Natural Language Engineering (2023), **29**, pp. 1481–1494
doi:10.1017/S1351324922000353

Cambridge University Press

ARTICLE

Quantifying the impact of context on the quality of manual hate speech annotation

Nikola Ljubešić[1,2,*], Igor Mozetič[1] and Petra Kralj Novak[1,3]

[1]Jožef Stefan Institute, Ljubljana, Slovenia, [2]Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia, and [3]Central European University, Vienna, Austria
[*]Corresponding author. E-mail: nikola.ljubesic@ijs.si

(Received 15 October 2021; revised 11 March 2022; accepted 19 May 2022; first published online 22 August 2022)

Abstract

The quality of annotations in manually annotated hate speech datasets is crucial for automatic hate speech detection. This contribution focuses on the positive effects of manually annotating online comments for hate speech within the context in which the comments occur. We quantify the impact of context availability by meticulously designing an experiment: Two annotation rounds are performed, one in-context and one out-of-context, on the same English YouTube data (more than 10,000 comments), by using the same annotation schema and platform, the same highly trained annotators, and quantifying annotation quality through inter-annotator agreement. Our results show that the presence of context has a significant positive impact on the quality of the manual annotations. This positive impact is more noticeable among replies than among comments, although the former is harder to consistently annotate overall. Previous research reporting that out-of-context annotations favour assigning non-hate-speech labels is also corroborated, showing further that this tendency is especially present among comments inciting violence, a highly relevant category for hate speech research and society overall. We believe that this work will improve future annotation campaigns even beyond hate speech and motivate further research on the highly relevant questions of data annotation methodology in natural language processing, especially in the light of the current expansion of its scope of application.

Keywords: Hate speech; Manual annotation; Inter-annotator agreement; Impact of context

1. Introduction

Natural language processing techniques have recently become prominent methods in many academic and application areas (Chen *et al.* 2020), which includes the task of hate speech detection (Basile *et al.* 2019; Zampieri *et al.* 2020). Given that the current paradigm in natural language processing heavily relies on supervised machine learning (Mohammad 2019), the annotation quality of the training and evaluation data has a direct impact on the resulting research and applications relying on natural language processing methods.

In this contribution, we investigate the impact of the discussion context when manually annotating online comments for occurrences of hate speech. Given that most of the manually annotated datasets for hate speech are annotated outside of the discussion context (Pavlopoulos *et al.* 2020), our central question is: What is the impact on the annotation quality if the data gathering and the manual annotation process were only moderately modified to include context?

In this work, we use the notion of hate speech as an umbrella term for various types of socially unacceptable online communications (Schmidt and Wiegand 2017), for example, offensive speech, toxic speech, hateful speech, inappropriate speech. While we use the term 'hate speech'

https://doi.org/10.1017/S1351324922000353 Published online by Cambridge University Press

primarily because of its frequent use in the literature, we are aware of its terminological ambiguity and refer the interested reader to a nuanced discussion in (Fortuna and Nunes 2018).

Most hate speech datasets, including recent standard-setting shared tasks on hate speech identification (Basile *et al.* 2019; Zampieri *et al.* 2019, 2020), are annotated without taking into account the wider narrative in which the online comments occur. However, there are also exceptions to this common practice. The English dataset described in Gao and Huang (2017) takes into account, during the annotation, the previous comment and the title of the post. The Arabic dataset described in Mubarak, Darwish, and Magdy (2017) takes into account only the title of the post. The English, Slovenian and Croatian datasets described in Ljubešić, Fišer, and Erjavec (2019, 2021a) perform annotation of full Facebook discussion threads in-context, but do not quantify in any way the impact of the context-sensitive annotation campaign. In Pavlopoulos *et al.* (2020), 250 English comments from Wikipedia Talk Pages are annotated for toxicity in-context and out-of-context. The in-context annotation is performed by giving access to the title and the preceding comment. Different annotators perform the in-context and the out-of-context annotation campaign. The main finding of this paper is that the toxicity of comments is significantly higher when annotated in-context compared to out-of-context, but the quality of the annotations is not analysed. The Civil Comments in Context (CCC) dataset (Xenos, Pavlopoulos, and Androutsopoulos 2021) uses the English Civil Comments (CC) dataset not annotated in-context and annotates a 10,000 comments subset of the CC dataset in-context, releasing both the original out-of-context and the new in-context annotation as part of the CCC dataset. This dataset, however, contains in-context and out-of-context labels that were obtained in different projects by different researchers, using different crowdsourcing platforms and, highly likely, different annotators. The main finding of this research, similar to Pavlopoulos *et al.* (2020), is that in-context annotations result in more toxic annotations. Finally, the CONDA dataset (Weld *et al.* 2021) annotates in-gaming communication for toxicity in full conversations, focusing primarily on slot and intent annotation, and disregarding our main research focus.

