Mining Annotator Perspectives from Hate Speech Corpora

[♥]Michael Fell, [♣]Sohail Akhtar, and [♦]Valerio Basile

[♥♣♦]Università degli Studi di Torino, Turin, Italy
[♥]mic.fell@gmail.com, [♣♦]{firstname.lastname}@unito.it

Abstract. Disagreement in annotation, traditionally treated mostly as noise, is now more and more often considered as a source of valuable information instead. We investigate a particular form of disagreement, occurring when the focus of an annotated dataset is a subjective and controversial phenomenon, therefore inducing a certain degree of polarization among the annotators' judgments. We argue that the polarization is indicative of the conflicting perspectives held by different annotator groups, and propose a quantitative method to model this phenomenon. Moreover, we introduce a method to automatically identify shared perspectives stemming from a common background. We test our method on several corpora in English and Italian, manually annotated according to their hate speech content, validating prior knowledge about the groups of annotators, when available, and discovering characteristic traits among annotators with unknown background. We found several precisely defined perspectives, described in terms of increased sensitivity towards textual content expressing attitudes such as xenophobia, islamophobia, and homophobia.

Keywords: Linguistic Annotation · Perspective Identification · Annotator Bias · Hate Speech · Polarization of Opinions

1 Introduction

Most modern approaches to Natural Language Processing (NLP) tasks rely on supervised Machine Learning. This is true, among other tasks, for text classification tasks such as abusive language and hate speech detection [35,6]. However, while bias in datasets has been investigated [33,27], the bias in the annotation of the datasets used for training hate speech models is relatively less studied.

Recent works highlight the importance of a "perspectivist turn", i.e., a change of paradigm in supervised machine learning moving away from datasets aggregated by majority vote, and towards frameworks that consider multiple annotator perspectives in data creation, model training, and evaluation [7]. Taking an inclusive stance towards disagreement in data annotation does not only have ethical implications, but rather it has practical impact on the performance of predictive systems [30] and on the reliability of the evaluation [8].

We focus on hate speech (HS), and in general abusive phenomena in online verbal communication, for several reasons. Firstly, hateful discourse online is growing at a worrying rate [36], and it is linked to an increase of violence and hatred towards vulnerable communities, with strong negative social impact [14,21,22]. Moreover, hate speech is a highly *subjective* phenomenon. While no phenomenon is neither totally subjective nor totally objective, the position of hate speech on a hypothetical *inter-subjectivity* spectrum [18] is far from the center, as its judgment is influenced by factors such as socioeconomic background, ethnicity, gender, among others [31]. Moreover, the hatred is typically directed towards targets carrying specific socio-economic, cultural, or demographic traits, which are likely aligned to the factors influencing the judgments of hateful messages by human annotators. Indeed, messages containing hateful content are often *controversial*, that is, they reference events, people, and issues that prompt very different reactions depending on the recipient of the message [26].

In the area of hate speech detection, Akhtar et al. [1] introduced a quantitative measure of the polarization of the annotation induced by the controversiality of the messages. They show how in presence of highly subjective phenomena like hate speech, systematic patterns emerge that suggest a diversification of the annotators' perspectives beyond the mere disagreement. In a follow-up work, Akhtar et al. [2] leveraged the polarization in the annotation to create multiple perspective-encoding classifiers, boosting the classification performance in the process. In this work, we further explore the polarization of annotation, and particularly at providing a methodology to qualitatively study emerging groups of annotators holding different, and sometimes conflicting, perspectives.

More specifically, in this work we deal with **shared perspectives**, that is, the set of factors that cause a certain annotation by a group of human annotators (each holding an **individual perspective**). By analyzing the annotation with computational methods, we aim at i) distinguishing groups of annotators holding different shared perspectives, and ii) identifying the nature of the shared perspectives, providing a human-readable description. First, we provide a formal definition of *perspective* in the context of the annotation of NLP datasets, hinging on the difference between *label agreement* and the novel concept of *feature agreement* We then empirically demonstrate the emergence of perspectives in real datasets of hate speech, computed with a straightforward yet effective procedure, and illustrated in the form of important words and selected examples.

2 Related Work

Most research related to the identification of perspectives mainly focuses on the perspective of the author of the messages. The literature is typically concerned with subjective phenomena in natural language, such as abusive language, where an abundance of expressions of emotions, opinions and sentiments is found. Subjective language is considered a catalyst for multiple perspectives [32,28] and varying opinions at sentence level [34]. Political discourse analysis is an important research area and many researchers worked in identifying different perspectives on political topics including election campaigns as a qualitative analysis task [24]. Lin et al. [17] automatically identified perspectives at the document and sentence level with high accuracy by developing statistical models and learning algorithms on articles about the Israeli-Palestinian conflict.

