasks. For Turkish, we used BERTurk¹⁰ which is a version trained on Turkish web data, while for Arabic we employed a similar model adapted for Arabic language.

In this section, we provide explanations about the models used for different aspects of hate speech detection. In addition, we explain the data preprocessing step and the media monitoring process.

2.1. Data preprocessing and paralinguistic features

Since social media posts are usually informal and include mentions, URLs, and paralinguistic features (e.g., emojis and hashtags) that make understanding the semantics of the text challenging. Therefore, preprocessing of such textual data is commonly employed to remove some of these elements and reduce linguistic variance. We also followed this strategy and preprocessed the tweets before using them in the machine learning models. Since URLs and usernames do not generally provide useful information regarding hate speech detection or classification, we removed them. As a positive side effect, removing URLs and usernames ensures that individuals who post hate speech are not directly targeted and this helps to maintain user privacy and is an ethical manner to handle sensitive data.

Emojis are Unicode graphic symbols that are used as abbreviations for thoughts and emotions. Graphic emojis have become an integral part of communication today. For instance, a thumbs-up/thumbs-down emoji can indicate the speaker's agreement or disagreement on a subject without recourse to words. Hashtags are very important in social media because they allow messages to be linked around a particular topic. In language processing studies, hashtags are often removed during preprocessing in an effort to simplify the subsequent modeling. However, hashtags are sometimes used as words in the middle of a sentence and removing

9 Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, 4171–86.

10 Schweter, Stefan. 2020. "Berturk - BERT Models for Turkish." *Zenodo*, April 27, 2020. <https://doi.org/10.5281/zenodo.3770924>.

them can completely alter the meaning of the sentence. In order to study the effect of these paralinguistic features (emojis and hashtags) on the performance of hate speech detection, we built different machine learning models that use tweets with these features either kept within the tweets or removed from them. In this way, we tested the models with different tweet configurations where the tweets include or do not include emoji tokens, emoji textual aliases, and hashtags, in addition to the text of the tweet. In our preliminary tests, we achieved best performance (accuracy) with “text + emoji tokens + hashtags” configuration and used it to train our BERT classifier models.

2.2. Hate speech detection and classification

In the first experiment, we developed machine learning models to detect whether a given tweet contains hate speech and to categorize the tweet using the categories shown in Table 2. Hate speech detection is a 2-class classification, where the classes are “no hate speech” and “hate speech”. For categorization, we used two different settings. One of the settings involves 6 classes, which are the “no hate speech” class and the four categories in Table 2. In the other setting, we designed a 4-class classification model by merging the categories 2 and 3 as a single category, and 4 and 5 as a single category in order to simplify the task.

We use about 80% of data for training and 20% of data for testing. Table 4 shows the performance of these models using the accuracy metric (the ratio of the number of correct classifications to the total number of data). As expected, the performance increases when the task becomes simpler, reaching 84.87% accuracy in the 2-class (hate vs non-hate) problem.

Table 4. Result of hate speech classification models

| Model | 6-class | 4-class | 2-class |
|----------|---------|---------|---------|
| Accuracy | 80.46% | 80.55% | 84.87% |

2.3. Hate speech strength prediction

In the second experiment, we formulated a regression problem using BERTurk to measure the strength of hate speech on a scale from 1 to 10. The model achieved a Root Mean Squared Error (RMSE) of 1.67. RMSE quantifies the difference between predicted and actual values, placing greater emphasis on larger errors by squaring them before averaging. A lower RMSE indicates more accurate predictions, demonstrating the model’s effectiveness in capturing the varying intensity of hate speech.

2.4. Target identification

Another important task is to identify both the general target group(s) and specific group(s) of the hate speech discourse. Recognizing the target group and specific group of such messages is vital for evaluating the potential harm to these various identity groups. We identify the target group category (e.g., gender, nationality) and the specific group category (e.g., women, refugees, LGBTI+s) within each hate-filled tweet. Target Group Classification involves identifying whether a text targets a group and determines the general target group category (e.g., gender, nationality). Specific Group Classification further specifies the individual group within the identified target group category (e.g., women within gender, refugees within nationality, LGBTI+s within sexual orientation).