The main insights gained from the literature overview are the following. While most datasets are not annotated in-context (Basile *et al.* 2019; Zampieri *et al.* 2019, 2020), most works that do annotate in-context do not investigate in any way the impact of such an annotation mode (Gao and Huang 2017; Mubarak *et al.* 2017; Ljubešić *et al.* 2019; Weld *et al.* 2021). Recently, the interest in the question of the context relevance for manual data annotation for hate speech has risen. The question, mostly answered by now, is what the impact is of the in-context annotation mode on the hate speech label distribution, with a trend of in-context annotations containing more instances annotated as hateful than out-of-context annotations (Pavlopoulos *et al.* 2020; Xenos *et al.* 2021). The datasets that these measurements are based on have a series of shortcomings, for example, their size in case of (Pavlopoulos *et al.* 2020) and the lack of control over the annotation process, that is, using different annotators in each annotation round (Pavlopoulos *et al.* 2020; Xenos *et al.* 2021), and even relying on pre-existing annotations from other projects (Xenos *et al.* 2021). Given the high complexity of the task at hand, our position is that to accurately measure the impact of performing annotations in different modes (in-context or out-of-context), all the remaining factors, such as the annotators, the annotation platform, have to be controlled as much as possible. To the best of our knowledge, up to this point, no investigation of the impact on the quality of the annotations in the two modes was performed. While observing a different distribution of hate speech labels does show that the results of the annotation processes in the two modes are different, it does not state anything about the quality of these annotations. One could assume that annotations performed in-context are better, but this should not be taken without any evidence, nor without quantifying the level of such a potential effect. One can easily identify an argument against this assumption: if annotators are given more information at their disposal, they might get overloaded, with a possible negative impact on the annotation quality.

https://doi.org/10.1017/S1351324922000353 Published online by Cambridge University Press

In this work, we push the research of the impact of context further by quantifying the quality of the resulting annotations depending on whether the annotation is performed in-context or out-of-context in the following manner:

- We perform two annotation rounds, in-context and out-of-context, on identical data, by using the same annotators, making sure that annotators never annotate the same comments in the two annotation rounds. With this, we control for the data selection and annotator selection bias.
- We measure the annotation quality through inter-annotator agreement (IAA). To obtain a maximally objective assessment of the annotation quality, we make sure that each comment in each annotation round is annotated by two annotators and that the intersection of annotations between any two annotators is balanced. With this, we minimize the annotator difference bias.

The goal of this paper is to answer the following research questions:

- Q1: Is the quality of hate speech annotations in-context equivalent to the out-of-context annotations? We hypothesize that this is not the case and that the in-context annotations are of higher quality.
- Q2: Is the hypothesized difference in the quality of hate speech annotations (Q1) higher among replies to comments than among the original comments? We hypothesize, given the higher level of contextual dependence of replies, that the annotation quality difference will be higher for replies than for comments.
- Q3: What are the differences in the hate speech labels between the in-context and out-of-context annotation rounds? We hypothesize that the in-context label distribution has more hateful labels, as already demonstrated in previous work (Pavlopoulos *et al*. 2020; Xenos *et al*. 2021), but also investigate additional differences between the label assignments.

Our results show that the in-context annotations are of higher quality than out-of-context annotations. Replies show to have lower annotation quality in comparison to comments in each of the two scenarios, especially when the annotations are performed out-of-context. Finally, out-of-context annotations favour the no-hate-speech class, especially among violent comments. We claim that these findings are very important for the field as the costs of performing annotations in-context and out-of-context are comparable.

2. Materials and Methods

The source of our data is YouTube comments occurring under videos about the COVID-19 pandemic published at the beginning of May 2020.[a] In particular, we performed a keyword search, using the official YouTube Data API, for videos matching COVID-related keywords, for example, {*coronavirus, nCov, coronavirus, corona-virus, covid, SARS-CoV*}. An in-depth search was then performed by crawling the network of related videos as provided by the YouTube algorithm. We filtered the videos that match our set of keywords in the title or description from the gathered collection. Finally, we collected the comments posted under these videos. The titles, the video descriptions and the comments are in English, according to Google's cld3 language detection service.[b] The set of videos covers the first week of May 2020, while the comments range from May 1, 2020, to January 23, 2021. Out of the whole data collection, 100 discussion threads of lengths

[a] Such a specific time frame was used for this dataset to be temporally disjoint with the training data for hate speech classification model, published between February 2020 and April 2020. We do not use the training data in this research.

[b] https://github.com/google/cld3

https://doi.org/10.1017/S1351324922000353 Published online by Cambridge University Press

varying between 10 and 200 comments were randomly sampled, resulting in a dataset of 10,759 YouTube comments and replies.

The internal structure of each discussion thread on YouTube is as follows. First, the YouTube video title and video description are provided. Then, the discussion thread follows, each consisting of comments and replies to a specific comment. There can be multiple replies to the comment, but not replies to replies. While the ordering of replies to the comment (we refer to these as micro-threads) is temporal, the order between comments, that is, micro-threads themselves, is very much unclear, especially because of different orderings of comments depending on the platform used to interact with the YouTube content. For this reason, we assume that the order inside micro-threads is known, while for the relation between micro-threads, a bag of micro-threads structure has to be assumed.

Ten human annotators were taking part in the annotation campaign. Their demographics are the following. They are all in their twenties, and master students from the University of Ljubljana, Slovenia, enrolled in programmes from the area of social sciences and humanities. Six annotators declared themselves as male, while the remaining four annotators as female. The annotators are native speakers of Slovenian with a high level of proficiency in English. They have previously taken part in a series of hate speech annotation campaigns, including data in Slovenian, Croatian (a closely related South-Slavic language) and English. Each of the annotators has previously annotated about 22,000 comments in Slovenian, 10,000 comments in Croatian and 20,000 comments in English.