In NLP, the task of *stance detection* [20] aims at identifying points of view, judgments or opinions on a given topic of interest in natural language. The social and political issues on which individuals tend to express their opinions are usually controversial in nature, causing polarization among people [4]. Beigman-Klebanov et al. [10] worked on perspective identification in public stance on controversial topics such as abortion.

Highly Controversial topics, such as hate speech, are a rich source to identify and analyze conflicting perspectives in online environments. When social media users express different opinions on topics or social issues, the text depicts high level of controversy due to varying perspectives [26]. When such phenomena are manually annotated by human judges, high controversy is bound to have an impact on such annotations, in terms of agreement between the human judges.

In the aforementioned work, Akhtar et al. [1] developed a novel method to measure the level of polarization in conflicting annotations on social media corpora. The authors developed a quantitative index, called *polarization index*, to measure the level at which polarized opinions appear in individual messages. The authors extended their work [2] by developing perspective-aware models based on automatically clustered groups of annotators. State-of-the-art machine learning models are trained on gold standard training sets based on this division, successfully picking up the divergence of opinions in group-based test data. The same authors recently developed a novel multi-perspective abusive language dataset [3] on Brexit to identify and model perspectives expressed by annotators with different ethnic, cultural and demographic background. In contrast to traditionally published NLP corpora, this dataset provides a natural grouping of the annotators into groups of similar backgrounds.

It is noteworthy that disagreement in annotation is a relevant topic also in more *objective* tasks such as POS tagging [8] and even outside the scope of NLP; for instance, Basile et al. [7] describe the high disagreement in the annotation of medical images by experts.

3 Mining Perspectives in Annotations

We postulate that annotators and their individual perspectives influence how they annotate different items related to a given topic. This is particularly relevant to annotation tasks that exhibit a high degree of *subjectivity* as here the influence of the perspective on the ratings may be higher. According to a common definition, a judgment is considered *subjective* when it is mainly “based on, or influenced by, personal feelings, tastes, or opinions.”; we usually contrast this concept with that of *objective*, a term that characterizes judgments that, ideally, are not influenced by personal feelings or idiosyncrasies and which, on a practical level, the vast majority of people would see and label in the same way. For instance, in hate speech detection, different annotators have been shown to diverge highly in their ratings and are polarized [5]. Furthermore, the offensiveness of words depends on the context in which the words are uttered [23]. For example, consider the difference between the use of the word "nigga" in a Rap song, where it is considered as lowly offensive, as opposed to using such words in a political discourse, where it is understood as highly offensive. We assume that annotators implicitly or explicitly take perspectives on topics, and we model this as described in the following.

In order to mine shared annotator perspectives in a given dataset, we postulate a two-step procedure. First, we detect perspectives that are shared among annotators. To this end, we measure how much the annotators agree on item labels, the *label agreement*. Second, we measure to what extent annotators agree on the importance of linguistic features of the items, the *feature agreement*. Combined, our method ensures that annotators in the same shared perspective label items similarly and do so for similar reasons. In this work, we only use *unigrams* as linguistic features to allow simpler explanations. For instance, annotators holding the perspective that the word "fag" is especially hateful, tend to always label items containing this word as *hate speech*, i.e., they exhibit a high feature agreement on this unigram.[1] We finally perform analysis on shared perspectives consisting of annotators that are similar both in label agreement and feature agreement. Such annotators tend to agree both on their item labels and the importances they give to the item features (unigrams).

Individual Perspectives Given a list of items, an annotator A judges these items, according to their opinion on each of them. We call this labeling the *individual perspective* of annotator A on the items. Formally, given n items, assume there are possible opinions $0, 1, ..., c$ for each item. Then, an annotator A takes a perspective p_A by holding an opinion on each item. We call $p_A \in \{0, 1, ..., c\}^n$ the perspective of A (on the items). By modelling annotator perspectives as vectors, we can compare them quantitatively.

In order to identify perspectives in annotations, we require items to have disagreeing annotations. This is only possible in the case where annotations have not previously been aggregated into a single label, what in [11] has been called *diamond standard*. This is in contrast with the usual *gold standard* paradigm where multiple annotations are harmonized into one gold label, often implemented by majority voting. Under the paradigm of annotator perspectives that we have introduced above, the reduction of multiple labels (annotator opinions) into a gold label by majority vote is equivalent to taking the *majority perspective*.

Shared Perspectives While each annotator takes their own perspective, we are more interested in finding perspectives which are shared among annotators. We

[1] Note that our method is agnostic to the type of features extracted from the messages, and it could therefore be used in conjunction with other, more refined features.

call perspectives p_A, p_B shared based on their similarity. We employ clustering to find clusters of annotators that share perspectives. While shared perspectives arise from an agreement of annotators on item labels (label agreement), we also aim to understand how shared perspectives are linguistically defined. To this end, we analyze the importance that different annotators give to different linguistics features, i.e. which words annotators in a shared perspective agree to be important (feature agreement).

