Counter-TWIT: An Italian Corpus for Online Counterspeech in Ecological Contexts

Pierpaolo Goffredo[♢] **Valerio Basile**[♡] **Bianca Cepollaro**[♠] **Viviana Patti**[♡]
[♢] Université Côte d'Azur, France / Inria, CNRS, I3S, France
[♠] Università Vita-Salute San Raffaele, Italy
[♡] Dipartimento di Informatica, Università degli Studi di Torino, Italy
[♢]pierpaolo.goffredo@inria.fr, [♡]{name.surname}@unito.it,
[♠]cepollaro.biancamaria@hsr.it

Abstract

This work describes the process of creating a corpus of Twitter conversations annotated for the presence of *counterspeech* in response to toxic speech related to axes of discrimination linked to sexism, racism and homophobia. The main novelty is an annotated dataset comprising relevant tweets in their context of occurrence. The corpus is made up of tweets and responses captured by different profiles replying to discriminatory content or objectionably couched news. An annotation scheme was created to illustrate the relevant dimensions of toxic speech and counterspeech. An analysis of the collected and annotated data and of the Inter-Annotator Agreement (IAA) that emerged during the annotation process is included. Moreover, we report about preliminary experiments on automatic *counterspeech* detection, based on supervised automatic learning models trained on the new dataset. The results highlight the fundamental role played by the context in this detection task, confirming our intuitions about the importance to collect tweets in their context of occurrence.

1 Introduction

Billions of users are active every day on the main social media platforms and they are regularly exposed to toxic discourse, i.e. speech that inflicts psychological or emotional harm and/or incites people to participate in bigoted practices ranging from sexism to homophobia, to racism. To protect users from online toxicity, social media providers have been increasingly implementing censorship-based measures. Such measures are highly controversial and only targeted to the most extreme and explicit forms of toxic speech. Implicit toxic contents are particularly dangerous because they can go under the radar, they are hard to question, and may end up being accepted without conversation participants fully realizing it.

The question arises: how can we counter online toxic speech? Recent studies in social philosophy of language investigated the strategy that consists in engaging in interventions aimed at avoiding that toxic contents get (wittingly or unwittingly) accepted by the conversation participants. Such strategy is often dubbed *counterspeech* and has been mostly analyzed by taking into account face-to-face exchanges. Philosophers of language (Lepoutre, 2017; Langton, 2018) have focused on how counterspeech could work in idealized conversational models. In particular, they have focused on speech that counters implicit toxic contents by (i) spelling out, unpacking, articulating the objectionable contents implicitly conveyed by a given utterance and then (ii) challenging, questioning, rejecting, disputing, confronting it. This counterspeech strategy seems very costly. The first move is cognitively costly: it's hard to unpack implicit content on the spot. The second move is about social cost: it may be tough to go and take a confrontational attitude.

Interestingly, certain features of how communication works on social networks make social media particularly interesting venues to easily observe real instances of counterspeech in ecological contexts. For counterspeech to succeed in face-to-face interactions, the counterspeaker needs to be ready to intervene saying the right thing, in the right place, at the right moment. On social networks, on the other hand, counterspeech can well be asynchronous: this may lighten its cognitive load. As for the social cost of counterspeech, note that social network users enjoy a bit of anonymity in their online intervention and online interactions follow a different etiquette than face-to-face exchanges in terms of interruption of the "conversation". This may possibly lighten the social cost associated with counterspeech. A further interesting aspect is that online counterspeech can reach many more people than offline interventions. In fact, users often challenge offline contents (newspapers articles, pieces of public speeches, reported conversations, passages of textbooks, and so on) on social networks,

Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), pages 57 - 66

in order to give their conversational moves more attention.

Studying counterspeech online comes with the added benefit of enabling the researcher to build computational models of language interactions involving toxic speech and counterspeech. By leveraging the most recent Natural Language Processing techniques, a corpus of counterspeech represents the first step towards automated systems to detect, support or even generate effective responses to toxic speech online.

The exploratory theoretical investigation conducted in philosophy raises many empirical questions. In our work, we address a few ones. For instance: do people on social networks ever employ such an idealized model where in order to reject implicit toxic content one has to first make explicit what was wrong with it? Or do users prefer less sophisticated strategy, like insulting and attacking bigoted contributions? Does the use of irony make the counterspeaker sound more or less aggressive? Do users support counterspeakers with reactions and comments or is it a solitary enterprise? Many more questions are still left unanswered, but this work paves the way for illuminating further the nature and working of online counterspeech.

