How do we counter dangerous speech in Italy?

How do we counter dangerous speech in Italy?

Vittoria Tonini[1,2,*], Simona Frenda[2,3], Marco Antonio Stranisci[1,2] and Viviana Patti[1]

[1]Computer Science Department, University of Turin, Torino, Italy
[2]aequa-tech, Torino, Italy
[3]Interaction Lab, Heriot-Watt University, Edinburgh, Scotland

Abstract

The phenomenon of online dangerous speech is a growing challenge and various organisations try to prevent its spread answering promptly to hateful messages online. In this context, we propose a new dataset of activists' and users' comments on Facebook reacting to specific news headlines: AmnestyCounterHS. Taking into account the literature on counterspeech, we defined a new schema of annotation and applied it to our dataset, in order to examine the most used counter-narrative strategies in Italy. This research aims to support the future development of automatic counterspeech generation. This paper presents also a comparative analysis of our dataset with other two datasets in Italian (Counter-TWIT and multilingual CONAN) containing dangerous speech and counter narratives. Through this analysis, we will understand how the environment (artificial vs. ecological) and the topics of discussions online influence the nature of counter narratives. Our findings highlight the predominance of negative sentiment and emotions, the varying presence of stereotypes, and the strategic differences in counter narratives across datasets.

Keywords

Counter narrative, Linguistic analysis, Abusive language, Italian language

1. Introduction and Background

Recently, the attention about dangerous speech (DS) online has increased in different sectors, ranging from initiatives for monitoring the DS' spread in particular in Italy (e.g., by VOX[1], or by researchers like Capozzi et al. [1]) to prevent the escalation of DS online using methods of detection and removal of dangerous contents (e.g., following the policies of social platforms). Moreover, specific actions of countering DS online like the Amnesty Task Force on Hate Speech[2], that reassembles specialized activists who actively intervene writing counterspeech, were promoted[3] in response to potential or effective dangerous speech or news on various topics. In this context, the new techniques of Natural Language Understanding (NLU) and Natural Language Generation (NLG) can play a very important role. On DS detection, the literature is vast [4, 5] and covers various nuances of DS [6, 7], different types of manifestation (i.e., explicit and implicit, [8]) and co-occurrences with other psychological and linguistic phenomena, like stereotypes [9] and sarcasm [10]. Regarding works on countering DS, some studies focused on imitating the operators of Non-Governmental Organizations (NGO) in their intervention in online discussions, or selecting the most suitable responses from a database [11] or creating generative models able to reply automatically to hateful content using counter narratives (CN) avoiding hallucinations [12]. The development of NLU and NLG models are mainly based on data-driven approaches, that imply the creation of a specific dataset to detect DS or generate adequate CN. According to the survey by Bonaldi et al. [2], in literature, the available datasets in languages different from English are very few. Among them, currently, only two datasets contain Italian texts: CONAN [13] and Counter-TWIT [14].

The creation environment of CONAN is artificial (i.e., activists have been asked to write CN to specific hateful comments) and the one of Counter-TWIT is entirely ecological (i.e., collection of tweets written by users). In this scenario, in our work we propose a new dataset, **AmnestyCounterHS**, that differently from the existing ones, reflects the real action of activists online. Indeed, our dataset, compiled from Facebook, includes interactions guided by the Amnesty Task Force on Hate Speech (HS), representing an ecological and spontaneous context. Here, the intervention of counterspeech is guided by Amnesty International activists who decided to intervene under certain posts potentially dangerous spread by online newspapers or users (e.g., verbal attacks to

CLiC-it 2024 – Tenth Italian Conference on Computational Linguistics, 4 – 6 December 2024, Pisa, Italy
*Corresponding author.
vittoria.tonini@edu.unito.it (V. Tonini); simona.frenda@aequa-tech.com (S. Frenda); marco.stranisci@aequa-tech.com (M. A. Stranisci); viviana.patti@unito.it (V. Patti)
0000-0002-6215-3374 (S. Frenda); 0000-0001-9337-7250 (M. A. Stranisci); 0000-0001-5991-370X (V. Patti)

[1]http://www.voxdiritti.it/la-nuova-mappa-dellintolleranza-7/ (webpage visited on july 2024)
[2]https://www.amnesty.it/entra-in-azione/task-force-attivismo/ (webpage visited on july 2024)
[3]As reported in Bonaldi et al. [2], the terms 'counterspeech' and 'counter narratives' are used interchangeably in Natural Language Processing field (NLP), and both can be considered as "communicative actions aimed at refuting hate speech through thoughtful and cogent reasons, and true and fact-bound arguments" [3].

women, immigrants, and so on).

Moreover, inspired by existing strategy taxonomies [15, 13, 14], we mapped a more complete taxonomy inclusive of both existing and new strategies found in our dataset. This new resource allows us to analyze the used strategies of CN in the Italian language across different types of messages and contexts (CONAN, Counter-TWIT, AmnestyCounterHS). By comparing these datasets, we propose to examine: 1) which strategy of CN is the most used in the different contexts and discussions online; 2) which the differences are in terms of sentiments, emotions, and the presence of stereotypes, between potentially dangerous messages posted online and the counterspeech produced by activists/users in all the datasets.

The importance of understanding how these strategies of CN are used relies on the need to raise social awareness about real events, the necessity to be correctly informed about facts (avoiding fake news), as well as to be conscious of the consequences of dangerous speech in the target groups [16].

2. Datasets

In this section, we describe existing dataset of CN in Italian (CONAN and Counter-TWIT), and the creation of AmnestyCounterHS.