3.1 Label Agreement

We measure label agreement in terms of inter-annotator agreement. We use Krippendorff's alpha reliability [15] and cluster the annotators based on the label agreement. We proceed as follows:

- Given n annotators that label the same k items, we obtain a label matrix $V \in \mathbb{R}^{n \times k}$, where $V_{i,j}$ is the rating of annotator i of item j. We compute the similarity matrix $A \in \mathbb{R}^{n \times n}$, where $A_{i,j} = \alpha(V_{i,:}, V_{j,:})$ that encodes the pairwise agreement between annotators i, j, where α is Krippendorff's alpha reliability, and $V_{i,:}$ is the label vector of annotator i. Then, the distance matrix $D = 1 - A$ induces a clustering of the annotators. The distances in D are the pairwise disagreements between annotators.
- We use an off-the-shelf clustering algorithm to cluster the annotators based on their distances D to one another, into groups of annotators with high intra-group label agreement and low inter-group label agreement.
- A high label agreement $\alpha(i, j)$ indicates that annotators i, j tend to give similar labels on the items (texts).

Note that Krippendorff's alpha is also defined for incomplete annotation, i.e., where not all annotators covered all the instances. This is a typical scenario in crowdsourcing, but could happen with other annotation procedures as well.

3.2 Feature Agreement

We want to measure whether annotators agree on the importance of linguistic features of the textual items. In this paper, we use a simple *bag of words* (BOW) to model the texts; the features are unigram counts. Feature agreement between annotators i, j arises when i and j give similar *importance* to features. We measure the importance of each feature to an annotator by computing the chi-square (χ^2) statistics between the feature distribution and the label distribution in the annotator, following a univariate feature selection approach. This measures how the annotator's label depends on the presence of a word in an item. For instance, the presence of the unigram *bitch* often coincides with the label *hate speech*, while this is not the case for the word *sunny*. The χ^2 statistics captures this; it is much higher for *bitch* than it is for *sunny*. When annotators tend to agree on the importance of words, they exhibit an overall high feature agreement. Specifically, we compute feature agreement as follows:

- We extract k features from the corpus. Since we employ the BOW model, given a corpus of n documents, we compute the term-document matrix $F \in \mathbb{N}^{n \times k}$, where $F_{i,j}$ indicates the count of word j in document i. The columns of F are the features, as $F_{:,i}$ is the word counts of word i over the documents.
- For each feature $f_i = F_{:,i}$ and annotator r, we compute the importance imp of feature f_i to annotator r as $imp(f_i, r) = \chi^2(f_i, V_{r,:}^T)$, where V is the previously introduced label matrix.
- We define the feature agreement between annotators i, j by comparing their feature importances. To this end, let $I \in \mathbb{R}^{k \times n}$ be the importance matrix, where $I_{i,j} = imp(f_i, j)$. Then, the vector of all importances of annotator j is given by $I_{:,j}$. The feature agreement β between annotators i, j is then computed by the cosine similarity of their importances vectors: $\beta(i,j) = cosine(I_{:,i}, I_{:,j})$, where $cosine(x,y) = x \cdot y \cdot (\|x\| \cdot \|y\|)^{-1}$.
- A high feature agreement $\beta(i,j)$ indicates that annotators i, j tend to give similar labels when similar words are present.
- Given n annotators, we compute the similarity matrix $B_{i,j} = \beta(i,j)$ that encodes the pairwise feature agreement between annotators i, j. Analogously to the label agreement case, we use the distance matrix $D = 1 - B$ to cluster the annotators into groups of annotators with high intra-group feature agreement and low inter-group feature agreement.

Since the χ^2 statistics requires a dense label matrix, if an annotator has not labelled an item, we insert the negative label (i.e., *not hate speech*). Truly unimportant words then correctly get low importance, while truly important words get assigned a somewhat diminished importance.

3.3 Label Feature Agreement

We consider two different ways of clustering the annotators: by label agreement and by feature agreement. These two clusterings sometimes differ, for instance, when two annotators agree on the item labels (label agreement), but do not agree on the importance of words related to those labels (feature agreement). Since our goal is to find annotators that label similarly and do so for similar reasons, we analyze all annotators that cluster in the exact same way in both labels and features. Specifically, we perform two clusterings for the annotators a_i. First, we cluster the a_i into k different clusters {1, 2, ..., k} according to label agreement, assigning each a_i a cluster $Lab(a_i) \in \{1, 2, ..., k\}$. Analogously, each a_i is assigned a cluster $Feat(a_i) \in \{1, 2, ..., k\}$ according to feature agreement. Then, we only consider such annotators a_i that cluster in the same way, i.e., label feature agreement is defined as $\{a_i : Lab(a_i) = Feat(a_i)\}$.