The contributions of this article can be summarized as follows:

- A novel corpus of toxic speech and counterspeech in a conversational context from Italian social media, covering different target groups.
- A novel annotation schema encoding a fine-grained classification of toxic speech and argumentative relations between utterances.
- A pilot experiment on automatic counterspeech detection, showing the importance of taking the conversational context into account rather than modeling single utterances in isolation.

2 Related Work

There is a growing concern among the ICT (Information and Communication Technologies) companies leading the development of Social Networks about toxic speech: as it can undermine the image of such social environments as "safe" place, they must implement methods to cut off this phenomenon (Mathew et al., 2019). Some countries started to considered hate speech as a crime and sentencing it as such[1]. In other cases, institutions invited the ICT companies to subscribe codes of conduct concerning hate speech moderation and censorship on their platforms. This is the case of the Code of Conduct issued by the EU Commission in 2016 (EU Commission, 2016). Moreover, Social Networks regulated *hateful conduct*, publishing guidelines to avoid harmful behaviors subscribed by users as part of their terms of service[2]. However, such measures don't seem to suffice to effectively combat the phenomenon (Gagliardone, 2015).

Approaches to counterspeech have been investigated by the Computational Linguistics community, suggesting that counterspeech can reduce or limit the hateful content on the Web, especially in Social Networks (Mathew et al., 2018). However, especially from a computational point of view, the development of corpora and models for the automatic detection and generation of counterspeech is still underdeveloped, while most of the efforts have been devoted to the detection of various forms of toxic speech, hate speech included (Poletto et al., 2021; Jurgens et al., 2019).

Most literature focuses on English language and considers toxic speech data collected from specific templates, which limits the coverage of explicit toxic speech and leaves out implicit toxic speech altogether. Chung et al. (2019) recently created a large multilingual corpus of short texts in English, French and Italian, called CONAN, consisting of <hate speech (HS) - counterspeech (CS)> pairs created ad hoc in the context of the HateMeter project[3], with the effort of more than 100 operators from NGOs and with a special focus on Anti-Muslim hatred online in different European countries. Annotated corpora like CONAN enable a systematic study of Counter-Narratives (CNs), a study which is still in its beginnings, but differs from the one we presented here. In particular, counterspeech in CONAN is not observed in an ecological setting, which is the perspective we hold in the current study.

A similar work to Chung et al. (2019) is realized by Chung et al. (2020), where off-the-shelf

[1] https://en.wikipedia.org/wiki/Hate_speech_laws_by_country

[2] Twitter's measures: https://help.twitter.com/it/rules-and-policies/hateful-conduct-policy and Facebook's measure: https://www.facebook.com/communitystandards/hate_speech

[3] http://hatemeter.eu/

NMT models are used to synthesize silver data from other languages using the CONAN dataset as kick-start for generation to overcome the scarcity of gold standard data for training and the lack of huge datasets made of counter narratives in Italian language. The accomplishment is done under different resource conditions, testing the effect of using (i) silver data, (ii) gold standard data, and (iii) their combination. Tekiroğlu et al. (2020) investigate methods to obtain high quality Counter-Narratives while reducing efforts from experts trained by some Non-Governmental Organizations (NGOs) to intervene in online hateful conversations.

Orbach et al. (2020) created benchmark data for training and evaluating the performance of an automatic detection system of counterspeech debates in order to introduce a novel NLU task. Mathew et al. (2019) propose a study to understand how the counterspeech phenomenon is related to statistics of comments collected from YouTube. Menini et al. (2021) present experimental results obtained considering different methods with and without context referring to abusive vs. not abusive tweets.

Unlike the related works presented in this section, the contribution of this work in Automatic Counterspeech Detection is the development of a multi-layer corpus of Italian Twitter data in the context of their conversation thread.

3 The Counter-TWIT corpus

We developed a novel corpus, called **Counter-TWIT**, to study counterspeech online in an ecological setting, based on Twitter conversation threads in the Italian language.

3.1 Collecting Conterspeech Twitter Data

We collected a new dataset of tweets. Counter speech is rare across all of social media, and we considered several strategies for ensuring there were sufficient instances in our dataset.

We chose Twitter as the source platform, in particular collecting tweets and their replies, because of the accessibility of its API.