Our annotation schema distinguishes between four classes of speech: acceptable (normal, not hateful), inappropriate (obscene or vulgar expressions), offensive (including offensive generalization, contempt or dehumanization) and violent (inciting physical violence toward some target group or individual). For more details, see the annotation guidelines (Novak *et al.* 2021) or the dataset itself (Ljubešić *et al.* 2021b).

The annotation procedure of the YouTube comments was performed in two annotation rounds. In the first round, comments were annotated in-context, that is, each annotator was annotating full discussion threads under specific videos, with micro-threads sorted in temporal order. The second annotation round was performed out-of-context, that is, the annotators were given comments and replies in isolation, sampled from multiple discussion threads and ordered randomly.

The order of annotation rounds is probably irrelevant in our case as our annotators took part in five comparable annotation campaigns before this one. However, for additional methodological rigour, we decided to perform the in-context annotation first, and then, two weeks later, the out-of-context annotation. This design ensures that any annotator training effect, considered very unlikely in our case, works against our hypothesis that the in-context annotations are of higher quality than the out-of-context annotations.

The following criteria were used to divide the threads of comments between the ten annotators:

- Each comment must be annotated twice by two different annotators;
- Each annotator should get approximately the same number of comments to annotate;
- Each pair of annotators should have approximately the same overlap;
- Each annotator should have both long and short threads;
- In the second, out-of-context round, no comment annotated by an annotator in the first round should be given to the same annotator.

The overlap of annotations between different annotators is crucial for this research as it enables the calculation of the IAA, which is our quantification of the annotation quality.

We consider and adhere to the following necessary conditions for a fair comparison of annotating in-context and out-of-context:

https://doi.org/10.1017/S1351324922000353 Published online by Cambridge University Press

- We annotate the same comments in-context and out-of-context, thereby mitigating the data selection bias.
- We use very well-trained annotators that worked already on a series of annotation campaigns with the same annotation schema.
- We commission the same annotators for both annotation rounds, thereby mitigating the annotator selection bias.
- We make sure that no data leakage between annotation rounds occurs, that is, that a specific annotator never annotates the same comment in both annotation rounds.
- We assure a similar overlap in annotation between every two annotators in each annotation round, thereby obtaining an as good as possible assessment of annotation quality through annotator agreement.

The annotator agreement is calculated by Krippendorff's Alpha reliability (Krippendorff 2018). Alpha generalizes several specialized agreement measures (such as Scott's π, Fleiss' K, Spearman's rank correlation coefficient and Pearson's intraclass correlation coefficient), takes ordering of classes into account and has the agreement by chance as the baseline.

We consider the four classes of speech ordered, from acceptable speech to increasing levels of hate speech: acceptable, inappropriate, offensive and violent. We actually calculated the inter-annotator Alpha agreement for both, the nominal and ordinal variable type. However, in the remainder we present only the results for the ordinal case, as the relations and the conclusions are very much the same for both variable types. One final argument for the variable to be considered ordinal is the fact that in our experiments Alpha, regardless of the annotation mode, is always higher when the variable is considered ordinal than nominal. This gives experimental evidence that the annotators perceive the proposed speech classes as ordered.

We operationalize the quality of an annotation round as a distribution of the Alpha agreement of each annotator with the remaining annotators. Namely, each pair of annotators has a similar overlap of annotated comments. Each annotation campaign therefore is described through a distribution of Alpha scores of size $N = 10$.

3. Results

In this section, we provide answers to the three research questions, each in its own subsection. We start by presenting the comparison of the IAA distribution when annotating in-context and when annotating out-of-context. We continue by comparing these distributions across comments and replies to comments. We conclude the results section by analysing the label differences between the two annotation rounds.

3.1. *Comparison of in-context and out-of-context IAA*

To set off the analyses over the in-context and out-of-context distributions of IAA, we first check both distributions for normality, to define the type of descriptions and statistical tests to be used. We apply the Shapiro-Wilk normality test (Shapiro and Wilk 1965) over each distribution and obtain the result of ($W = 0.98, p = 0.99$) for the in-context distribution and the result of ($W = 0.66, p = 0.007$) for the out-of-context distribution. The results indicate that for the out-of-context distribution, we can reject the null hypothesis that the data came from a normal distribution; therefore, we continue our data analysis via non-parametric means.

The two distributions are described in Table 1 and plotted as boxplots in Figure 1. In the tabular representation of the distributions, we calculate min and max due to the small number of instances in the distribution ($N = 10$) and to better understand the variability of IAA between the annotators, while the quartiles of the distribution can be inspected in the boxplot. The calculated median,

https://doi.org/10.1017/S1351324922000353 Published online by Cambridge University Press

Table 1. Comparison of the in-context and out-of-context agreement distributions

	Median	Min	Max
In-context IAA distribution	0.617	0.507	0.708
Out-of-context IAA distribution	0.582	0.407	0.633