3.4 Cluster Analysis

Given the clusters of annotators we obtain, we analyze certain cluster properties statistics, how the clusters differ and which words are important to each cluster. Specifically, we perform the following analyses.

Quantitative cluster description: the number of annotators in the cluster, the positive label rate %, the label agreement α, the number of features, and the feature agreement β. We compare the cluster numbers also with the numbers for all annotators disregard of their cluster affiliation.

Qualitative cluster description: we inspect the most characteristic unigrams for the clusters, i.e. the words with the highest *relative importance* R to the cluster. We measure $R_C(w)$ of a word w to a cluster C as $R_C(w) = \frac{1+med\{imp(w,i)\}_{i \in C}}{1+med\{imp(w,i)\}_{i \in \neg C}}$, i.e., the median importance to all annotators inside the cluster vs. the median importance to all annotators outside the cluster. We inspect examples that are polarized between the clusters, i.e., they are annotated with disagreement between the clusters. These examples often carry *important words* as vocabulary.

4 Datasets

The experiments described in this paper are conducted on several hate speech corpora, consisting of Twitter messages (tweets); they are published in various research studies on hate speech. This section provides details about the datasets, such as the annotation process with scheme and guidelines, and information on the annotators.

4.1 HS Dataset on Brexit

The hate speech dataset on Brexit was recently published [3]. Originally, the authors gathered the data from a study on stance detection in political debates [16] where around 5 millions tweets were collected during the Brexit voting period, June 2016. The authors developed a multi-perspective dataset to automatically detect abusive language on social media with the intention to model annotator perspectives and polarized opinions. The collected tweets are filtered with a list of selected abusive keywords based on a previous study [19]. 1,120 tweets were randomly sampled and annotated for hate speech, Aggressiveness, Offensiveness and Stereotype, following the scheme described in [25,29]. In total, six annotators contributed to the dataset. Three of the annotators were researchers with western background and experience in linguistic annotation who volunteered to annotate the data. The other three volunteers were first- or second-generation immigrants and migrants as students from the developing countries to Europe and the UK, of Muslim background. The group of migrants is named *Target* and the locals are named *Control*. The dataset is unique in the way that it involves migrants as the victims of abuse on social media. Personal details of all annotators such as cultural and demographic background and ethnicity are known and considered a valuable source of information for perspective-aware abusive language detection. For the current study, we only used the hate speech label.

4.2 HS Dataset in Italian

The hate speech dataset in Italian language [13] (HS Italian) consists of 3,200 tweets collected from TWITA [9] in 2017, partially overlapping with the Italian

Hate Speech Corpus [29]. The collection was filtered by the authors of the original dataset with a list of handcrafted keywords related to migrants and ethnic and religious minorities in Italy. The tweets were annotated on the Figure Eight platform[2]. A minimum of three annotators annotated the whole corpus, subsequently aggregated by the crowd-sourcing platform to create a gold standard dataset. We requested and obtained the dataset from the authors.

4.3 HS Dataset in English

Davidson et al. [12] developed a hate speech dataset to perform automatic hate speech detection as a multi-classification task. The authors gathered around 85.4 million tweets from a total of 33,458 Twitter users. A hate speech lexicon containing hateful words and phrases was used to query the tweets. This lexicon was compiled by Hatebase.org and the hateful words in the lexicon were identified by internet users. The authors randomly selected about 27,000 tweets from the dataset by using the keywords from the hate speech lexicon. CrowdFlower (now Appen) workers were hired to manually annotate the tweets. The annotation scheme comprises the labels *hate speech*, *offensive but not hate speech*, and *neither offensive nor hate speech*. The authors developed detailed guidelines with their own definitions of different hate speech terms including the context in which the words were used. Each tweet in the dataset was annotated by three or more annotators. The Davidson dataset is only available for download in an aggregated gold standard form[3], therefore we requested and obtained the non-aggregated dataset from the authors.[4]

5 Perspective Mining Experiments

We performed analysis on all the datasets that we introduced in the previous section. An important factor for our experiments is what prior knowledge we have about the annotators that annotated the datasets. Where such background information is given, we can confirm or reject our findings by comparing our empirically found annotator clusters (shared perspectives) with groupings of human annotators. As stated in the dataset section before, we have the following information on the dataset annotators. On the Brexit dataset: the personal details of all annotators such as cultural and demographic background and ethnicity are known. On all other datasets: no background information on the annotators is available. Note that the HS Italian and Davidson datasets are sparsely annotated, as annotators have only labeled a fraction of the instances. This is in opposition to the Brexit dataset which has a dense annotation matrix.