As a third experiment, we developed a general target group identification model across four targets (Please see the “Target identification” section in the guidelines for detailed information.):

- 0: Target group not specified or not present
- 1: Country/nationality/race/ethnicity
- 2: Religion
- 3: Gender/sexual orientation

Table 5. Explanations for general target identification

| Target Identification | Explanation |
|--|--|
| 0: Target group not specified or not present | Discourses in which the target identity is vague or not explicitly defined. |
| 1: Country/nationality/race/ethnicity | Discourses in which the individual(s)/group are targeted due to their country/nationality/race/ethnicity. The categories in this project are listed as follows: Refugees, Israel-Jews, Greeks, Armenian, Kurdish, Arab (see Section 2.5 for the complete list) |
| 2: Religion | Discourses in which the individual(s)/group are targeted due to their religious identity. The categories in this project are listed as follows: Jews, Alevi (see Section 2.5 for the complete list) |
| 3: Gender/sexual orientation | Discourses in which the individual(s)/group are targeted due to gender and/or sexual orientation. The categories in this project are listed as follows: LGBTI+, women (see Section 2.5 for the complete list) |

Table 6 shows the target identification results for each group and also for the whole dataset in terms of class-based and overall F1-scores. The F1-score is a metric that combines the precision score (how much of the predictions for a class actually belong to that class) and recall (how much of the data that belong to that class are correctly predicted) in a single score. In classification problems, the F1-score is preferred to accuracy when the number of samples in different classes differ significantly. As shown in Table 6, the average F1-score (macro average) for our target detection model is 60.0% and the accuracy (which is the same as the micro averaged F1 score) is 73.0%. The target group involving country, nationality, race or ethnicity is reliably identified, whereas identifying religion or sexual orientation groups are less successful.

Table 6. Result of multi-label general target group identification model

| | F1-score | Support (size) |
|---|-------------|----------------|
| Target group not specified or not present | 0.70 | 870 |
| Country/Nationality/Race/Ethnicity | 0.82 | 1349 |
| Religion | 0.46 | 256 |
| Gender/Sexual Orientation | 0.43 | 49 |
| Average (micro average) | 0.73 | 2524 |
| Average (macro average) | 0.60 | 2524 |

2.5. Specific group identification

As the fourth experiment, we developed an specific group identification model across 11 categories as follows:

- 0: No-group
- 1: Refugees
- 2: Jews
- 3: Greeks
- 4: Armenians
- 5: Alevis
- 6: Kurds
- 7: Arabs
- 8: LGBTI+s
- 9: Women
- 10: Other-groups

Table 7 shows the individual group classification results of the developed model.

Table 7. Results of specific group classification

| 11-class classification model | |
|--------------------------------------|------|
| Accuracy | 0.96 |

2.6. Span detection

While the goal of hate speech detection is to identify whether a given text contains hateful content, span detection aims to pinpoint the location of hate speech indicators within the text, in order to provide better insight.

To develop the span detection model, we formulated span detection as a token classification task, where each token (subwords in this case) is labeled with a tag that denotes whether it is part of a hateful span. In order to train the BERTurk model with this objective, we need the tweet dataset labeled with the hate speech spans from the tweets. However, the annotations were conducted at the tweet level, meaning that individual tweets were labeled rather than the spans, and we needed to label the tweets so we could obtain spans that indicate hate speech. To achieve this, we selected tweets targeting distinct groups (Armenian, Greek, Jew, Arab, Immigrant/Refugee, LGBTI+, Alevi, Kurdish) of varying sizes from tweets annotated by the three annotators who were all in agreement. We omitted Arabic tweets from our analysis due to their lack of diversity—they only targeted immigrants/refugees and were limited in number. We then automatically extracted spans indicating hate speech through prompting the GPT-4 large language model, filtering hallucinated spans and resolving minor text variations. Two annotators reviewed the GPT-4 spans, and disagreements were resolved by a third annotator selecting the most appropriate annotation. This process resulted in 3697 tweets, shown in Table 8.