CONAN[4] is a multilingual and expert-based dataset of DS/CN pairs in English, French and Italian, focused on Islamophobia. The original dataset consists of 4078 pairs over the 3 languages. The dataset has been augmented through translation (from Italian/French to English) and paraphrasing, which brought the total number of pairs to 14.988. The dataset was created by Chung et al. [13] in an artificial environment and consists of expert-based data. The DS/CN pairs were collected through niche sourcing from three different NGOs in the United Kingdom, France, and Italy. Consequently, both the responses and the dangerous speech content are expert-based, composed by operators specifically trained to counteract online dangerous speech. For this paper we considered only the Italian pairs, which are 3,213 in total. Here is an example of a pair from the CONAN dataset:

1) DS: *"Noi li ospitiamo nel nostro paese, forniamo un aiuto economico e loro ci uccidono: sono da considerarsi più simili agli animali che alle persone."*[5]

CN: *"I criminali sono in tutti i popoli e di tutte le religioni, per fortuna una minoranza, non si deve mai generalizzare. Lei è italiano quindi mafioso?"*[6]

Counter-TWIT[7] dataset is made up of 624 pairs of tweets and their replies. Data were collected in an ecological environment using keywords to take texts from profiles of activists, organisations, or pages especially devoted to calling out common instances of discrimination. In this data we encounter both DS(16) and CN(81), but they are not DS/CN pairs such as in CONAN, but rather consist of tweets and their replies.

2) Tweet: *"In Italia spesso funziona cosi: La vittima diventa automaticamente il colpevole."*[8]

Reply: *"Nelle violenze in particolare"*[9]

AmnestyCounterHS is a collection of posts and relative comments gathered from Facebook. The data collection strategy was driven by the work of the Amnesty Task Force on HS, a group of activists that produce CN against discriminatory contents spread by online newspapers and users. During the task force, the activists identified some posts containing news headlines that probably convey or incite hate speech and assigned them a topic based on the specific target of the news headline. Among the various topics covered in the dataset are: women, migrants, LGBTQIA+, solidarity, and environmental issues. During their activities they built a database of hateful contents against which they got activate between 2020 and 2023. Starting from this database, we collected all the news headlines detected by activists in the March 2020, 2021, 2022, and 2023. Then we gathered and anonymized all the comments in reply to them, for a total of 39,582 users' comments and 2,010 activists' comments. For our work, we used only 10,670 users' comments selected from users who replied at least 5 times. This approach allowed us to focus on users with more interactions. Table 1 reports the information of all corpora. This enabled us to obtain three collections of text: *i.* a set of news headlines that incite the use of dangerous speech; *ii.* a set of comments written by activists replying to users or written directly under post; *iii.* a set of comments written by users replying to activists or other users, or written directly under posts. Table 2 shows the number of comments written by users and activists per type of interaction.

3) Headline: *"Migranti, riprendono gli sbarchi. E il coronavirus ora avanza in Africa"*[10]

Comment: *"salve, legga l'articolo per favore, non sono ripresi gli sbarchi, in realtà stanno diminuendo costantemente, non si preoccupi....è "il Giornale" che fa gli scherzoni"*[11]

[4] https://github.com/marcoguerini/CONAN

[5] "We host them in out country, provide them an economic aid, and they kill us: they should be considered more like animals than humans."

[6] "Criminals exist among all people and religions, fortunately as a minority, one should never generalise. You are italian, so are you a mafioso?"

[7] https://github.com/pierpaologoffredo/Counter-TWIT/blob/main/Readme.md

[8] "In Italy it often works like this: the victim becomes guilty."

[9] "Particularly in cases of violence"

[10] "Migrants, the landings resume. Coronavirus is now spreading in Africa"

[11] "Hi, please read the article, landings did not resumed, in fact they are decreasing, don't worry ... "il Giornale" is playing tricks"

Dataset	# Pairs	Pair type	Environment	Topic
CONAN [13]	3,213	dangerous speech - counterspeech	artificial	islamophobia
Counter-TWIT [14]	624	tweet - reply	ecological	multiple
AmnestyCounterHS	12,714	news headline - comment	hybrid	multiple

Table 1
Information about CONAN, Counter-TWIT and AmnestyCounterHS datasets.

Type of interaction	Number of interactions
User replying to user	16,423
User replying to activist	909
User replying to post	22,016
Activist replying to user	1,521
Activist replying to post	489

Table 2
Number of interactions by type.

Schema of annotation The proposed annotation schema[12] includes different layers focused on the identification of linguistic style, support of CN or DS, and detection of textual spans that encode CN's strategies or DS implicit and explicit manifestation.

The annotation is made up of four layers[13]:

1. Determine if the text is written in a **formal or informal style**[17]. This helps understand the most used style of language for both DS and CN.
2. Identify if the comment is **supporting another DS or a CN** comment. This layer distinguishes between direct DS or CN and comments that support them.
3. Identify if the comment contains DS and specify if it is **explicit or implicit**. This is important because implicit DS can sometimes be hard for machines to recognise [8].
4. Identify if the comment is a CN and which **counter narrative strategy** has been used. This helps us to identify the most frequently used strategies of CN.

We have identified nine possible CN strategies: **Informative** that is a comment with a statement that seeks to debunk or fact-check the claims made by the attacker, **Alternative** when alternatives to the statement made by the attacker are proposed, **Suggestion**, **Explicitation** in the case of a comment that explicitly clarifies something that was implicit in the DS comment, **Question** made to cause reflections in the writer of the DS comment, **Denouncing and explaining** when the writer explains why things said by the perpetrator are not acceptable, **Positive** in the case of a polite comment, **Hostile** when the writer uses aggressive tone and words,

[12]The guidelines and the dataset have been released in https://github.com/aequa-tech/external-resources.

[13]You can see some examples of the various annotation layers in Table 7 in Appendix A.

Counter-TWIT	CONAN	AmnestyCounterHS
-	Facts	Informative
Alternative suggestion	-	Alternative
-	-	Suggestion
Explicitation	-	Explicitation
-	Question	Question
-	Denouncing	Denouncing and explaining
-	Consequences	Denouncing and explaining
-	Hypocrisy	Denouncing and explaining
-	Positive	Positive
-	Affiliation	Positive
Hostility	-	Hostile
Irony/Humour	Humour	Humour
Others	-	-

Table 3
Annotation scheme mapping

and **Humour** strategy in case of humoristic, ironic or sarcastic statements (further descriptions and examples of CN strategies are presented in Appendix B). We have created this mapping, based on the annotation schemes from the existing resources in Italian [13, 14], as shown in Table 3. We cross-referenced the strategies from both schemes and added the **Suggestion** category. By using this strategy, the writer suggests actions to the attacker to encourage them to rethink their views. Here are some examples of texts where we can see this strategy: *"Legga l'articolo per favore"*[14] or *"Vada a consultare i documenti storici che parlano di loro e verifichi cosa hanno fatto"*[15].