Collecting counterspeech in an ecological setting is a very challenging task, since there are not obvious keyword-based strategies to filter out the relevant tweets from the ones that are posted everyday and that can be collected by relying on the Twitter API. Let us recall that the creation of the novel corpus was a stage, necessary to the following preliminary experimental phase, where the corpus will be exploited for training a machine learning model able to recognize automatically counterspeech discourse on misogyny, homophobia and racism. We initially selected the profiles of activists, organizations, or pages especially devoted to calling out common instances of bigotry. Users interacting in such contexts are likely to comment on hate speech and thus engage in counterspeech. Such profiles are not as popular as those of public figures such as actresses and politicians. In some cases, however, a few comments are enough to start an interesting conversation thread. In such pages users often highlight how certain news are presented in troublesome ways implicitly conveying discriminatory contents. In addition, these profiles allow their followers to reply in order to share their personal opinion giving rise to counterspeech as a *collective enterprise*, which is an interesting trait.

For collecting data different tools for Python language have been used in favor of rebuilding the conversation tree.

3.2 Data annotation

To annotate the tweets we developed a custom annotation platform. Expert annotators were selected among bachelor's, master's and PhD students and university researchers, within disciplines related to Humanities and Social Sciences such as philosophy and psychology, with some specific background in the study of hate speech and counterspeech.

The annotators were trained in various areas of language sciences, ranging from philosophy of language to computational linguistics. Therefore, they were trained to be sensitive to the relevant distinctions at play in the annotation, e.g., between explicit and implicit communication, irony, and so on. The annotation scheme was applied by seven annotators to a collection of 624 messages, including 344 root tweets and their replies (280 posts). The annotators were provided clear and detailed guidelines[4].

At first, the annotators tested a preliminary version of the platform on a small sample of tweets and replies, sharing comments and discussing doubts and controversial issues that needed explanation or modification. This process led to settling on the final version of the annotation scheme and guidelines.

[4]Guidelines are available at https://github.com/pierpaologoffredo/Counter-TWIT/blob/main/Readme.md (in Italian).

Figure 1: Screenshot of the annotation interface of Counter-TWIT.

The annotation process was based on two layers: firstly, annotators were called to judge whether a tweet or reply could be considered as (Yes/No): TOXIC SPEECH, COUNTERSPEECH, SUPPORT TO COUNTERSPEECH. All of these are binary questions and not mutually exclusive. Figure 1 shows a screenshot of the annotation interface.

In case a tweet or reply is marked as "counterspeech", the annotator is asked to annotate the type of counterspeech and the target group considered (Misogyny, Homophobia, Racism and Other [5]), as a second annotation layer.
Counterspeech often denounces the nature of the discriminatory content it aims to counter. There are several possible labels that can be used for marking different classes of counterspeech, also based on previous studies (Mathew et al., 2019). After a careful discussion and inspired by the reflections in (Cepollaro, 2021), we decided to select four labels associated to the different type of counterspeech: EXPLICITATION, HOSTILITY, IRONY/HUMOR, ALTERNATIVE. In the second-level each label is binary and they are not mutually exclusive, except for hostility that is rated on a scale from 1 to 10. In the following all the layers included in our annotation scheme are described.

Toxic Speech Toxic speech promotes discrimination or deprives people of important powers of self-determination and social and civic participation. Racist, sexist and homophobic slurs count as systemic toxic discourse that generally worsens its targets' well being. Furthermore, note that toxic speech is not about impolite language or vulgar expressions: speech can be toxic and damage people's dignity without employing "bad" words.

Therefore, we call toxic speech the discourse that explicitly or implicitly expresses or promotes unjust discrimination on the basis of gender, ethnicity, geographical origin, sexual orientation, the presence of disabilities, and so on. The **toxic speech** label applies both to explicit and obvious cases, and to implicit and more difficult to grasp cases. What distinguishes toxic speech is that it implicitly or explicitly conveys content that contributes to extant social injustice, e.g., those due to sexism, homophobia, and racism. This could be in principle performed via aggressive as well as non-aggressive speech. Take for instance a scenario where one attacks their interlocutor with a racial insult: this is aggressive toxic speech. Then take a scenario where one claims that the members of a given group should not benefit from certain rights: this is toxic speech too because of its content, but it is not aggressive in the sense of the former. In other words, the feature of aggressiveness or hostility does not primarily concern the content but the form of a contribution. This said, it appears clear how a counterspeech intervention can also display a different degree of aggressiveness or hostility in its form. Counterspeech in general (at least of the kind we considered in this study) is confrontational in character, for it challenges a piece of discriminatory content. But confrontation can be carried out in more or less aggressive ways. What's the difference between toxic speech and counterspeech hostility? Possibly none, but this does not blur the divide between the two notions: while the former conveys discriminatory content, the latter challenges it.