Table 8. Number of tweets used for span detection per target group

| Target Group | Number of Tweets |
|-------------------|------------------|
| Jewish | 1132 |
| Greek | 1119 |
| Armenian | 628 |
| Arab | 337 |
| Immigrant/Refugee | 242 |
| Alevi | 127 |
| Kurdish | 63 |
| LGBTI+ | 49 |

We combined this data with non-hate filled tweets annotated by the three annotators in agreement to train the span identification model. The resulting model showed a 41% F1-score in detecting hate filled spans, meaning it can correctly detect almost half of this kind of span, maintaining a balance between accuracy and completeness. Our tool leverages this model to annotate words that signal hate speech in any given text. The model’s moderate performance can be attributed to the small size of the annotated dataset and the challenging nature of the problem.

2.7. Hate speech detection in Turkish print media

Print media is another outlet where hate speech targeting specific groups and identities in Turkey persists. To enable effective content moderation and combat hate speech in this medium, we developed a suite of models that analyze distinct aspects of hate speech in Turkish news articles. These models include one for detecting the presence of hate filled discourse, another for categorizing the type of hate speech, and a third for identifying the specific target group. The model for detecting hate filled discourse is a binary classification model that predicts whether an article contains hate filled or non-hate filled content. The hate speech categorization model is a multi-class classification model that classifies content in three categories. Notably, this model aligns with our tweet-based hate speech categories (given in Table 2) as it *excludes only the exclusionary/discriminatory discourse* category due to its absence in the news dataset. The target group detection model identifies the specific group targeted by hate speech, formulated as an 11-class classification task aligned with the tweet-based model.

All models were built on BERTurk, leveraging its understanding of Turkish text, and were fine-tuned using the Foundation’s news media archive. The archive was originally an image-based format that required preprocessing to convert it into high-quality text. This transformation was achieved using EasyOCR for image-to-text conversion, followed by post-processing with GPT-4, which significantly enhanced the quality of the dataset. The resulting dataset contains diverse articles of varying lengths and target groups, including ethnicities, nationalities, and religious communities. The dataset includes over 200 target groups represented, sourced from 1,210 media outlets, covering both local and national news. To create a balanced training set, the dataset includes hate filled articles complemented by a random sample of non-hate filled news articles taken from the same time frame. The dataset contains the following distribution of hate filled news articles: 10,198 articles categorized as *exaggeration/generalization/attribution/distortion*, 1,199 as *symbolization*, and 121 articles containing both categories (treated as a combined class). The dataset also includes 2,279 articles classified as *threat of enmity/war/attack/murder/harm*, 644 as *swearing/insult/defamation/dehumanization*, and 13 articles containing both (forming another combined class). Additionally, the hate speech category model includes 14,715 non-hate filled articles as a separate class.

The models performed robustly in evaluation. The binary hate speech detection model achieved an F1-score of 87.49%, demonstrating its strong ability to differentiate hate filled from non-hate filled content. The target group detection model achieved an F1-score of 82.38%, performing lower than its tweet counterpart. This difference can be attributed to two main factors: the class imbalance in the dataset and the inherent discrepancy between tweets and news content. News articles typically contain longer, more complex text with multiple potential target groups and more nuanced language compared to the more direct and concise nature of tweets, making the classification task more challenging for news content. The hate speech category prediction model, while achieving a moderate but reasonable F1-score of 78.57%, highlighted challenges in capturing the nuances of complex hate speech categories. These results underscore the models’ effectiveness in understanding and analyzing the polarized and often hostile discourse in Turkey’s print media, providing essential tools for automated content moderation and analysis.

2.8. Media tracking & analysis

Another important step in combating hate speech is real-time and continuous monitoring and analysis of the potential sources. To this end, we enhanced our tool with a component that continuously tracks the content on the X platform for specific keywords, namely “Syrian” and “Refugee”. This component fetches tweets

containing these keywords every 90 minutes using the X API, then labels them using our hate speech detection models, demonstrating their functionality and facilitating easy analysis and reporting of this discourse. The output is displayed as a graph showing the daily percentage of tweets containing hate filled content, with days on the x-axis and percentages on the y-axis. While this component is promising and helpful for tracking content and further analysis, it is limited by the small number of tweets retrieved daily due to the quota and policies of the X platform. We believe it could be valuable for analyzing different discourses when increased accessibility to the potential content is given.