Looking at the comments, we noticed that some of them are offensive and impolite but not dangerous towards certain categories. They reflect the intensity of discussions on specific topics, displaying **hostility towards the interlocutor** rather than targeting specific categories. For instance:

4) Comment: *"come scusa, forse non è consapevole di essere lei stessa non saper utilizzare la punteggiatura, continui pure fare figure di merda, i commenti sono pubblici"*[16]

[14]"Read the article, please"

[15]"Go consult the historical documents about them and verify what they have done"

[16]"Excuse me, perhaps you are not aware that you yourself do not know how to use punctuation, keep making an ass of yourself, comments are public"

5) Comment: *"Ormai mi limito a ridere, rispondere a certi commenti è un insulto verso noi stessi"*[17]

Another interesting observation regards the presence of negative **stereotypes** that in various cases have been identified as implicit dangerous speech:

6) Comment: *"un figlio che sia campione di moto o una figlia che faccia la ballerina"*[18]

7) Comment: *"Non chiede di sbarcare...ordina di sbarcare il che è diverso. Loro decidono dove sbarcare e quando sbarcare altrimenti speronano"*[19]

These examples illustrate how stereotypes and implicit biases are embedded in the discourse, often contributing to the perpetuation of harmful stereotypes. This is one of the reasons why we decided to do an analysis of stereotypes in our comparative analysis.

Finally, we noticed that various comments are featured with **irony**. Irony is frequently used to convey dangerous or offensive sentiments in a less direct manner [10]:

8) Headline: *"Il Giornale Pescara, magrebino aggredisce e deruba 63enne fuori dal supermercato"*[20]

Comment: *"Adesso vediamo di dargli anche la medaglia sto disgraziato"*[21]

Annotation and inter-annotator agreement The annotation has been carried out for 307 comments by two annotators with linguistics background using the LabelStudio platform (Figure 2 in Appendix C). The Cohen's kappa was computed to examine the inter-annotator agreement for all labels obtaining the results shown in Table 4. The highest results were obtained for the counter-narrative (0.66) and dangerous speech (0.62) labels. For counter-narrative strategies, the easiest to identify was **Question**, followed by **Positive**, and **Informative**. There were some difficulties related to the **Support** label. For instance, the sentence: "nessun problema, si boicotta la Disney."[22] was annotated as dangerous speech support by one annotator, while the other one did not consider it as such. It would be helpful to provide further information about this label in the annotation scheme.

Label	Cohen's kappa
Style	0.44
Presence of CN	0.66
Presence of DS	0.62
Support	0.11
Question	0.65
Informative	0.57
Positive	0.57
Hostile	0.42
Denouncing and Explaining	0.41
Humour	0.29
Explicitation	0.22
Alternative	0.20
Suggestion	0.16
Explicit DS	0.43
Implicit DS	0.33

Table 4
Cohen's kappa values for inter-annotator agreement across labels.

3. Comparative Analysis

In order to investigate the differences in terms of sentiments, emotions, and the presence of stereotypes, between potentially dangerous messages posted online and the counterspeech produced by activists/users in all the datasets, we performed three different types of analysis.

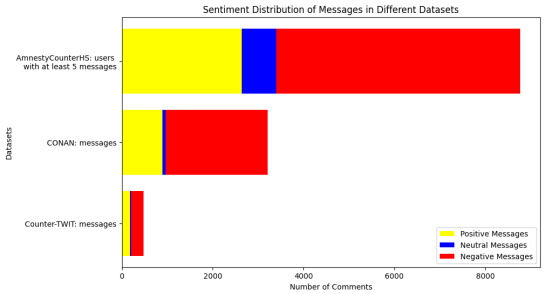
Affective: to determine which sentiment and emotion feature the intervention of who wrote CN (activists or other users) respect to other messages.
Stereotype: to understand if not only user comments contained stereotypes but also if activists or non-activists who wrote CN somehow contributed to spreading them.
Strategies: to identify the most used strategies in CN depending on the context and topic of discussion online.

3.1. Affective Analysis

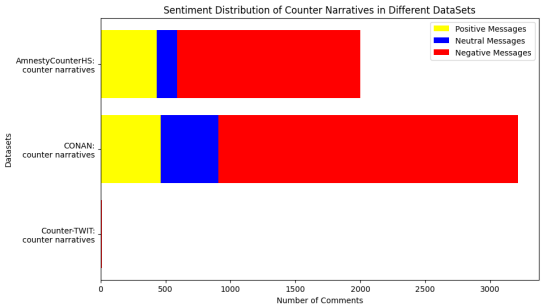
The affective analysis (Figure 1) has been performed automatically, detecting sentiment (positive, negative and neutral) and emotions (joy, sadness, fear, and anger) inferring labels from the following fine-tuned models available on the HuggingFace hub: lxyuan/distilbert-base-multilingual-cased-sentiments-student for sentiment, and Taraassss/sentiment_analysis_IT for emotion. In order to compare sentiment and emotions identified in potential dangerous speech and CN, we selected: 3,213 DS and 3,213 CN from CONAN; 543 tweets and 81 replies annotated as CN from Counter-TWIT; 10,670 users' comments and 2,010 activists' comments from AmnestyCounterHS[23].