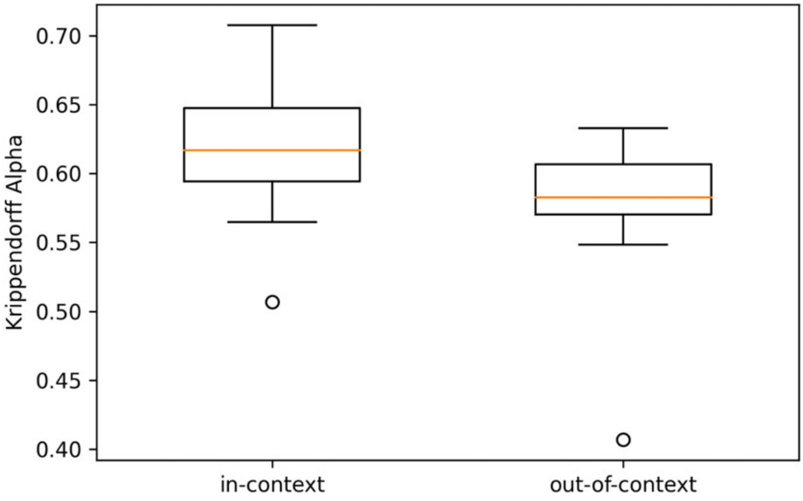

Figure 1. A boxplot comparison of the in-context and out-of-context agreement distributions.

min and max of each distribution are a clear sign of higher IAA in the in-context distribution. Both distributions show a reasonable variability of IAA between the annotators, with an absolute difference between the least- and the most-agreeing annotator around 0.2. The agreement in general is in line with our results from previous work with the same annotation schema (Ljubešić *et al.* 2019; Evkoski *et al.* 2022) and higher compared to some other hate speech annotation campaigns (Salminen *et al.* 2019).

Given that the two described distributions contain only ten measurements, it is paramount to perform statistical testing of the observed differences. We can consider the two distributions to be paired as they contain IAA estimates for the same ten annotators. We therefore perform the two-tailed paired data Wilcoxon sign-ranked test (Wilcoxon 1992), which tests the null hypothesis that two related paired samples come from the same distribution. In particular, it tests whether the distribution of the differences of measurements is symmetric around zero. It is a non-parametric version of the paired t-test. The test statistic T, in the two-tailed version, calculates the sum of the ranks of the differences above or below zero, whichever is smaller. The test produces a result of ($T = 4.0, p = 0.014$). The T statistic encodes that the smallest and the third smallest absolute difference in the measurements were the only two, out of 10, negative differences, that is, where the IAA in the out-of-context campaign was higher than in the in-context campaign, resulting in a statistic of $T = 1 + 3 = 4$. The p-value, which encodes the probability of obtaining such a statistic if both samples came from the same distribution, is small enough so that we can safely reject the null hypothesis.

We calculate the common language effect size (CLES) (McGraw and Wong 1992) in order to quantify the effect of the difference observed between the two distributions. CLES compares the

https://doi.org/10.1017/S1351324922000353 Published online by Cambridge University Press

Table 2. Comparison of the in-context and out-of-context agreement distributions for comments and for replies

		Median	Min	Max
In-context IAA distribution	comments	0.617	0.507	0.705
Out-of-context IAA distribution	comments	0.607	0.390	0.645
In-context IAA distribution	replies	0.600	0.464	0.711
Out-of-context IAA distribution	replies	0.548	0.434	0.620

two paired distributions and calculates in what percentage of pairs the expected difference, in-context IAA being higher than out-of-context IAA, is observed. CLES for the two distributions is 0.8, that is, for 8 out of 10 annotators IAA is higher when annotating in-context in comparison to out-of-context.

The reported measurements show a clear difference between IAA when performing in-context and out-of-context annotations. The difference is not striking and is expected given the previous measurements showing that a rather small percentage of instances are annotated differently in-context vs. out-of-context (Xenos *et al.* 2021). However, the higher quality of annotations in-context is obvious and cannot be easily discarded given a comparable cost of the in-context and out-of-context annotation procedures.

3.2. Comparison of the IAA on replies and on comments

Next, we address the second research question, whether the difference between IAA in the in-context and out-of-context annotation campaigns is greater for replies to the comments than for the comments themselves. We hypothesize that the difference is greater for the replies as these should be more context-dependent than comments. While comments do rely on the overall context of the video and the whole discussion thread, the replies are additionally very directly related to the comments and earlier replies in the specific micro-thread.

We perform this analysis by splitting the annotations in both annotation rounds to those made on comments and on replies. We then calculate, for each annotator and each annotation round, the Alpha IAA separately over comments and separately over replies. Now there are four distributions (in-context vs. out-of-context, and comments vs. replies), one of which did not pass the normality test. Therefore, we again opt for the safer non-parametric descriptions and tests over the data. A non-parametric description of these distributions is given in Table 2, and a boxplot is presented in Figure 2.

There are two observations to be made: (1) the difference in the in-context and out-of-context IAA distribution appears to be larger for replies than for comments, and (2) replies seem to be harder to annotate overall, obtaining lower IAA regardless of the annotation mode. While the first observation goes towards answering our second research question, the second observation was not envisaged during the question stating phase of our research.