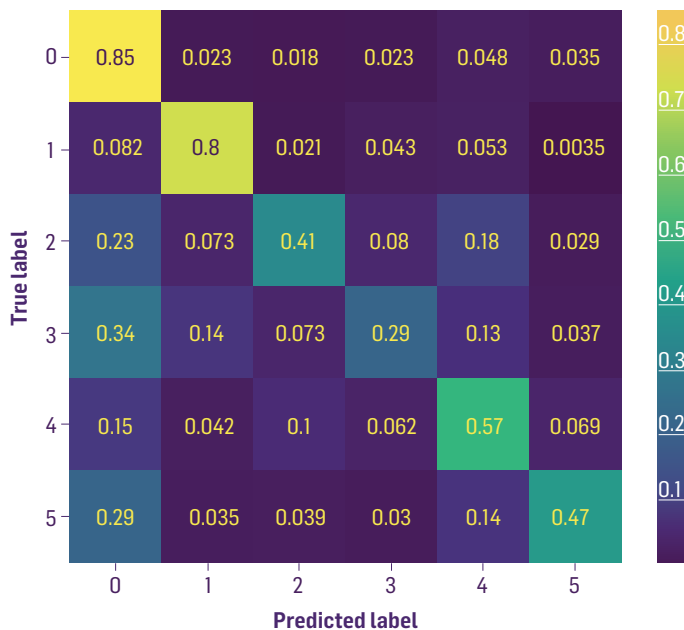
3. ERROR ANALYSIS

To better understand the performance of the developed AI tool, we analyzed its performance on the test set from different perspectives. We conducted detailed error analysis on the experiment of Section 2.2 with the 6-class setting.

Figure 1 displays the confusion matrix for hate speech categorization, which illustrates the percentage-wise misclassification between different categories. It is important to note that each row adds up to 1, meaning that each value in the row represents the percentage of data with the corresponding True Label that was classified as the corresponding Predicted Label. For example, the value in the entry corresponding to True Label 2 (Symbolization) and Predicted Label 4 (Swearing, Insult, Defamation, Dehumanization), which is 0.18, indicates that 18% of the tweets that actually belong to class 2 were incorrectly classified as class 4. Similarly, the value in the entry corresponding to True Label 3 (Exaggeration, Generalization, Attribution, Distortion) and Predicted Label 3, which is 0.29, indicates that 29% of the tweets that belong to class 3 are correctly classified as class 3.

Overall, the model exhibits a tendency to classify tweets into class 0 (no hate speech), as evidenced by the 85% accuracy for this class. This observation is further supported by the high confusion rates for classes 2, 3, 4, and 5 being incorrectly predicted as class 0. Class 1 (Exclusionary, Discriminatory Discourse) also achieves a relatively high accuracy of 80%. However, the other classes demonstrate lower accuracy levels, with class 3 (Exaggeration, Generalization, Attribution, Distortion) showing the lowest performance at 29%. Additionally, there is notable confusion between class 2 (Symbolization) and class 4 (Swearing, Insult, Defamation, Dehumanization), as one of them is often misclassified for the other.

Figure 1. Confusion matrix for hate speech categorization



We then analyze annotator agreement for tweets that were correctly and incorrectly classified. To assess annotator agreement, we again use Krippendorff’s alpha coefficient, where higher values indicate stronger agreement among annotators. The results presented in Table 9 show that tweets correctly classified by the model tend to have higher annotator agreement. This suggests that the model also finds it challenging to correctly classify tweets that are difficult for annotators to categorize. As an additional figure, we also measured the accuracy of hate speech categorization on the subset of tweets for which all three annotators assigned the same categories. The model resulted in 90.17% accuracy, a much higher accuracy when compared to the 80.46 % accuracy (Table 4) across all tweets.