[2] https://www.figure-eight.com/, now Appen.

[3] https://github.com/t-davidson/hate-speech-and-offensive-language

[4] As we needed non-aggregated data for our work, we only found aggregated gold standard data on author's GitHub repository. Therefore, we requested the authors to provide us with pre-aggregated data and we are grateful to them for providing us the required format of the dataset.

	Brexit			HS Italian			Davidson			
Cluster	A	all	B	A	all	B		A	all	B
Annotators	3	6	2	7	14	7		45	111	41
Pos. labels %	**20.5**	12.9	**5.8**	30.9	26.6	22.3	(off.)	77.2	71.2	65.6
							(HS)	**4.1**	10.0	**15.4**
Label agr. α	**0.58**	0.35	**0.44**	**0.03**	0.23	**0.42**		0.64	0.58	0.60
Features		266			623				2366	
Feature agr. β	**0.86**	0.7	**0.86**	0.34	0.35	**0.48**		0.22	0.29	**0.46**

Table 1. Quantitative cluster statistics for different datasets. Note that in the Davidson dataset there are two kinds of positive labels, one for "offensive language" content (off.) and one for "hate speech" (HS, stronger label)

Table 1 gives an overview of the quantitative statistics and differences between the clusters. As for the qualitative analysis, we provide examples where important words are shown in context. For all clustering experiments, we used the experimental setup as described below.

5.1 Experimental Setup

- Preprocessing: we removed URLs and Twitter handles (@username) from the tweets, tokenized them using the NLTK[5] Tweet Tokenizer and lemmatized them using spaCy[6].
- BOW features: we created the BOW feature space with the scikit-learn[7] CountVectorizer, where we set the minimum document frequency to 10. This number was decided based on the fact that some tweets occured duplicate or near-duplicate, because of the dialog structure of Twitter, users will cite each other. We alleviate the problem by setting a rather high minimum document frequency of 10. Furthermore, we counted each word once per document ("binary") and extracted solely unigrams.
- Clustering: we used the KMeans algorithm with different numbers k of clusters, we settled to $k = 2$ which appeared most reasonable based on the inspection of 2D-PCA embeddings of the datasets. This parameter choice makes us conform with the polarization paradigm, i.e. we analyze two conflicting/polarized perspectives.
- Important words: for each cluster, the top 20 words with highest relative importance are considered. The polarized examples are extracted using the polarization index method [1].

5.2 Perspectives in the Brexit Corpus

In the Brexit datasets, the inter-annotator agreement is measured as $\alpha = 0.35$. The positive label rate is 12.9%. We extract 266 features from the corpus and obtain $\beta = 0.7$ as feature agreement between all annotators.

[5] https://www.nltk.org
[6] https://spacy.io
[7] https://scikit-learn.org

After label feature agreement (see Section 3.3), we obtain the clusters $A = \{3, 4, 5\}$ and $B = \{1, 2\}$. Since we know the annotator backgrounds, we know that A corresponds to the migrants with muslim background (*target group*) and that B corresponds to the non-migrants (*control group*). This result effectively validates our clustering methodology based on label and feature agreement to extract perspectives empirically.

Quantitatively, we find the following differences between the clusters: i) The positive label rate is much higher in A (20.5%) as compared to positive labels in B (5.8%), indicating the annotators in A are more sensitive in this task (all annotators 12.9%). ii) The label agreement is higher in cluster A ($\alpha_A = 0.58$) as compared to $\alpha_B = 0.44$, indicating that cluster A holds more coherent opinions as B. Both values are much higher than the average, meaning that the groups hold polarized opinions. iii) The feature agreement is higher in both clusters ($\beta_A = 0.86 = \beta_B$) compared to the dataset feature agreement ($\beta = 0.7$), indicating polarization of the feature agreements of the clusters as well.

Qualitatively, we find that certain words are highly correlated with the positive label in both groups, and some words are specific to the annotator clusters. The shared vocabulary contains words such as “islam”, “kill” and hashtags related to US president Donald Trump (#maga, #trump2016). When inspecting the corpus, we find examples such as the following that exemplify the use of the words; matched words are **bold**. And indeed, in this example, both annotator groups give the positive label.

> 🐦 *RT @*[redacted]*: The U.K. Must ban* **Islam** *and close all mosques! URL*

> 🐦 *London should kick all Muslim Refugees out before they all* **kill** *them.* **#Trump2016** *URL*

From these shared vocabulary examples, we can see that since “islam” is one of the important words, the hate speech in this corpus appears to be at least partially islamophobia. An inspection of the important words for the potential target group of islamophobia, cluster A , supports this claim. We find a specific and distinctive vocabulary related to muslims, invasion, terrorists.[8] The following examples illustrate the important words for cluster A. The examples got the positive label in A and the negative label (“no, this is not hate speech”) in cluster B. Interestingly, while “islam” is a shared top word, we found it in the combination “radical islam” typically in cluster A.