As can be clearly seen from the sentiment analysis graphs, both in the message datasets and in the counter narrative datasets, there is a predominance of **negative** polarity. Regarding emotions, **anger** is the most prevalent emotion. Therefore, we observed this notable trend, despite the different origins of the datasets. However, it is important to point out that anger is not always a purely negative sentiment. While it often reflects strong emotions associated with dissatisfaction or conflict, it

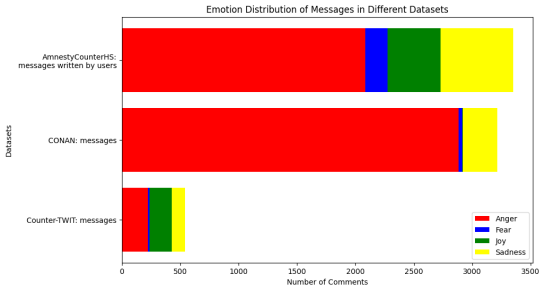
[17] "Nowadays, I just limit myself to laughing, answering certain comments is an insult to ourselves"
[18] "a son who is a motorcycle champion or a daughter who is a dancer"
[19] "They don't ask to land...They order to land, which is different. They decide where and when to land, otherwise they ram"
[20] "Maghrebian assaults and robs 63-year-old outside the supermarket"
[21] "Now let's also give this miserable a medal"
[22] "No problem, we'll boycott Disney."
[23] The assumption that these texts from activists are counter-narratives is based on the way the data was collected (activist-comment): the data collection strategy was influenced by the methodology established by the Amnesty Task Force on HS.



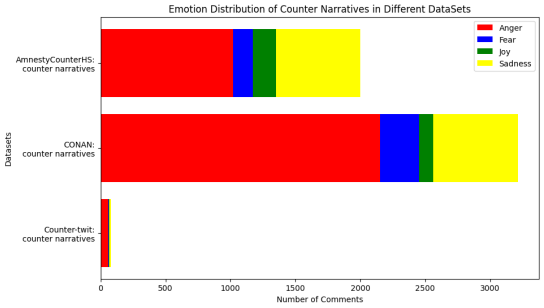
(a) Sentiment distribution in messages



(b) Sentiment distribution in CN



(c) Emotion distribution in messages



(d) Emotion distribution in CN

Figure 1: Affective analysis results.

can also highlight important debates and drive positive change, such as in the following example: "un po' di vergogna per un commento fuori luogo come il suo davanti a tanto dolore, no?"[24]. The comment, despite containing a provocation, aims to be constructive because it tries to spark a reaction in the user's thinking. In many cases, anger can be a powerful force for tackling issues and making progress. So, the anger seen in these datasets might not just show the seriousness of the issue but also the possibility for meaningful discussion and action.

For AmnestyCounterHS, we also wanted to carry out a sentiment analysis by dividing the comments based on the year of publication to see if the sentiment of the users who wrote various comments, and thus interacted more with the activists, changed over time. We expected their behaviour could become more positive after several interactions with activists. Unfortunately, we did not observe significant changes over the years, as can be seen in the figures provided in Appendix D).

3.2. Analysis of Stereotype

Like in previous analysis, the presence of stereotypes (see Table 5) has been performed automatically, inferring labels from the fine-tuned model aequa-tech/stereotype-it available on the HuggingFace hub. The set of examined data is the same of affective analysis.

Dataset	Type of text	% stereotype
CONAN	DS	85.6%
CONAN	CN	47.5%
Counter-TWIT	Tweet	12.2%
Counter-TWIT	Reply	29.6%
AmnestyCounterHS	Users' Comments	17.6%
AmnestyCounterHS	Activists' Comments	20.4%

Table 5
Percentage of presence of stereotypes.

In Table 5, we can see that in the CONAN dataset, dangerous speech messages may be more likely to contain stereotypes, while responses often serve oppositions to stereotypes present in the original messages. This pattern is not the same in Counter-TWIT and AmnestyCounterHS. Indeed, these two datasets, containing data extracted from ecological environments (respectively, Twitter and Facebook), reflect the spontaneous interaction between users and activists, where the activists themselves can explicitly mention stereotypes to oppose them or may be contributing to the creation or amplification of stereotypes.

[24] "a little shame for a comment out of place like yours in the face of so much pain, no?"

Dataset	Informative	Alternative	Suggestion	Explicitation	Question	Denouncing and explaining	Positive	Hostile	Humour
CONAN	48.3%	-	-	-	16.1%	22.7%	7.8%	-	5.1%
Counter-TWIT	-	6.3%	-	8.4%	-	-	-	61.1%	24.2%
Amnesty CounterHS	34.8%	6.7%	4.3%	4.4%	11.2%	19.8%	4.8%	5.9%	8.1%

Table 6
Percentage of different strategies

3.3. Analysis of CN Strategies

The third type of analysis focuses on the various types of counter narrative strategies used across all three datasets. Firstly, we had to map the strategy types to our guidelines, adapting the strategy labels from the different datasets to match the labels in our dataset (see Table 3). Secondly, we examined the distribution of strategies across datasets considering the type of environment (ecological, artificial) and the different topics.

In an artificial context such as that of the CONAN dataset, the most commonly used strategy is **informative**. This prevalence is expected because, in controlled environments, there is often a focus on providing factual information and raising awareness to counteract misinformation effectively. This is also the most used strategy in our dataset, where CN were written by activists. In an ecological context like that of the Counter-TWIT dataset, the most frequently used category is **hostile**. This is understandable, **as real-world interactions often involve more emotional and aggressive responses**, reflecting the more spontaneous and less regulated nature of online discourse. The use of this CN strategy is interesting, because usually it is not suggested to use it. Despite this, it can happen that ones get irritated when facing dangerous speech. The **hostile** strategy can be considered somewhat the opposite of **positive**, which instead represents a very polite attitude. Moreover, we wanted to see also which the most used strategies were according to the topic. Analysing our dataset we obtained that for the topics LGBTI, migrants and solidarity, the most frequent strategy was **informative**. For the topic "women", the most used strategy was **alternative**, while for the topic "environment", the prevalent strategy was **denouncing and explaining**.

We also conducted a manual analysis of the corpus to understand if there were any interactions between users and activists that proved more effective than others. In particular, we observed that an activist who employed the **Polite** strategy in some comments managed to engage quite well with a user. An example of a comment written by the activist is: *"interessante. Mi permetta, senza polemica, di puntualizzare alcune inesattezze che ha riportato, forse nella velocità"*[25]

[25]"Interesting. Allow me, without being argumentative, to point out a few inaccuracies you mentioned, perhaps due to haste."