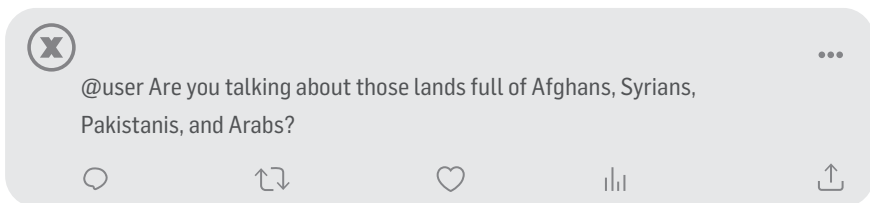
Table 9. Krippendorff’s alpha coefficients for correctly and incorrectly classified tweets

| | Krippendorff’s Alpha Coefficient |
|------------------------|---|
| Correctly Classified | 0.387 |
| Incorrectly Classified | 0.202 |
| All Test Data | 0.335 |

These results and analysis show that detecting and classifying hate speech is a challenging task, especially for discourse where there is annotator disagreement.

As a final analysis, we examined a selection of tweets that were misclassified by the model and we provided potential explanations for the challenges the model faced in correctly classifying them. Automatic classification of hate speech is a challenging task: in addition to the standard difficulties in natural language understanding, hate speech detection in tweets and media contains some further challenges, summarized below:

- **Lack of context:** Understanding the full intent of the author is a challenging task of natural language understanding. When considering short tweets that are studied without any context/background (as when they are annotated or machine-classified in isolation), this lack of context becomes the main challenge. The issue is exacerbated for tweets that are replies to previous tweets. For instance:



Without available context, it is unclear whether the author is referring to specific countries or to Turkish territories inhabited by people from these countries. Whether the tweet is hate filled or not relies on such distinctions. Therefore, the absence of context, as seen in the tweet above, makes accurate annotation more difficult.

- **Hidden intent:** Some tweets contain hate speech elements, but the content itself is not actually hate filled; rather, the author is actually condemning those who employ hate speech that is directed at a specific identity. While this is a typical problem in natural language understanding, it is more challenging to understand user intent in the case of short tweets.



as if the Syrians did not experience the same pain as us, as if they did not suffer losses... They say let's hang the Syrians and slaughter the Afghans while having fun.



This can also be achieved through sarcasm. By using sarcastic language, users can disguise their harmful intent, making their statements seem less aggressive or even humorous. Sarcasm allows users to avoid accountability by framing their remarks as jokes or irony. As a result, it can contribute to the normalization of hate speech, making it more difficult to identify and address.



You always say Kurdish voters, please do not discriminate, Turkish voters, Georgian voters, Circassian voters, Bosnian voters, Arab voters, Armenian voters, Romani voters, Alevi voters, Sunni voters etc. are the only ones that remain.



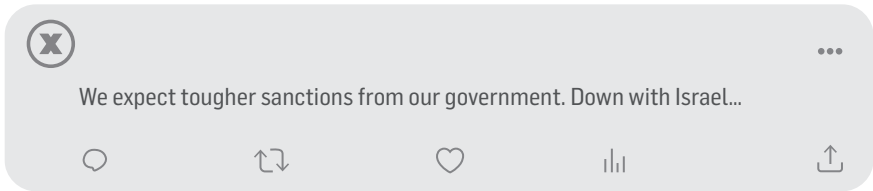
- **Opposing text and hashtag:** Some tweets include a hateful hashtag but express views that oppose the sentiment of the hashtag. In such cases, the labeling becomes ambiguous, but the model tends to classify the content as hate filled.



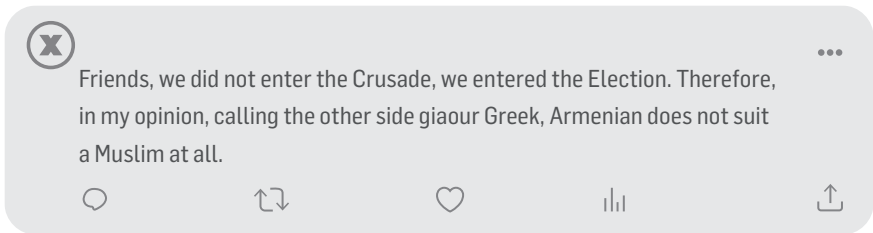
The looter you think is Syrian may also be a child of your own country.
Filth exists in every nation #idontwantrefugeesinmycountry



- **Incorrect labeling:** Despite a thorough and careful annotation process, some tweets may still be mislabeled by the annotators. These mislabeled tweets in the training data can confuse the model and result in degraded performance.