> 🐦 *FYI world, the ppl of GB supporting #Brexit know if they don't control their own immigration/borders* **radical Islam** *will end their lives.*

Stealing jobs, a well-known negative prejudice towards foreigners, is also among the examples that are important for cluster A:

> 🐦 *Bloody foreigners coming here & taking our* **jobs** *though! #Brexit URL*

[8] Words with highest relative importance for cluster A: radical, job, illegal, invasion, love, can, let, merkel, mayor, then.

Identified Perspectives Overall, we find two polarized groups, both by label and feature agreement. Cluster A - the target group - is much more likely to give the positive label and this group of annotators consistently bases their opinion on a specific and distinct vocabulary which can be described as Islamophobic. Given the background information we have on all annotators, we identify cluster A as the *Muslim perspective* on the topic, highly sensitive to Islamophobic content. In opposition, for cluster B we did not find a characteristic vocabulary, those annotators form more a counter position to the migrant group, therefore we describe them as control group or *non-muslim perspective*. We conclude, annotators in cluster A are very sensitive towards islamophobic and, more general, xenophobic textual content.

5.3 Perspectives in the HS Italian Dataset

In this dataset, we found large differences between the number of items annotated by the different annotators. To avoid biasing our model, we only analyze annotators with a high rating count[9]. When clustering using both α and β agreement, we obtained the same clustering into the two clusters A, B of each 7 annotators.

Quantitatively, we found an anomaly here, as the label agreement in cluster A is almost zero ($\alpha_A = 0.03$), whilst in cluster B it is rather high ($\alpha_B = 0.42$). This already indicates that A is a *cluster of outliers*. Furthermore, the feature agreement is higher in cluster B ($\beta_B = 0.48$) as compared to A ($\beta_A = 0.34$). The latter appears to be due to noise only.

Qualitatively, we found that degrading talk about immigrants get positive labels from both clusters. For cluster B, we found examples with complains about immigrants driving up public costs by living in "hotels" as well as concerns about "sicurezza"(security) being diminished in the country after immigration.[10]

Identified Perspectives: in this dataset, we found a defined perspective in cluster B. The annotators tend to label a large spectrum of content - from critical, over conservative, nationalistic, to openly hateful tweets, all as hate. Hence, annotators in cluster B are very sensitive towards xenophobic textual content.

5.4 Perspectives in the Davidson Dataset

Analogously to the HS Italian dataset, we only analyze annotators with a high rating count[11]. We obtained different clusters according to α and β agreement. After computing the label feature agreement, we obtained cluster A of size 45 and cluster B of size 41. Note that, in contrast with all previous datasets, we

[9] this means for this dataset at least 800 ratings per annotator

[10] Words with highest relative importance for cluster B: hotel, #immigrati, spesa, se, clandestino, #gabbiaopen, giusto, tangere, succedere, #sicurezza (hotel, #immigrants, expense, if, illegal alien, #opencage, right, touch, succeed, #safety).

[11] for this dataset, at least 500 annotations per annotator

have two kinds of positive labels in this dataset, one for “offensive language” content and one for “hate speech” (stronger label).

Quantitatively, we found cluster B to have a much higher hate speech label rate (15.4%) over cluster A (4.1%). The base rate is 10%. While both cluster have comparable positive label rates, this indicates that cluster B has a tendency to give the hate speech label when the offensive label would have been an option. as compared to cluster A. Further, feature agreement is much lower in cluster A ($\beta_A = 0.22$) as opposed to cluster B ($\beta_B = 0.46$), indicating the annotators in cluster B agree much more on their important words.

Qualitatively, we found some words are understood by both clusters as hateful. As cluster A has a much lower positive label rate, A was rarely more critical than B. For cluster B we find several examples with the same keywords, centered around homophobic slurs such as “faggot”.[12]

Identified Perspectives: in this dataset, we find a defined perspective for cluster B. Annotators in this group give harsher labels when homophobic slurs are present in a tweet, as compared to annotators in A. We conclude that the annotators in B are highly sensitive towards homophobic textual content.

6 Conclusion

In this paper, we analyzed a number of annotated hate speech corpora, showing how the opinions of the annotators, reflected in their annotation, are far from uniformly distributed. In fact, the annotation of hate speech tends to be polarized, and our methodology is able to highlight the groups of annotators sharing similar opinions. We identified perspectives in the datasets, defined as increased sensitivity towards certain types of textual content (xenophobic, islamophobic, homophobic). Further, we introduced an automated method to support the manual exploration of the perspectives emerging from a polarized annotation of hate speech, resulting in consistent patterns describing why certain groups of people are more or less keen on judging a message as hateful.