4. Discussion and Conclusion

In this paper, we examine the strategy of CN used in various contexts, looking at their characteristics and typology across different datasets in Italian: CONAN, Counter-TWIT, and AmnestyCounterHS. Thanks to this comparative analysis, we noticed that different environments and topics affect the type of strategy used by activists or users who want to counter DS [18].

One of the main points that we want to underline is the importance of the conversational context [19, 20, 21, 22]. In our dataset, AmnestyCounterHS, the annotators showed difficulties to understand the position of the author of the message, without the entire conversational thread. For instance, let us consider this comment written under some news about COVID-19: *"Infatti. Ampiamente dimostrato"*[26]. Without the full conversation, it is challenging to determine whether this comment is supporting or contradicting an argument about COVID-19. Similarly, let us take a look at the comment: *"Grande argomentazione, scuola di Demostene? #posailfiasco"*[27] written under this newstitle: *"Un milione di profughi sono ostaggio di Erdogan"*[28]. We can clearly see that the comment is ironic, but we cannot understand its stance on integration. For this reason, future developments in automatic counterspeech generation should focus on incorporating comprehensive conversational threads to enhance accuracy and relevance. This approach will be fundamental to create effective AI-driven counter-narrative systems.

5. Ethical Statement and Limitation

The data in the corpus was collected from public pages and has been anonymised. IDs were created by us, and the links from which the comments were taken have been removed, therefore it is not possible to trace the original comments. Moreover, in the released version, the identities of the annotators are not revealed. An ethical concern is related to the characteristics of the annotators

[26]"Indeed. It's been extensively demonstrated"
[27]"Great argument, is it from the school of Demosthenes? #giveitup"
[28]"One million of refuges are hostage to Erdogan"

participating in data annotation. Data were annotated by two young Italian females with a background in linguistics. The limited diversity among annotators may narrow the variety of perspectives included, and their personal biases could influence the data annotation process.

Acknowledgments

Thanks to Dr. Martina Rosola and all the activists of Amnesty Task Force on HS for supporting us in the collection and creation of the AmnestyCounterHS dataset.

References

[1] A. T. E. Capozzi, M. Lai, V. Basile, C. Musto, M. Polignano, F. Poletto, M. Sanguinetti, C. Bosco, V. Patti, G. Ruffo, G. Semeraro, M. Stranisci, Computational linguistics against hate: Hate speech detection and visualization on social media in the "Contro L'Odio" project, in: R. Bernardi, R. Navigli, G. Semeraro (Eds.), Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019, volume 2481 of *CEUR Workshop Proceedings*, CEUR-WS, 2019. URL: http://ceur-ws.org/Vol-2481/paper14.pdf.

[2] H. Bonaldi, Y.-L. Chung, G. Abercrombie, M. Guerini, NLP for counterspeech against hate: A survey and how-to guide, in: K. Duh, H. Gomez, S. Bethard (Eds.), Findings of the Association for Computational Linguistics: NAACL 2024, Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 3480–3499. URL: https://aclanthology.org/2024.findings-naacl.221.

[3] C. Schieb, M. Preuss, Governing hate speech by means of counterspeech on facebook, in: 66th ICA annual conference, at Fukuoka, Japan, 2016, pp. 1–23.

[4] Z. Waseem, T. Davidson, D. Warmsley, I. Weber, Understanding abuse: A typology of abusive language detection subtasks, in: Proceedings of the First Workshop on Abusive Language Online, Association for Computational Linguistics, Vancouver, BC, Canada, 2017, pp. 78–84. URL: https://aclanthology.org/W17-3012.

[5] C. Bosco, F. Dell'Orletta, F. Poletto, M. Sanguinetti, M. Tesconi, Overview of the EVALITA 2018 hate speech detection task, in: T. Caselli, N. Novielli, V. Patti, P. Rosso (Eds.), Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018, volume 2263 of *CEUR Workshop Proceedings*, CEUR-WS, 2018. URL: http://ceur-ws.org/Vol-2263/paper010.pdf.

[6] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Computing Surveys 51 (2018) 85:1–85:30. URL: https://doi.org/10.1145/3232676.

[7] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: A systematic review, Language Resources and Evaluation 55 (2021) 477–523. URL: https://rdcu.be/cCdaB.

[8] T. Caselli, V. Basile, J. Mitrović, I. Kartoziya, M. Granitzer, I feel offended, don't be abusive! Implicit/explicit messages in offensive and abusive language, in: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 6193–6202. URL: https://aclanthology.org/2020.lrec-1.760.

[9] M. Sanguinetti, G. Comandini, E. D. Nuovo, S. Frenda, M. Stranisci, C. Bosco, T. Caselli, V. Patti, I. Russo, Haspeede 2 @ EVALITA2020: Overview of the EVALITA 2020 hate speech detection task, in: V. Basile, D. Croce, M. D. Maro, L. C. Passaro (Eds.), Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020, volume 2765 of *CEUR Workshop Proceedings*, CEUR-WS, 2020. URL: http://ceur-ws.org/Vol-2765/paper162.pdf.

[10] S. Frenda, V. Patti, P. Rosso, When sarcasm hurts: Irony-aware models for abusive language detection, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, Cham, 2023, pp. 34–47.

[11] Y.-L. Chung, S. Sinem Tekiroğlu, S. Tonelli, M. Guerini, Empowering ngos in countering online hate messages, Online Social Networks and Media 24 (2021) 100150. URL: https://www.sciencedirect.com/science/article/pii/S246869642100032X. doi:https://doi.org/10.1016/j.osnem.2021.100150.

[12] Y.-L. Chung, S. S. Tekiroğlu, M. Guerini, Towards knowledge-grounded counter narrative generation for hate speech, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 899–914. URL: https://aclanthology.org/2021.findings-acl.79. doi:10.18653/v1/2021.findings-acl.79.

[13] Y.-L. Chung, E. Kuzmenko, S. S. Tekiroglu,

M. Guerini, CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2819–2829. URL: https://aclanthology.org/P19-1271. doi:10.18653/v1/P19-1271.