Counterspeech Counterspeech is a second-round speech expressing disagreement with a content or attitude. The type of counterspeech we are

[5] We did not constrain the definition of the main axes of discrimination in place, because we wanted annotators to be aligned with the folk understanding of such notions. We introduced the category "Other" to collect any other targets, with the idea of qualitatively analyzing any choices on this item. The small number of such selections (only 33 within the entire corpus) seems to confirm that the choice of targets was reasonable.

interested in is the one that tries to combat discriminatory or stereotyped contents (e.g., sexist, homophobic, racist, etc.) occurring in another post, comment, newspaper article, song, film, etc. expressed using a toxic language. In our framework, counterspeech is meant to be used to address toxic speech, rather than merely false speech. It is particularly interesting when it is exploited to address *implicit* rather then explicit toxic speech (speech conveying toxic contents via implications, presupposition, and the like): "implicit toxic contents are particularly dangerous: they can go under radar, they are hard to question, and may end up being accepted in the common ground without conversation participants fully realizing it. They may be immune to censorship, slipping through it" (Cepollaro, 2021).

Support to counterspeech Support consists in giving resonance and visibility to a certain counterspeech intervention (inside or outside the Twitter thread), in expressing approval and support for another user's intervention. For example, in this exchange[6]:

> ***-"Miley Cyrus video reveals all the sexualization of lesbians."***
> *-"Quite right!"*

The answer expresses approval and support for the counterspeech intervention, therefore it counts as support for the counterspeech.

Explicitation The explicitation of the implicit meaning unpacks, articulates and brings out what was implicit in a message (Sbisà, 1999). This typology is particularly interesting because discriminatory contents are often conveyed. Social media users sometimes employ explicitation to point out how certain apparently harmless interventions actually communicated discriminatory contents. Explicitation, by articulating what is implicit, opens up the possibility that implicit content will be criticized or questioned.

The practice of explicitation highlights implicitly transmitted information monitors and filters the influence that the implicit meaning can have on. Here is an example of what the practice of explicitation looks like:

> ***-"Emma Watson is beautiful but smart"***
> *-"What does 'but' mean, that a beautiful woman is not smart?!"*

In this case the second speaker challenges the first's assumption that there would be a contrast for a woman between being beautiful and being smart.

Hostility In engaging in counterspeech, users can express various degrees of hostility and antagonism. This is often carried out through (but is not limited to) the use of aggressive and insulting language. For instance:

> *"Good giant? What a bunch of morons"*

The speaker in the example gets angry at the newspaper that called "good giant" a man who murdered a lesbian woman for rejecting him.
To conceptualize and then measure the efficacy of counterspeech is still an open question. Among the most promising candidates, we find its capabilities to change people's minds and raise awareness about discrimination in the toxic speaker and in the audience. It is also an open question what modulates counterspeech efficacy. It may well be that hostility backfires, and that less confrontational counterspeech styles obtain better effects, but it is not said. This could easily depend on the context and the kind of content that counterspeech aims to reject. For this reason, our study is not yet concerned with counterspeech efficacy, but rather on the ways in which it is performed and perceived. A further step in this research is then to conceptualize and measure its efficacy, relying on a classification of its most salient features.

Alternative In engaging in counterspeech, users can propose an alternative to the main topic being discussed: they may for instance object to the way a newspaper title an article and come up with an alternative that in their view would avoid the troublesome contents conveyed by the actual one.

This kind of correcting interventions typically targets the wording of the text or some aspects of its content, suggesting a more "fair" point of view or providing a more detailed description of the facts.

> *The news to report is not that there are baby prostitutes in Parioli, but that there are pedophile customers in Parioli. Stop blaming the victims!*

The speaker in the example suggests that newspaper shouldn't talk about "baby prostitutes" but "pedophile clients" since their way of couching the news implicitly blames victims.