We first focus on the first observation and perform a two-tailed paired data Wilcoxon sign-ranked test between the in-context IAA distribution of comments and the out-of-context IAA distribution of comments, yielding a test result of ($T = 10.0, p = 0.084$). When we perform the same test between the in-context and out-of-context IAA distribution of replies, we get a result of ($T = 4.0, p = 0.014$). This result supports the conclusion about a significant difference in IAA between replies annotated in-context and out-of-context. On the other hand, for comments the difference is on the verge of statistical significance. In terms of the effect size, CLES is 0.8 for replies and 0.7 for comments. Given the rather large value of CLES for comments (7 out of

https://doi.org/10.1017/S1351324922000353 Published online by Cambridge University Press

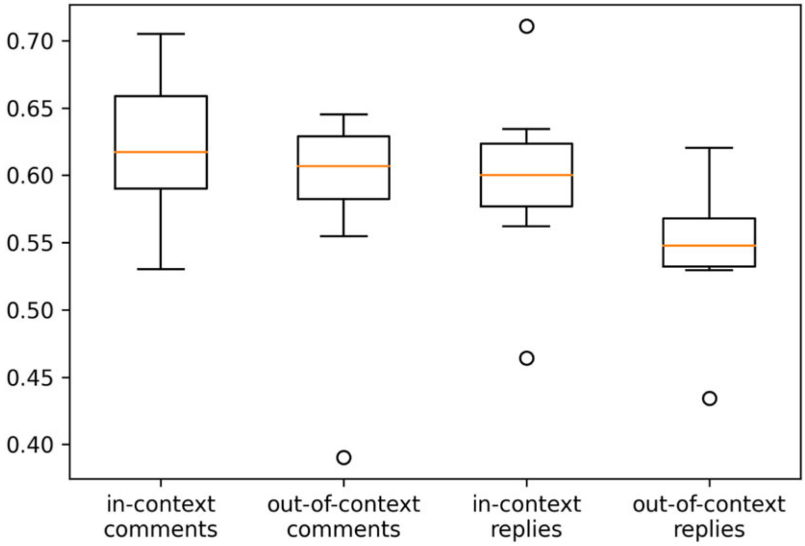

Figure 2. A boxplot comparison of the in-context and out-of-context agreement distributions.

10 annotators have higher IAA in-context than out-of-context), we can assume that the lack of statistical significance is primarily due to small number of data points (10 annotators only). However, also from the point of view of the effect size, the in-context annotations seem to be more important when annotating replies than when annotating comments, which supports our initial hypothesis.

Our second observation is that comments appear to be easier to annotate overall than replies. To properly check whether this observation is supported by statistics, we perform a test over merged, in-context and out-of-context IAA distributions. We thereby directly compare the IAA on comments on one side and on replies on the other, regardless of the annotation mode. The two-tailed paired data Wilcoxon sign-ranked test shows a significant difference ($T = 29.0, p = 0.003$), and the CLES effect size is 0.75 (in 15 out of 20 cases, annotations yield higher IAA for comments than for replies).

Regarding the reason for the difference in IAA between comments and replies regardless of the annotation mode, our first hypothesis is that replies might be shorter than the original comment; therefore, their meaning might be harder to decode for human annotators. It is well known that both automatic and manual annotations are harder for short than for longer texts simply due to the lack of information (Tiedemann and Ljubešić 2012).

Another hypothesis is that the information present in the replies is intertwined with the information from the original comment, but potentially also with the previous replies to the same comment, making thereby its status of acceptability harder to decode.

While we can only assume that the second hypothesis holds to some extent, the first hypothesis about the difference in the length of comments and replies can be checked by investigating the length distribution of the comments and replies. We can reject the first hypothesis already via descriptive statistics since the replies are on average actually longer (mean 177 characters, median 100 characters) than the comments (mean 147 characters, median 84 characters).

The fact that replies are on average longer than comments supports the second hypothesis that the content of replies is more intertwined with the previously posted content, thereby making decoding of the acceptability of the reply even harder.

This short discussion on the interesting and unexpected observation that replies are harder to annotate than comments, regardless of the annotation mode, should be considered just

Table 3. Comparison of the in-context and out-of-context hate speech probability distributions

	Acceptable	Inappropriate	Offensive	Violent
In-context distribution	71.6%	1.0%	27.0%	0.5%
Out-of-context distribution	74.1%	0.9%	24.7%	0.4%

preliminary, and more rigorous experiments with human participants should be performed before any conclusions are to be drawn.

3.3. Difference in the hate speech annotations between the in-context and out-of-context annotation rounds

In this subsection, we focus on the third research question: how do the hate speech annotations change between the in-context and out-of-context annotation modes? We first present the probability distribution of the final annotations (we use all the annotations provided by all the annotators) in Table 3. The distributions show a small, but clear tendency of annotators labelling the comments with more unacceptable classes if context is available, as compared to if context is omitted in the annotation campaign.

All three unacceptable classes (inappropriate, offensive and violent) obtain a smaller probability mass out-of-context, which results in the acceptable class getting 2.5 absolute percentage points more. We test the significance of this difference, more precisely whether the variables of acceptability and context availability during annotation are independent, by a chi-square test over the raw counts of the two distributions. The result ($\chi = 35.33, p = 1.03e - 07, df = 3$) shows that the variables of acceptability and context availability during annotation are dependent. To quantify the measure of association between the two variables, we calculate the Cramer's V effect size (Cramér 1946), a quantification of the correlation between two nominal variables ranging between 0 and 1. The result of 0.03 indicates a very low effect size, as is expected given the rather minor, but still significant differences in the two distributions.

The impact of the observed distribution changes on downstream natural language processing tasks and applications relying on manually annotated hate speech data is not to be underestimated. If the training data are of low quality, this influences both the reliability of evaluation results, but also the downstream applications. Our results, as well as previous research, show that out-of-context annotations favour assigning non-hate-speech labels. This implies that out-of-context annotation campaigns disguise hate speech utterances and introduce non-hate-speech-favouring bias. This particular bias has a very undesirable side effect of hiding hate speech utterances, both from the machine learning model in the evaluation phase, as well as during the downstream application phase. When it comes to the application of such a model to unseen data, a lower hate speech detection rate can be expected. This is a very undesirable behaviour for a hate speech detection model in a real-world scenario, where a model is used for screening vast amounts of content and signalling for potentially hateful content to the human moderator.