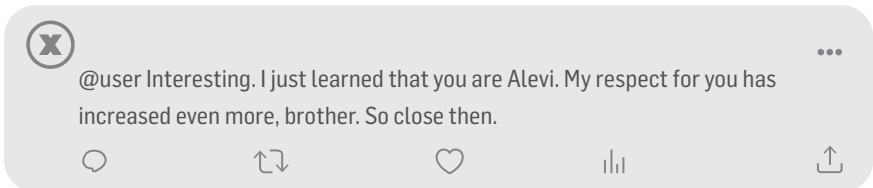
This tweet was labeled as “No hate” by the annotators whereas the model outputs “Threat of enmity, war, attack, murder, harm”, which is in fact the correct categorization for this tweet.





This tweet was labeled as “No hate” by the annotators whereas the model outputs “Symbolisation”, which is in fact the correct categorization for this tweet.






As part of the comparison between true and predicted categories, we also show some tweets that are correctly classified by the model for each category:

No hate:










Exclusionary, discriminatory discourse:

 I rejected the requests of Arab farms who wanted to join my neighborhood in Hayday 



    






Symbolization:

 You make big calculations, Fatih, this makes you the lawyer of the Greek seed İmamoğlu... 



    






Exaggeration, generalization, attribution, distortion:

 We are also patriots, rest assured. Let's see who will be right. I think that you have been deceived by these tales and that great misery awaits the Turkish nation. May it be good for our homeland. I hope that you will be right and the Turkish nation will not become needy for bread under Arab occupation. 

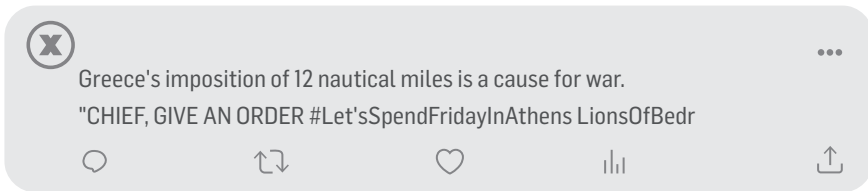
    

Swearing, insult, defamation, dehumanization:

 I can't believe it, they included Israel instead of a dog, who are you to equate a big dog with Israel? The dog demands its rights from you, in the other world, Israel cannot be the feces of a dog 😞 

Threat of enmity, war, attack, murder, harm:



4. LIMITATIONS

While developing the hate speech detection tool, several limitations were encountered that may impact the performance and generalizability of the model. One significant challenge was data collection, largely due to the evolving policies and quota restrictions put in place by X (Twitter). These changes limited access to a comprehensive datasets by reducing the variety and volume of tweets that could be collected, and this may affect the model's ability to generalize across different types of hate speech.

Another limitation arose from the annotation process. Despite training, achieving consistent annotator agreement proved to be difficult. Annotators came from diverse backgrounds and each brought their own perspectives and interpretations to the task, which led to variability in how hate speech is identified. This variability underscores the subjective nature of detecting hate speech and highlights the complexities involved in reaching a unified understanding, even among trained individuals.

Additionally, the dataset was made up of individual tweets rather than of complete threads, and this often resulted in a lack of context. This lack of contextual information made it harder for annotators to accurately assess the intent and nuance behind each tweet, as isolated statements can be ambiguous or misleading when extracted from a broader conversation. Furthermore, we did not analyze the spread or virality of the tweets, thereby missing an opportunity to explore how hate speech propagates and intersects with issues like disinformation. Understanding the dissemination patterns of hate filled content could have provided valuable insights into its impact and reach, which are crucial considerations for comprehensive hate speech mitigation strategies. In addition, during the development stages of our tool, images were not incorporated into the training process. In certain cases, we observed that only when a tweet is analyzed in conjunction with its accompanying image does it become evident that hate speech is present. Therefore, the examination of visual content is of significant importance. We acknowledge that this remains a crucial area for future research.