[6]The main tweet is in **bold**, while the reply is in *italic*, the tweets are translated into English by the authors.

Irony/Humor Irony detection consists in reporting if a text contains traces of irony. In this context we call "irony" a plethora of phenomena, such as humor, something witty, black humor, sarcasm, etc.

Irony can be expressed in many ways and there is no single definition of what is ironic and what is not. In this task users are asked, expanding as much as possible the definition of irony, to note as ironic any humorous, sarcastic, ironic intent, be it positive or negative.

> *"And thank goodness he's a good giant. If he was bad that he did, would he eat it?"*

This tweet ironically remarks how ridiculous it is to call "good" someone who murdered a woman for rejecting it. Note that the labels on this layer are not mutually exclusive: more then one typology label could be selected during the annotation.

3.3 Annotation Results

For each tweet, the gold label was obtained by aggregating the results of the individual judgments, by applying simple mathematical operations: majority vote for binary labels and arithmetic mean for labels with numeric values (only *Hostility* in our scheme). Figure 2 shows the distribution of the gold standard labels. 3.04% of tweets were labeled as both Counterspeech and Support, while no overlap was found between Toxic and the other labels.

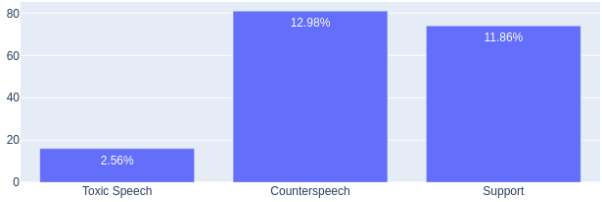

Figure 2: Distribution of the Layer 1 labels over the Counter-TWIT corpus.

The labels are not evenly distributed between tweets and replies. It is possible to observe in Figure 3 that TOXIC SPEECH is more present in replies (3.5%) than in tweets (1.7%), as well as SUPPORT (17.5% in replies and 7.2% in tweets). The opposite is true for the COUNTERSPEECH label, present in 16.2% of the tweets and 8.9% of the replies.

Interestingly, the presence of counterspeech at the root tweet level is significant. This indicates that tweets classified as counterspeech have led users to comment to support counterspeech. These

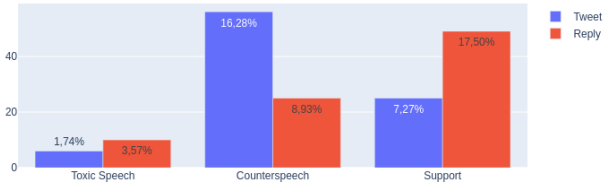

Figure 3: Distribution of Layer 1 labels (root tweets and replies).

first analysis results confirm that collecting data from target profiles is effective for the purpose of filtering samples of counterspeech in the wild, given that the phenomenon is very sparse and a simple keyword-based or hashtag approach is harder to be applied. We can also see that in the debate generated around these profiles there is often an attempt of countering toxic speech generated elsewhere (news, TV, etc). This is interesting because it allows us to analyze the phenomenon of toxic speech in social media (and its reactions) in more comprehensive way such as by investigating cross-references between various media, and framing the overall debate in the context of a media ecosystem. This latter includes social media but also others toxic information sources to be countered. As a consequence, the support label among annotated replies is also significant.

Figure 4 shows the distribution of the gold standard labels for the second level of annotation considering the whole corpus made of 642 tweets.

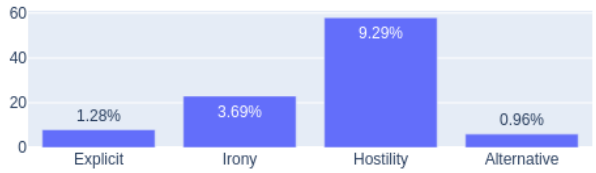

Figure 4: Distribution of the counterspeech typology labels over the Counter-TWIT corpus

Also in this case it is possible to notice that a tweet or a reply can be considered belonging to different type of counterspeech rather than a single one as illustrated in the Figure 5.

Regarding the neutral class, this is represented by all those tweets and replies that are not classified as toxic, counterspeech and support to counterspeech. It includes 472 tweets and replies. This imbalance in the data highlights once again how difficult it can be to collect these types of tweets and replies and subsequently categorize them.