As future work, we plan to test our methods with deeper and more refined linguistic features, to abstract away from individual words and therefore provide a more robust analysis. We also plan on investigating other NLP tasks traditionally considered less subjective, but recently found to contain informative disagreement [30], as well as non-linguistic tasks such as image labeling.

Finally, we note how this was was only possible thanks to the availability of non-aggregated datasets. In line with [5] and the *Perspectivist Data Manifesto*[13], we consider this factor crucial for research like ours.

[12] Words with highest relative importance for cluster B: hypocrite, til, mike, warn, fag, spread, faggot, jealous, tat, texas.

[13] https://pdai.info/

References

1. Akhtar, S., Basile, V., Patti, V.: A new measure of polarization in the annotation of hate speech. In: Alviano, M., Greco, G., Scarcello, F. (eds.) AI*IA 2019 – Advances in Artificial Intelligence. pp. 588–603. Springer International Publishing, Cham (2019)
2. Akhtar, S., Basile, V., Patti, V.: Modeling annotator perspective and polarized opinions to improve hate speech detection. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing **8**(1), 151–154 (Oct 2020), https://ojs.aaai.org/index.php/HCOMP/article/view/7473
3. Akhtar, S., Basile, V., Patti, V.: Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection (2021), https://arxiv.org/abs/2106.15896
4. AlDayel, A., Magdy, W.: Stance detection on social media: State of the art and trends. Information Processing & Management **58**(4), 102597 (2021). https://doi.org/https://doi.org/10.1016/j.ipm.2021.102597, https://www.sciencedirect.com/science/article/pii/S0306457321000960
5. Basile, V.: It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In: 2020 AIxIA Discussion Papers Workshop, AIxIA 2020 DP. vol. 2776, pp. 31–40. CEUR-WS (2020)
6. Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F.M., Rosso, P., Sanguinetti, M.: SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 54–63. Association for Computational Linguistics, Minneapolis, Minnesota, USA (Jun 2019). https://doi.org/10.18653/v1/S19-2007, https://www.aclweb.org/anthology/S19-2007
7. Basile, V., Cabitza, F., Campagner, A., Fell, M.: Toward a perspectivist turn in ground truthing for predictive computing. In: Proceedings of the XVIII Conference of the Italian chapter of AIS - Digital Resiliance and Sustainability: People, Organizations, and Society. pp. 1–16. Association for Intelligent Systems, Trento (Oct 2021), http://www.itais.org/itais2021-proceedings/pdf/21.pdf
8. Basile, V., Fell, M., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M., Uma, A.: We need to consider disagreement in evaluation. In: Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future. pp. 15–21. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.bppf-1.3, https://aclanthology.org/2021.bppf-1.3
9. Basile, V., Lai, M., Sanguinetti, M.: Long-term social media data collection at the university of turin. In: CLiC-it (2018)
10. Beigman Klebanov, B., Beigman, E., Diermeier, D.: Vocabulary choice as an indicator of perspective. In: Proceedings of the ACL 2010 Conference Short Papers. pp. 253–257. Association for Computational Linguistics, Uppsala, Sweden (Jul 2010), https://aclanthology.org/P10-2047
11. Campagner, A., Ciucci, D., Svensson, C.M., Figge, M.T., Cabitza, F.: Ground truthing from multi-rater labeling with three-way decision and possibility theory. Information Sciences **545**, 771–790 (2021)
12. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proceedings of the 11th International AAAI Conference on Web and Social Media. pp. 512–515. ICWSM '17 (2017)