We further investigate how exactly the annotations relate among the four classes between the in-context and the out-of-context annotation mode. In Figure 3, we show a confusion matrix to compare the two annotation modes. While investigating the third research question, we use all the annotation interactions between the two modes (around 20,000 annotations per mode), which results in about 40,000 data points. The raw counts in the confusion matrix show a consistent flow from unacceptable classes in the in-context mode to the acceptable class in the out-of-context mode. This tendency is rather obvious if one compares the lower left part of the matrix (the number of annotation pairs for which the level of unacceptability is higher in-context compared to

https://doi.org/10.1017/S1351324922000353 Published online by Cambridge University Press

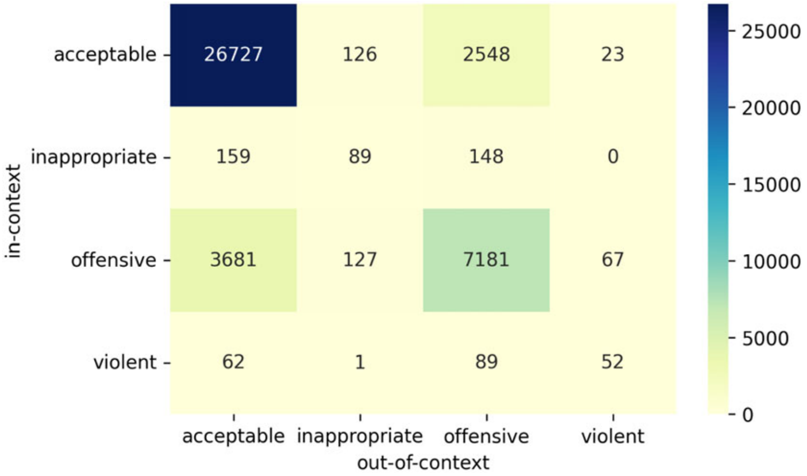

Figure 3. Confusion matrix between the in-context and out-of-context annotation rounds. The matrix comprises about 40,000 data units, accounting for two annotations of each of the 10,000 units, and two annotation modes.

the out-of-context mode) to the upper right part (the number of annotation pairs for which the level of unacceptability is higher in the out-of-context mode). Out of the six pairs of classes, only the (inappropriate, offensive) pair has a higher number of transitions inappropriate ↦ offensive from the in-context to the out-of-context mode (148 vs. 127).

An especially striking difference is between the annotations that are considered violent in one, but acceptable in another annotation round, with 2.7 times more violent cases in the in-context mode than in the out-of-context mode (62 vs. 23). In general, the violent class is very infrequent and the annotations in which the two modes agree on this label (52) is actually lower than the number of annotations that were annotated as violent in-context, but acceptable in the out-of-context mode (62). The violent class is the most problematic category content-wise and also the most interesting for legal repercussions of hate speech (Fišer, Erjavec, and Ljubešić 2017). Given the low frequency, but high importance in accurately annotating the violent class, this surely is another important argument for performing hate speech annotations in-context.

In the remainder of this section, we give a few examples of comments that were annotated as acceptable in one annotation round and violent in the other. We first give a few examples of comments annotated as violent in-context:

- The comment *'Take your country back my fellow citizens!'* was annotated as violent by both annotators in the in-context mode, but by just one annotator in the out-of-context mode, the other one chose the acceptable class. The comment was posted under a video on a right-wing populist channel discussing coronavirus-related job losses. The whole discussion thread is very loaded and aggressive.
- The comment *'Good time to disinfect the room with a bomb'*. was annotated as violent by both annotators in the in-context mode, and as acceptable by both annotators in the out-of-context mode. The video shows the US president meeting with the members of the congress. One could argue that the comment is quite likely violent in many contexts, but the annotators in the out-of-context mode were obviously not ready to make this assumption given the lack of information on what room the comment is referring to.
- The comment *'Anybody know how to get a gun in Canada? I'm ready to go to war'*. was annotated as violent by both annotators in the in-context campaign, but only by one annotator in the out-of-context campaign, the other annotator choosing the acceptable

https://doi.org/10.1017/S1351324922000353 Published online by Cambridge University Press

label. The comment occurs below a video reporting on a press conference of the Canadian prime minister, announcing large investments in healthcare, but also increased gun control. While one can assume that this comment might incite violence, going to war is still considered acceptable in most cultures, which did not allow one of the annotators to choose the violent class.

Some examples of comments annotated as acceptable in-context, but violent out-of-context are the following:

- The comment *'Drink disinfectant. . ..the President swears by it!'* was annotated by all annotators as acceptable, except for one annotator in the out-of-context mode, who was referring to the potential incitement to self-harm, that is, violence. The comment has been published under a video of a progressive YouTube channel reporting on Mike Pence staffer testing positive for COVID-19. Therefore, the comment was obviously sarcastic.
- The comment *'The deep state is behind this all. They need to be purged'*. was annotated by annotators in the in-context mode as offensive and violent, while in the out-of-context mode the annotators annotated the comment as acceptable and offensive. The comment was posted under a video reporting on a story of a coronavirus researcher killed in suspected murder-suicide. The video was posted on a channel well known for disinformation spreading, the discussion thread clearly mirroring such world view. This example depicts very clearly the complexity of the task at hand, not allowing annotators to reach very high overall agreement.