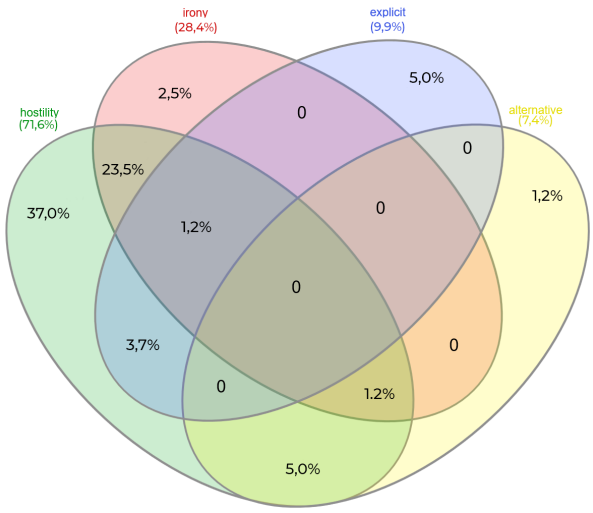

Figure 5: Intersection of counterspeech typology labels over the Counter-TWIT corpus (% refers to the total of tweets annotated as counterspeech).

3.4 Inter-Annotator Agreement

The quality of the gold standard is evaluated in terms of inter-annotator agreement using Krippendorff's α , a generalization of Cohen's Kappa to an arbitrary number of annotators applicable to incomplete question-answer matrices, which was suitable to our case (Artstein and Poesio, 2008). The analysis is limited to the binary labels.

Table 1: Krippendorff's α values for each label on tweets and replies.

Label	α **(tweets)**	α **(replies)**
TOXIC SPEECH	0.25	0.15
COUNTERSPEECH	0.46	0.03
SUPPORT	0.36	0.37
EXPLICITATION	0.38	0.02
IRONY	0.40	0.05
ALTERNATIVE	0.25	0.02

Table 1 shows that the annotation of the replies in particular is controversial and the issue deserves a deeper investigations. One possible reason could be that different annotators interpret the main tweet differently, and then, with a cascade effect, diverge more in assigning the label to the reply tweets. The agreement on the root tweets is, instead, generally higher, in particular on the core label COUNTERSPEECH.

In addition, the label which created disagreement the most has been EXPLICITATION. The annotators reported that during the annotation task it was very difficult to understand when a tweet or a reply could be marked with this tag, which highlighted a difficulty in reaching a common understanding of the meaning of the label. Recent literature postulates how disagreement stems from different sources. We hypothesize that in the case of this work, the disagreement on the main level of annotation (toxic/counterspeech) is dependent on the highly subjective nature of the annotation task. However, the disagreement on the finer-grained level may be due to the more difficult, ambiguous nature of the task, which needs greater knowledge of linguistic phenomena under observation.

Furthermore, a deeper analysis on that tweets (25) and replies (4) which have been considered as counterspeech by all three annotators reveals confusion in agreeing on EXPLICITATION as showed in Table 2.

Table 2: Krippendorff's α values for data considered counterspeech by all three annotators.

explicitation	**irony**	**alternative**
0.09790	0.41364	0.46749

Thus, the label which created a visible disagreement has been the **explicitation**. The annotators reported that during the annotation task it was very difficult to understand when a tweet or a reply could be marked with this tag, which highlighted a difficulty in reaching a common understanding of the meaning of the label.

However, the disagreement on the finer-grained level may be due to the more difficult, ambiguous nature of the task, which needs greater knowledge of linguistic phenomena under observation. The IAA results reflect the problems described.

4 Evaluation

We carried our a battery of experiments in order to perform three independent binary classifications: toxic vs. non-toxic speech, counterspeech vs. not counterspeech, and support to counterspeech vs. not support to counterspeech. We employ a supervised classifier based on BERT (Devlin et al., 2019) pre-trained on a large corpus of Italian tweets named AlBERTo (Polignano et al., 2019).

The metrics used to evaluate AlBERTo's performance are Precision, Recall, and F1-Score for the individual labels, and their macro-average.

The three experiments are 5-fold cross-validation experiments with 9 fine-tuning epochs and a learning rate of 10^{-5}. The results are shown

Table 3: Model performance over three binary classification using reply text as dataset for training. (0), (1), and (avg) refer respectively to positive class, negative class, and their macro-average.