[14] P. Goffredo, V. Basile, B. Cepollaro, V. Patti, Counter-TWIT: An Italian corpus for online counterspeech in ecological contexts, in: K. Narang, A. Mostafazadeh Davani, L. Mathias, B. Vidgen, Z. Talat (Eds.), Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), Association for Computational Linguistics, Seattle, Washington (Hybrid), 2022, pp. 57–66. URL: https://aclanthology.org/2022.woah-1.6. doi:10.18653/v1/2022.woah-1.6.

[15] S. Benesch, D. Ruths, K. P. Dillon, H. M. Saleem, L. Wright, Counterspeech on twitter: A field study, Dangerous Speech Project. Available at: https://dangerousspeech.org/counterspeech-ontwitter-a-field- study/. (2016) 1–39.

[16] W.-C. Hwang, S. Goto, The impact of perceived racial discrimination on the mental health of Asian American and Latino college students, Cultural Diversity and Ethnic Minority Psychology 14 (2008) 326–335. URL: https://pubmed.ncbi.nlm.nih.gov/18954168/.

[17] F. A. Sheikha, D. Inkpen, Learning to classify documents according to formal and informal style, Linguistic Issues in Language Technology 8 (2012). URL: https://journals.colorado.edu/index.php/lilt/article/view/1305. doi:10.33011/lilt.v8i.1305.

[18] S. S. Tekiroğlu, Y.-L. Chung, M. Guerini, Generating counter narratives against online hate speech: Data and strategies, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 1177–1190. URL: https://aclanthology.org/2020.acl-main.110. doi:10.18653/v1/2020.acl-main.110.

[19] X. Yu, E. Blanco, L. Hong, Hate speech and counter speech detection: Conversational context does matter, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 5918–5930. URL: https://aclanthology.org/2022.naacl-main.433. doi:10.18653/v1/2022.naacl-main.433.

[20] B. Vidgen, D. Nguyen, H. Margetts, P. Rossini, R. Tromble, Introducing CAD: the contextual abuse dataset, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 2289–2303. URL: https://aclanthology.org/2021.naacl-main.182. doi:10.18653/v1/2021.naacl-main.182.

[21] A. Albanyan, A. Hassan, E. Blanco, Finding authentic counterhate arguments: A case study with public figures, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 13862–13876. URL: https://aclanthology.org/2023.emnlp-main.855. doi:10.18653/v1/2023.emnlp-main.855.

[22] P. Möhle, M. Orlikowski, P. Cimiano, Just collect, don't filter: Noisy labels do not improve counterspeech collection for languages without annotated resources, in: Y.-L. Chung, H. Bonaldi, G. Abercrombie, M. Guerini (Eds.), Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA), Association for Computational Linguistics, Prague, Czechia, 2023, pp. 44–61. URL: https://aclanthology.org/2023.cs4oa-1.4.

A. Dataset details

Table 7 shows annotated examples extracted from our dataset.

B. Strategies of CN

1. **Informative**: the writer writes a statement that seeks to debunk or fact-check the claims made by the attacker. Example: *"Minoranze etniche è un termine usato in un contesto specifico, qui ad esempio, nel Regno Unito, le persone provenienti da questi paesi sono minoranze"*[39]

2. **Alternative**: the writer proposes alternatives to the statement made by the attacker and proposes corrections about some aspects of its content, suggesting a more "correct" point of view and giving

[39] "Ethnic minorities' is a term used in a specific context. For example, here in the United Kingdom, people from these countries are considered minorities."

Layers	Examples
Formal style	Comment: "salve, comprendo la sua polemica, ma non sono arrivati qui per "essere un peso", sono migranti, chi arriva dalla Libia, chi dalla Nigeria, [...]"[29]
Informal style	Comment: "stai tergiversando, situazioni diverse, qui si parla di omosessuali, completamente diverso dai giochi con talco e tutto il resto che hai citato. Ognuno però può fare quello che vuole non sono problemi miei. Ciao buona giornata"[30]
Dangerous speech support	Comment: "avrà tanti morti sulla coscienza, oltre ai nostri anche i migranti, dovete chiudere i porti"[31]
Counterspeech support	News title: "Disney, la carica dei 101 generi: "Entro il 2022 la metà dei personaggi sarà Lgbt"[32] Comment: "idealmente potrebbe essere vero che per una piena inclusione non ci dovrebbe essere bisogno di dare etichette, ma ognuno dovrebbe essere libero di essere chi è e amare chi vuole liberamente. Ma conviene con me che nelle società di [...]"[33]
Explicit dangerous speech	News title: "Il Giornale Pescara, magrebino aggredisce e deruba 63enne fuori dal supermercato"[34] Comment: "Adesso vediamo di dargli anche la medaglia sto disgraziato"[35]
Implicit dangerous speech	Comment: "Il suo desiderio da padre era quello di avere un figlio che giocasse rugby, come tanti che sperano di aver un figlio che sia campione di moto o una figlia che faccia la ballerina."[36]
Informative strategy of CN	Comment: "guardi che gli unici due sbarchi di Marzo sono stati subito controllati e messi in quarantena preventiva, non ci sono stati altri sbarchi tutto il mese, c'è eccome lo spazio per gestire questi pochi arrivati. Prima di accusare il prossimo[...]"[37]
Suggestion strategy of CN	Comment: "Mi perdoni, ma anziché ironizzare sugli altri o sentirsi addirittura più accorti degli altri, perché non cercare di argomentare il proprio pensiero? [...]"[38]

Table 7
Example of the annotation layers

a more detailed description of facts. Example: *"Non gigante buono, ma femminicida"*[40]

3. **Suggestion**: the writer suggests actions to the attacker to encourage them to rethink their views. Example: *"Le consiglio di leggere degli articoli sull'argomento"*[41]

4. **Explicitation**: the writer explicitates/reveals what was implicit in the statement made by the attacker. Example: *"Stanno equiparando la pedofilia all'omosessualità"*[42]

5. **Question**: questions that would challenge the speaker's chain of reasoning and compel them to either answer convincingly or recant their original remark. Example: *"Si potrebbe almeno riportare qualche fatto prima di trarre queste conclusioni?"*[43] Indirect questions should be annotated too. Example: *"mi dia qualche link che riporti esempi concreti di quanto afferma"*[44]