Interestingly, most gross disagreements, such as the ones shown above between the acceptable and violent class, are not based so much on the local micro-thread context, but the general topic of the video and the discussion thread as a whole. All the examples given here depict very clearly two facts about the task of manual annotation of hate speech in online comments: (1) the task is very hard as many comments carry a lot of implications with them, (2) the context of the comment is in some cases very important for the task, shifting the overall intent of the comment.

4. Discussion and conclusion

In this paper, we explore the impact of discussion context in manual annotation of online comments on the resulting annotation quality. We compare annotating comments for hate speech detection in full context and outside the context, the latter being by far the most frequent way of performing annotation campaigns in the hate speech research community. We measure the annotation quality through IAA. To the best of our knowledge, this is the first scientific work inspecting the dependence of manual annotation quality and context availability. Previous research only scarcely considered differences in the label distribution depending on the mode of the annotation campaign.

Our research relies on meticulous experimental design, where we use the same data and the same annotators in both annotation scenarios, with a uniform overlap of comments to be labelled between the annotators, and no data leak between the two annotation rounds. The participating annotators are highly experienced, with multiple similar annotation campaigns behind them.

Our findings show a very consistent positive impact of the available context on the quality of the manual annotations. The positive effect of the context availability is especially noticeable on annotations of replies to comments. Furthermore, comments are annotated more consistently than replies, regardless of the annotation mode. This is quite probably due to replies being more informationally intertwined with their context – the comments and previous replies, than the comments themselves.

https://doi.org/10.1017/S1351324922000353 Published online by Cambridge University Press

While in-context annotations do result in significantly higher IAA of 3.5 points of Alpha on average, the observed difference is in no way staggering. However, given the comparable costs of performing annotations in- and out-of-context, we simply cannot identify reasons not to perform annotations in-context when the context is available.

Comparing the resulting labels of the two annotation campaigns, we can confirm the previous findings that out-of-context annotations favour the category of acceptable speech, but also that the largest relative difference between the two annotation campaigns is on the highly infrequent and problematic violent class. A manual inspection of examples classified differently, as acceptable speech in one campaign and violent in another, shows that these drastic differences have roots in the unavailability of the overall context of the discussion thread, and not in the finesses of a specific micro-thread.

While this research casts important light on the question of interaction of context availability and annotation quality, there are still a number of limitations to be taken under consideration.

The first limitation is the way we measure annotation quality. We measure it via IAA, which surely is a relevant signal, but does not come close to measuring the phenomenon of annotation quality in full.

Another, even more distant signal besides IAA that could measure annotation quality could be the impact on the performance of a supervised classifier. We do not investigate this angle here, mainly for two reasons: (1) the current technology is still not capable of successfully exploiting the discussion context (Pavlopoulos *et al.* 2020), and (2) we cannot expect the computer to solve a task outside of context for which humans require this same context.

Recent studies show that annotations can differ for reasons such as socio-demographic factors and cultural differences (Akhtar, Basile, and Patti 2021). We do believe that we control for those factors by using master students born in the same small Central-European country, enrolled in similar studies, and living in an overall similar socio-economic situation. We do not know what the results of our study would be if annotators came from a different socio-demographic or cultural background, but we have no reason to believe that the conclusions would be much different.

Another question is to what level can we generalize beyond the social medium that we perform our experiments on. Our data are collected from YouTube, consisting of a bag of micro-threads (series of replies to a comment). We expect similar results on Facebook comments, given that these are also organized in bags of micro-threads. However, it is an open question to what extent our findings would hold for Twitter where only a portion of tweets represents replies to other tweets.

The question of topic also arises, where our YouTube comments are collected around the topic of COVID. Our expectation is that the topic does not have any significant effect on the usefulness of context while performing manual hate speech annotation.

Finally, we can also put forward the question to what level we can generalize from the case of hate speech detection to other annotation tasks on social media data. We conjecture that one can expect a similar level of context dependence of the annotation process for phenomena of similar complexity, for example, natural language processing tasks of emotion detection, stance detection, argumentation mining, etc.

All these limitations discussed above are primarily words of caution and an invitation for additional research. We do, however, stand by our main conclusion that no manual hate speech annotation should be performed outside the conversational context in which it occurs as this is mostly just a question of the annotation process organization, with the benefit of significant improvements to the obtained annotations.

Acknowledgement. The authors thank Matteo Cinelli for collecting the YouTube data.

https://doi.org/10.1017/S1351324922000353 Published online by Cambridge University Press

Financial support. The authors acknowledge financial support from the Slovenian Research Agency (research core funding no. P2-103 and P6-0411), the Slovenian Research Agency and the Flemish Research Foundation bilateral research project LiLaH (grant no. ARRS-N6-0099 and FWO-G070619N), and the European Union's Rights, Equality and Citizenship Programme (2014–2020) project IMSyPP (grant no. 875263). The European Commission's support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

References

Akhtar S., **Basile V.** and **Patti V.** (2021). Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. *CoRR*, abs/2106.15896.

Basile V., **Bosco C.**, **Fersini E.**, **Nozza D.**, **Patti V.**, **Rangel Pardo F. M.**, **Rosso P.** and **Sanguinetti M.** (2019). 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*. Minneapolis: Association for Computational Linguistics, pp. 54–63.