Label	Prec.(0)	Rec.(0)	F1 (0)	Prec.(1)	Rec.(1)	F1 (1)	Prec. (avg)	Rec.(avg)	F1 (avg)
COUNTERSPEECH	.914	.884	.898	.441	.408	.402	.661	.663	.650
TOXIC	.978	.985	.981	.295	.183	.186	.637	.584	.584
SUPPORT	.932	.929	.930	.550	.500	.501	.741	.714	.716

in Table 3. Despite the small size of the corpus and the representative items for each class, the classifiers for COUNTERSPEECH and SUPPORT perform reasonably well, while the classification of TOXIC SPEECH turned out to be a challenge, in particular for detecting the positive class.

The results are obtained with the model fine-tuned only with the tweet or reply text in isolation. We performed an additional experiment taking into account the root of the conversations where the replies belong. We do so by concatenating the text of the reply to the text of the original tweet it replies to, with the goal of observing how the performance of the model changes when considering the context of the reply. The results of this second experiment are shown in Table 4. The experiment is performed with the same hyperparameters of the previous experiment, in order to provide a consistent comparison.

Including context in the training improves the classification of counterspeech. This is due mainly to a higher recall on the positive class. This is true for all labels, and particularly for COUNTERSPEECH, which is about 65% higher. However, the extra training data seem to confuse the classifiers for the other two labels.

5 Error Analysis

In order to get some deeper insight about the difficulties in classifying a **counterspeech** content, we selected False Positives (FP), i.e., counterspeech tweets hat have not been classified as such by the model, and exploited the information included in the finer-grained annotation layer regarding counterspeech categories, namely EXPLICITATION, HOSTILITY, IRONY/HUMOR, ALTERNATIVE.

We considered all the FPs for the first annotation layer, counting all the data (tweets or reply) that were labeled as belonging to the counterspeech category from humans but not from the model. Thus, for those tweets we checked the values attached to the counterspeech typology labels in order to find a meaning among the classification errors and the counterspeech typologies' relation.

The proportion of False Positives over all the predictions obtained from the language model is the following: false positives represent about 7% of the total. Of these, the vast majority are *Ironic* (∼34%) and *Hostile* (∼76%), also considering that the labels are not mutually exclusive.

This qualitative analysis can lead to affirm that the model tends to confuse **hostile** and **ironic** content more than explicit and suggestion of alternative ones probably due to a higher cost from a cognitive and social point of view.

There are two layers of complexity that give rise to disagreement in classifying correctly the tweets. Detecting toxic speech depends on how each subject is sensitive to detecting each axis of discrimination (which often varies along demographic and psychological factors). A further source of disagreement stems from the relative unconstrained character of the notions deployed (toxic speech and counterspeech) (Basile et al., 2021).

Finally, we analyzed the False Positive Rate by counterspeech category. **Irony** and **Hostility** are by far the most difficult categories to predict, with a FP ratio of about 60% and 70% respectively, while next to no FPs are predicted for *Explicit* and *Alternative*.

6 Discussion and Conclusions

In this work we studied hate speech in online environments. To address the dangers of toxic speech, Social Networks defined policies that regulate speech inciting hatred, while some countries started to introduce norms to treat this phenomenon as a crime and sentenced as such. This way to address the problem showed some limitations as the main approaches consist in blocking or suspending the problematic content or the user account itself. Therefore several involved parties, such as institutions and organizations, started to consider counterspeech as an alternative to blocking (Gagliardone, 2015). Thus, adding "more speech" has been considered as a valid alternative to counter hate speech.

We collected and annotated data from Twitter in

Table 4: Model performance over three binary classification using reply text and root tweet for training. (0), (1), and (avg) refer respectively to positive class, negative class, and their macro-average.

Label	Prec.(0)	Rec.(0)	F1 (0)	Prec.(1)	Rec.(1)	F1 (1)	Prec. (avg)	Rec.(avg)	F1 (avg)
COUNTERSPEECH	.960	.883	.920	.466	.730	.564	.713	.807	.742
TOXIC	.979	.840	.903	.037	.283	.065	.508	.561	.484
SUPPORT	.922	.816	.865	.317	.544	.396	.620	.680	.630

order to create the Counter-TWIT Italian corpus to study counterspeech in an ecological setting. The corpus includes content that is considered to unleash hate speech and to receive replies in the form of counterspeech.