13. Florio, K., Basile, V., Lai, M., Patti, V.: Leveraging hate speech detection to investigate immigration-related phenomena in italy. In: 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW). pp. 1–7. IEEE (2019)
14. Izsak-Ndiaye, R.: Report of the special rapporteur on minority issues, rita izsak : comprehensive study of the human rights situation of roma worldwide, with a particular focus on the phenomenon of anti-gypsyism. Tech. rep., UN,, Geneva :. 2015-05-11 (May 2015), http://digitallibrary.un.org/record/797194, submitted pursuant to Human Rights Council resolution 26/4.
15. Krippendorff, K.: Estimating the reliability, systematic error and random error of interval data. Educational and Psychological Measurement **30**, 61 – 70 (1970)
16. Lai, M., Tambuscio, M., Patti, V., Ruffo, G., Rosso, P.: Stance polarity in political debates: A diachronic perspective of network homophily and conversations on twitter. Data & Knowledge Engineering **124**, 101738 (09 2019). https://doi.org/10.1016/j.datak.2019.101738
17. Lin, W.H., Wilson, T., Wiebe, J., Hauptmann, A.: Which side are you on? identifying perspectives at the document and sentence levels. In: Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X). pp. 109–116. Association for Computational Linguistics, New York City (Jun 2006), https://www.aclweb.org/anthology/W06-2915
18. Maul, A., Mari, L., Wilson, M.: Intersubjectivity of measurement across the sciences. Measurement **131**, 764–770 (2019)
19. Miller, C., Arcostanzo, F., Smith, J., Krasodomski-Jones, A., Wiedlitzka, S., Jamali, R., Dale, J.: From brussels to brexit: Islamophobia, xenophobia, racism and reports of hateful incidents on twitter. DEMOS. Available at www. demos. co. uk/wpcontent/uploads/2016/07/From-Brussels-to-Brexit_-Islamophobia-Xenophobia-Racism-and-Reports-of-Hateful-Incidents-on-Twitter-Research-Prepared-for-Channel-4-Dispatches-% E2 **80** (2016)
20. Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., Cherry, C.: SemEval-2016 task 6: Detecting stance in tweets. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). pp. 31–41. Association for Computational Linguistics, San Diego, California (Jun 2016). https://doi.org/10.18653/v1/S16-1003, https://aclanthology.org/S16-1003
21. Mossie, Z., Wang, J.H.: Vulnerable community identification using hate speech detection on social media. Information Processing & Management **57**, 102087 (07 2019). https://doi.org/10.1016/j.ipm.2019.102087
22. O'Keeffe, G.S., Clarke-Pearson, K.: The impact of social media on children, adolescents, and families. Pediatrics **127**, 800 – 804 (2011)
23. Pamungkas, E.W., Basile, V., Patti, V.: Do you really want to hurt me? predicting abusive swearing in social media. In: The 12th Language Resources and Evaluation Conference. pp. 6237–6246. European Language Resources Association (2020)
24. Pan, Z., Lee, C.C., Man, J., So, C.: One event, three stories. Gazette **61**, 99–112 (04 1999). https://doi.org/10.1177/0016549299061002001
25. Poletto, F., Stranisci, M., Sanguinetti, M., Patti, V., Bosco, C.: Hate speech annotation: Analysis of an italian twitter corpus. In: Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), Rome, Italy, December 11-13, 2017. CEUR Workshop Proceedings, vol. 2006. CEUR-WS.org (2017)
26. Popescu, A.M., Pennacchiotti, M.: Detecting controversial events from twitter. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management. pp. 1873–1876. CIKM '10, ACM, New York,

NY, USA (2010). https://doi.org/10.1145/1871437.1871751, http://doi.acm.org/10.1145/1871437.1871751

27. Razo, D., Kübler, S.: Investigating sampling bias in abusive language detection. In: Proceedings of the Fourth Workshop on Online Abuse and Harms. pp. 70–78. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.alw-1.9, https://www.aclweb.org/anthology/2020.alw-1.9
28. Riloff, E., Wiebe, J.: Learning extraction patterns for subjective expressions. In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing. pp. 105–112 (2003), https://www.aclweb.org/anthology/W03-1014
29. Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., Stranisci, M.: An italian twitter corpus of hate speech against immigrants. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018). European Language Resource Association (2018), http://aclweb.org/anthology/L18-1443
30. Uma, A., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M.: A case for soft-loss functions. In: Proceedings of the 8th AAAI Conference on Human Computation and Crowdsourcing. pp. 173–177 (2020), https://ojs.aaai.org/index.php/HCOMP/article/view/7478
31. Warner, W., Hirschberg, J.: Detecting hate speech on the world wide web. In: Proceedings of the Second Workshop on Language in Social Media. pp. 19–26. LSM '12, Association for Computational Linguistics, Stroudsburg, PA, USA (2012), http://dl.acm.org/citation.cfm?id=2390374.2390377
32. Wiebe, J., Wilson, T., Bruce, R., Bell, M., Martin, M.: Learning subjective language. Computational Linguistics **30**, 277–308 (09 2004). https://doi.org/10.1162/0891201041850885
33. Wiegand, M., Ruppenhofer, J., Kleinbauer, T.: Detection of Abusive Language: the Problem of Biased Datasets. 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). pp. 602–608. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1060, https://www.aclweb.org/anthology/N19-1060
34. Yu, H., Hatzivassiloglou, V.: Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In: EMNLP (2003)
35. Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., Çöltekin, c.: SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In: Proceedings of SemEval (2020)
36. Zhang, Z., Luo, L.: Hate speech detection: A solved problem? the challenging case of long tail on twitter. Semantic Web **Accepted** (10 2018). https://doi.org/10.3233/SW-180338