Chen X., **Xie H.**, **Cheng G.**, **Poon L. K. M.**, **Leng M.** and **Wang F. L.** (2020). Trends and features of the applications of natural language processing techniques for clinical trials text analysis. *Applied Sciences* **10**(6), 2157.

Cramér H. (1946). *Mathematical methods of statistics*. Princeton: Princeton University Press.

Evkoski B., **Pelicon A.**, **Mozetič I.**, **Ljubešić N.** and **Novak P. K.** (2022). Retweet communities reveal the main sources of hate speech. *PLoS ONE* **17**(3), e0265602.

Fišer D., **Erjavec T.** and **Ljubešić N.** (2017). Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in Slovene. In *Proceedings of the First Workshop on Abusive Language Online*. Vancouver: Association for Computational Linguistics, pp. 46–51.

Fortuna P. and **Nunes S.** (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)* **51**, 1–30.

Gao L. and **Huang R.** (2017). Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. Varna: INCOMA Ltd, pp. 260–266.

Krippendorff K. (2018). *Content Analysis: An Introduction to Its Methodology*, 4th edn. Thousand Oaks, CA: Sage Publications.

Ljubešić N., **Fišer D.** and **Erjavec T.** (2019). The FRENK datasets of socially unacceptable discourse in Slovene and English. In Ekštein K. (ed), *Text, Speech, and Dialogue. TSD 2019*. Lecture Notes in Computer Science, vol. 11697. Cham: Springer. Available at https://link.springer.com/chapter/10.1007/978-3-030-27947-9_9.

Ljubešić N., **Fišer D.** and **Erjavec T.** (2021a). Offensive language dataset of Croatian, English and Slovenian comments FRENK 1.0. Available at http://hdl.handle.net/11356/1433. Slovenian language resource repository CLARIN.SI.

Ljubešić N., **Mozetič I.**, **Cinelli M.** and **Kralj Novak P.** (2021b). English YouTube hate speech corpus. Available at http://hdl.handle.net/11356/1454. Slovenian language resource repository CLARIN.SI.

McGraw K. O. and **Wong S. P.** (1992). A common language effect size statistic. *Psychological Bulletin* **111**(2), 361–365.

Mohammad S. M. (2019). The state of nlp literature: A diachronic analysis of the acl anthology. *arXiv preprint, arXiv: 1911.03562.*

Mubarak H., **Darwish K.** and **Magdy W.** (2017). Abusive language detection on Arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*. Vancouver: Association for Computational Linguistics, pp. 52–56.

Novak P. K., **Mozetič I.**, **Pauw G. D.** and **Cinelli M.** (2021). IMSyPP deliverable D2.1: Multilingual hate speech database. Jožef Stefan Institute, Ljubljana, Slovenia. Available at http://imsypp.ijs.si/wp-content/uploads/2021/12/IMSyPP_D2.2_multilingual-dataset.pdf.

Pavlopoulos J., **Sorensen J.**, **Dixon L.**, **Thain N.** and **Androutsopoulos I.** (2020). Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online. Association for Computational Linguistics, pp. 4296–4305.

Salminen J., **Almerekhi H.**, **Kamel A. M.**, **Jung S.-g** and **Jansen B. J.** (2019). Online hate ratings vary by extremes: a statistical analysis. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, pp. 213–217.

Schmidt A. and **Wiegand M.** (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. Valencia: Association for Computational Linguistics*, pp. 1–10.

Shapiro S. S. and **Wilk M. B.** (1965). An analysis of variance test for normality (complete samples). *Biometrika* **52**(3-4), 591–611.

Tiedemann J. and **Ljubešić N.** (2012). Efficient discrimination between closely related languages. In *Proceedings of COLING 2012. Mumbai: The COLING 2012 Organizing Committee*, pp. 2619–2634.

Weld H., **Huang G.**, **Lee J.**, **Zhang T.**, **Wang K.**, **Guo X.**, **Long S.**, **Poon J.** and **Han S. C.** (2021). Conda: a contextual dual-annotated dataset for in-game toxicity understanding and detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online. Association for Computational Linguistics, pp. 2406–2416.

https://doi.org/10.1017/S1351324922000353 Published online by Cambridge University Press

Wilcoxon F. (1992). Individual comparisons by ranking methods. In *Breakthroughs in statistics*. New York: Springer, pp. 196–202.

Xenos A., **Pavlopoulos J.** and **Androutsopoulos I.** (2021). Context sensitivity estimation in toxicity detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*. Online. Association for Computational Linguistics, pp. 140–145.

Zampieri M., **Malmasi S.**, **Nakov P.**, **Rosenthal S.**, **Farra N.** and **Kumar R.** (2019). SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis: Association for Computational Linguistics, pp. 75–86.

Zampieri M., **Nakov P.**, **Rosenthal S.**, **Atanasova P.**, **Karadzhov G.**, **Mubarak H.**, **Derczynski L.**, **Pitenis Z.** and **Çöltekin Ç.** (2020). SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona (online): International Committee for Computational Linguistics, pp. 1425–1447.

Cite this article: Ljubešić N, Mozetič I and Kralj Novak P (2023). Quantifying the impact of context on the quality of manual hate speech annotation. *Natural Language Engineering* **29**, 1481–1494. https://doi.org/10.1017/S1351324922000353