Specifically, data were collected with the aim of observing counterspeech within the context of occurrence, i.e. collecting not only tweets in isolation, but conversation threads including a root tweet and the corresponding replies. Finally, we validated the corpus with cross-validation experiments.

We developed the **Counter-TWIT** corpus made of tweets and replies collected from accounts that has been selected after a deep research based on shared contents. All the data collected have been annotated, by exploiting a web-based annotation platform developed roughly from the scratch and published online[7], where a group of expert annotators were applying a novel multi-layer annotation scheme devoted to mark whether the tweets or replies were counterspeech, toxic speech or in support of counterspeech (layer 1). In case counterspeech was marked as present, users were asked to label the text as belonging to some typology of counterspeech for the sake of a deeper understanding of the phenomenon (Layer 2).

Thus, the annotated corpus has been used for training the **AlBERTo** neural language model for performing a battery of binary classification task related to the detection of toxic, counterspeech, and support to counterspeech. We used this language model since it has been trained and developed using an Italian vocabulary instead of using other multilingual model that presented limitations to the type of language learned and the size of vocabulary (Polignano et al., 2019).

We executed two type of experiments: one using only the replies of conversation tree and the second with also the "main" tweet. This approach has been designed in order to go deep into the intuition that this classification task needs the context. Results show that performance, Recall in particular, improves when conversation context data are provided, and this supports the original hypothesis that counterspeech must be studied in a context, which is intuitive given the definition of counterspeech as second-turn intervention aimed to contrast a previous contribution (Cepollaro, 2021), taken as reference definition in this work.

Finally, we performed a statistical and qualitative evaluation of the results obtained from the neural language model evaluating the number of data classified as not belonging to counterspeech class rather than being considered as such (False Positives data). We discovered that the model tends to confuse most with Irony and Hostility labels rather than Explicitation and suggestion to Alternative ones.

Given the promising preliminary results, we plan to expand the corpus in our future research. Furthermore, other qualitative analysis could be run by considering the correlation of types of counterspeech and the predictions made with a language model in order to understand in greater detail how the model behaves towards a specific counterspeech category. Indeed, annotating content as counterspeech is not an easy task, due to different shapes of the textual meaning based on the context and the language used. There is not a unique pattern to individuate and mark the tweet as belonging to a specific categories. A large annotated corpus will provide a more solid base for training the model in detecting counterspeech and, in possible future developments, for generating automatically counterspeech content in order to fight hate speech, which is another very interesting direction (Tekiroğlu et al., 2020).

Counter-TWIT[8] is made available online to further study the phenomenon described and other issue related to counterspeech classification in Italian Twitter.

[7] http://thesiscounterspeech.altervista.org/

[8] https://github.com/pierpaologoffredo/Counter-TWIT

References

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

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Bianca Cepollaro. 2021. Remedies to discriminatory contents: On and offline counterspeech. Talk at HaLO Workshop.

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

Yi-Ling Chung, Serra Sinem Tekiroglu, and Marco Guerini. 2020. Italian counter narrative generation to fight online hate speech. In *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021*, volume 2769 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Jacob Devlin, 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 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

EU Commission. 2016. Code of conduct on countering illegal hate speech online.

Iginio Gagliardone. 2015. *Countering Online Hate Speech - UNESCO*. UNESCO Publishing.

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

R. Langton. 2018. *Blocking as counter-speech*, pages 144–164. Oxford Scholarship Online.

Maxime Lepoutre. 2017. Hate speech in public discourse: A pessimistic defense of counterspeech. *Social Theory and Practice*, 43(4):851–883.

Binny Mathew, Navish Kumar, Ravina, Pawan Goyal, and Animesh Mukherjee. 2018. Analyzing the hate and counter speech accounts on twitter. *CoRR*, abs/1812.02712.

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):369–380.

Stefano Menini, Alessio Palmero Aprosio, and Sara Tonelli. 2021. Abuse is contextual, what about nlp? the role of context in abusive language annotation and detection. *CoRR*, abs/2103.14916.

Matan Orbach, Yonatan Bilu, Assaf Toledo, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2020. Out of the echo chamber: Detecting countering debate speeches. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7073–7086, Online. Association for Computational Linguistics.

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

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, volume 2481. CEUR.

Marina Sbisà. 1999. Ideology and the persuasive use of presupposition. *Language and ideology. Selected papers from the 6th International Pragmatics Conference*, 1:492–509.

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. Association for Computational Linguistics.