6. **Denouncing and explaining**: when you convey the impression that the opinions put forth by the hate speaker are not acceptable and you try to explain to the user why. Example: *"C'è un grosso errore di fondo in quanto scritto nell'introduzione di questo articolo. Rendere l'interruzione di gravidanza un diritto garantito dall'assistenza sanitaria pubblica non significa che lo Stato imponga alcunché."*[45]

7. **Positive**: a courteous, polite, and civil statement. Example: *"Insegnare ai bambini che ci sono tanti modi differenti per essere felici e che i loro sentimenti valgono è una cosa su cui concordo totalmente."*[46]

8. **Hostile**: the user expresses hostility, aggressiveness towards the initial content, using insults or aggressive words. Example: *"Bisogna davvero essere degli stupidi idioti retrogradi a credere alla negatività sull'Islam."*[47]

9. **Humour**: a strategy of counterspeech with an humoristic, ironic, sarcastic intent whether positive or negative. Example: *"E meno male che era buono. Se era cattivo che faceva, se la magnava?"*[48]

It is possible to identify more than a single counterspeech strategy in a single comment.

[40] "Not a good giant, but a femicide"
[41] "I suggest you to read some papers on the topic"
[42] "They are equating pedophilia with homosexuality"
[43] "Could you at least present some facts before drawing these conclusions?"
[44] "Please provide some links that present concrete examples of what you're claiming"
[45] "There's a big mistake in what's written in the introduction of this article. Making abortion a right guaranteed by public healthcare does not mean that the state is imposing anything."
[46] "Teaching children that there are many different ways to be happy and that their feelings matter is something I completely agree with."
[47] "One must truly be a stupid, backward idiot to believe the negativity about Islam."
[48] "Good thing he was nice. If he had been bad, what would he have done, eat her?"

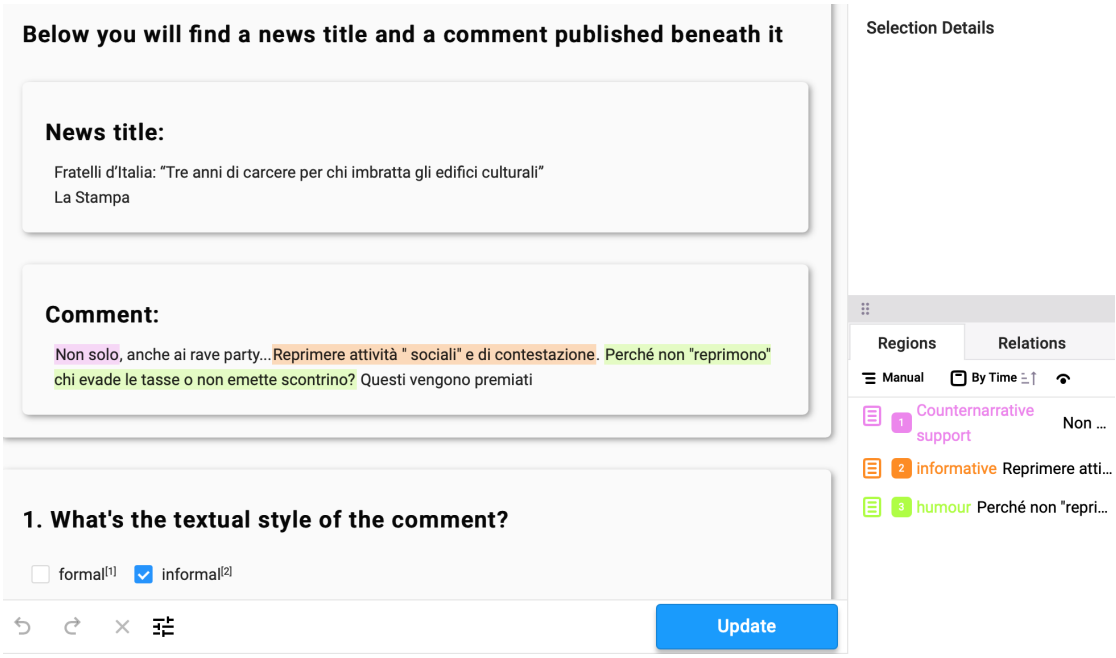


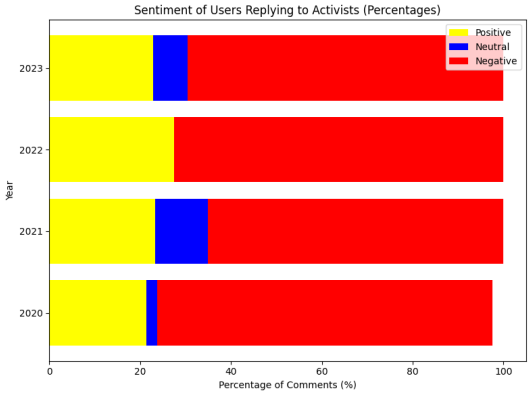
Figure 2: Screenshot of the annotation platform.

C. Annotation Platform

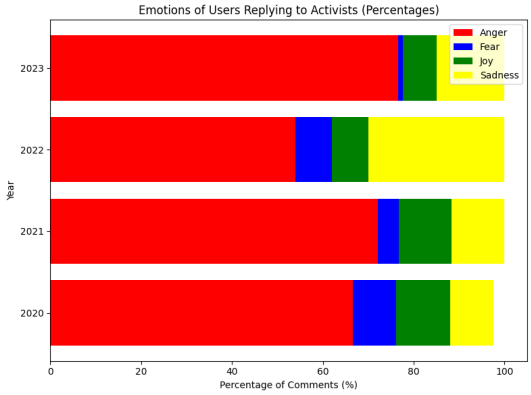
Figure 2 shows the layout of the annotation platform.