The dynamic and evolving nature of hate speech, including new slang, coded language, and cultural references, also presents an ongoing challenge for the detection tool. Adapting the model to keep pace with these changes requires continual updates and retraining with new data to maintain accuracy. These factors collectively highlight the need for ongoing refinements and the importance of understanding the broader limitations when interpreting the results of the hate speech detection tool.

3 CONCLUSION

This report was prepared by researchers of the Hrant Dink Foundation, Boğaziçi and Sabancı Universities as part of the **Utilizing Digital Technology for Social Cohesion, Positive Messaging and Peace by Boosting Collaboration, Exchange and Solidarity** project. It reflects the different perspectives and joint efforts carried out on hate speech of disciplines such as communication, linguistics, cultural studies and computer sciences. As can be seen from the examples in the labeling guide and the definitions made based on these examples, although hate speech data is often difficult to understand, full as it is of ambiguities and inconsistencies, our fundamental goal in this project is to gain a deeper and clearer understanding of hate speech in order to combat the problem more effectively by using the hate speech detection tool that we have developed. We hope that this method and the developed tool can be used in future studies or at least provide a foundational starting point for others.

Appendix A: Labeling interface

Tweet

@HDPgenelmerkezi a real Kurd would be a Muslim they would not let an Armenian traitor into their party. Armenians' purpose is to cause a war between Kurdish and Turkish and to found Western Armenia in Kurdish lands. They are all secret agents of Armenians, Jews. Do not send your kids to PKK, let them send theirs since they are Kurdish....

After clicking one of the following boxes, you can select words or phrases that cause hate speech. (Maximum 3 words or phrases can be selected)

Triggering Word ¹ Swearing/Insult ² Enmity Discourse ³

Overall Attitude and Stance

Please choose only 1 option

- Not Sure^[4] Anti-immigrant/
Refugee^[5] Neutral^[6] Irrelevant^[7]

Target Group

Multiple choice is available if the target group is more than one.

- Demographic/
Socioeconomic/
Race/Ethnicity^[8] Country/
Nationality^[9] Religion^[0] Gender^[a]
- Specific Opinion/Status/
Practice, Professional
Position Group^[e] Target group
is unclear or
absent^[t] Target group
is more than one^[a] Sexual
Orientation^[w]

Hate Speech Strength

Please choose only 1 option

- Not
Sure^[s] 0^[d] 1^[f] 2^[g] 3^[z] 4^[x]
- 5^[c] 6^[v] 7^[b] 8^[y] 9^[i] 10^[o]

Exclusive, Discriminatory Discourse

These are discourses in which a community is seen as negatively different from the dominant group in areas such as the benefit from rights and freedoms and inclusion in society.^[p]

Hate Speech Category

More than one category can be selected; if the target is unclear, the category should be selected according to the text content. If “not sure” is checked in the overall attitude/stance section, “not sure” should also be checked here.

| | | |
|---|--|--|
| <p><input type="radio"/> Not Sure ^[j]</p> | <p><input type="radio"/> There is no hate speech ^[k]</p> | <p><input type="radio"/> Symbolization</p> <p><i>These are discourses in which an element of identity itself is used as an element of insult, hatred or humiliation and the identity is symbolized in such manners.</i>^[l]</p> |
| <p><input type="radio"/> Exaggeration, Generalization, Attribution, Distortion</p> <p><i>These are discourses that draw larger conclusions and inferences from an event, situation or action, manipulate real data by distorting it, or attribute individual events to the whole identity based on their agents.</i>^[n]</p> | <p><input checked="" type="radio"/> Swearing, Insult, Defamation, Dehumanization</p> <p><i>Discourses that include direct profanity, insult, contempt towards a community, or insults by characterizing them with actions or adjectives specific to non-human beings.</i>^[m]</p> | <p><input type="radio"/> Threat of Enmity, War, Attack, Murder, or Harm</p> <p><i>These are discourses that include expressions about a community that are hostile, avoke war or express a desire to harm the identity in question.</i></p> |

Offensive Language

Please choose only 1 option

None

Low

High

© HRANT DINK FOUNDATION PUBLICATIONS, 2025

a5.u.lis DISCOURSE
DIALOGUE
DEMOCRACY
LABORATORY