D. Affective Analysis AmnestyCounterHS

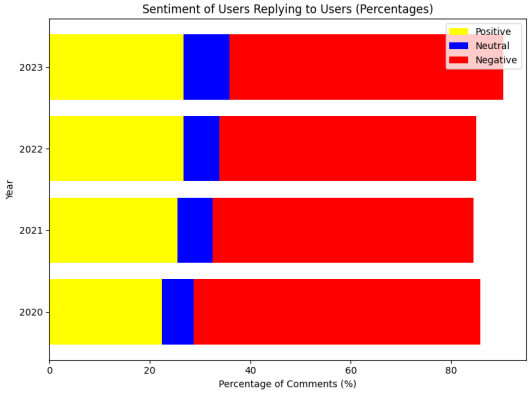
This section presents sentiment and emotion analysis of AmnestyCounterHS for four years: 2020, 2021, 2022, 2023.



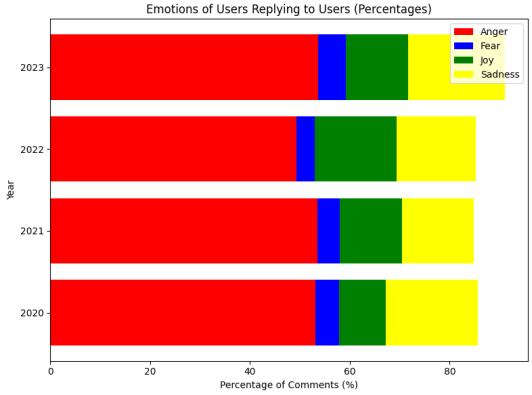
(a) Sentiment distribution of users replying to activists.



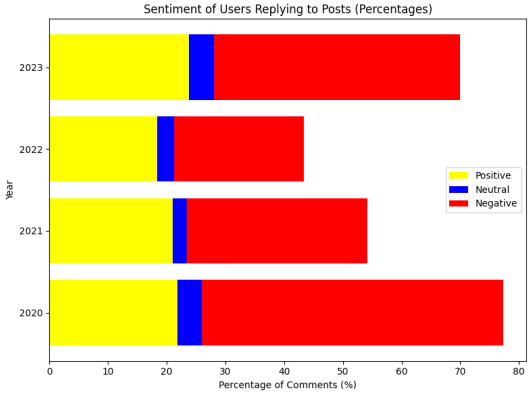
(b) Emotion distribution of users replying to activists.



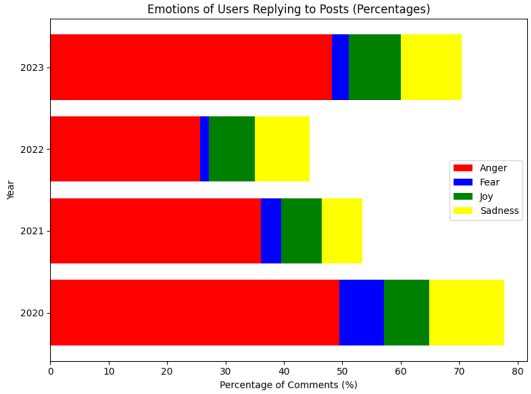
(c) Sentiment distribution of users replying to users.



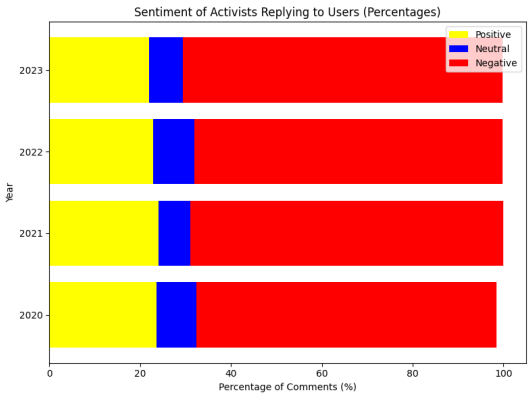
(d) Emotion distribution of users replying to users.



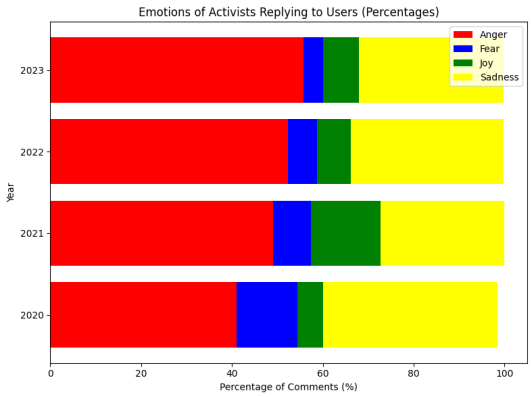
(e) Sentiment distribution of users replying to posts.



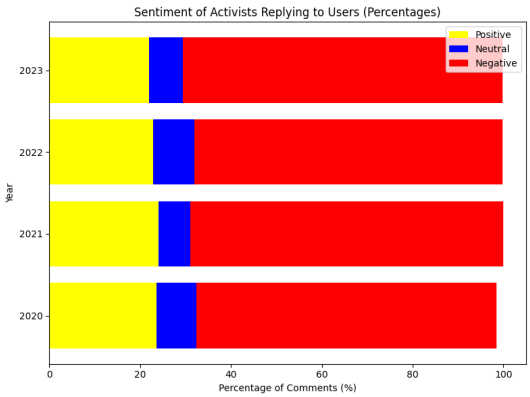
(f) Emotion distribution of users replying to posts.



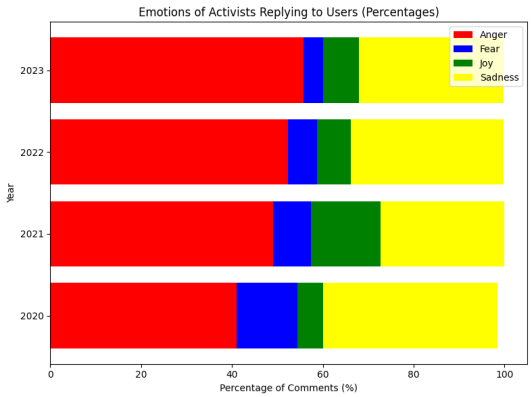
(a) Sentiment distribution of activists replying to users.



(b) Emotion distribution of activists replying to users.



(c) Sentiment distribution of activists replying to posts.



(d) Emotion distribution of activists replying to posts.