Annotating Hate Speech: Three Schemes at Comparison

Fabio Poletto, Valerio Basile, Cristina Bosco, Viviana Patti
Dipartimento di Informatica
University of Turin
{poletto,basile,
bosco,patti}@di.unito.it

Marco Stranisci
Acmos
marco.stranisci@acmos.net

Abstract

Annotated data are essential to train and benchmark NLP systems. The reliability of the annotation, i.e. low inter-annotator disagreement, is a key factor, especially when dealing with highly subjective phenomena occurring in human language. Hate speech (HS), in particular, is intrinsically nuanced and hard to fit in any fixed scale, therefore crisp classification schemes for its annotation often show their limits. We test three annotation schemes on a corpus of HS, in order to produce more reliable data. While rating scales and best-worst-scaling are more expensive strategies for annotation, our experimental results suggest that they are worth implementing in a HS detection perspective.[1]

1 Introduction

Automated detection of hateful language and similar phenomena — such as offensive or abusive language, slurs, threats and so on — is being investigated by a fast-growing number of researchers. Modern approaches to Hate Speech (HS) detection are based on supervised classification, and therefore require large amounts of manually annotated data. Reaching acceptable levels of inter-annotator agreement on phenomena as subjective as HS is notoriously difficult. Poletto et al. (2017), for instance, report a "very low agreement" in the HS annotation of a corpus of Italian tweets, and similar annotation efforts showed similar results (Del Vigna et al., 2017; Waseem, 2016; Gitari et al., 2015; Ross et al., 2017). In an attempt to tackle the agreement issue, annotation schemes have been proposed based on numeric scales, rather than strict judgments (Kiritchenko and Mohammad, 2017). *Ranking*, rather than *rating*, has also proved to be a viable strategy to produce high-quality annotation of subjective aspects in natural language (Yannakakis et al., 2018). Our hypothesis is that binary schemes may oversimplify the target phenomenon, leaving it uniquely to the judges' subjectivity to sort less prototypical cases and likely causing higher disagreement. Rating or ranking schemes, on the other hand, are typically more complex to implement, but they could provide higher quality annotation.

A framework is first tested by annotators: inter-annotator agreement, number of missed test questions and overall opinion are some common standards against which the quality of the task can be tested. A certain degree of subjectivity and bias is intrinsic to the task, but an effective scheme should be able to channel individual interpretations into unambiguous categories.

A second reliability test involves the use of annotated data to train a classifier that assigns the same labels used by humans to previously unseen data. This process, jointly with a thorough error analysis, may help spot bias in the annotation or flaws in the dataset construction.

We aim to explore whether and how different frameworks differ in modeling HS, what problems do they pose to human annotators and how suitable they are for training. In particular, we apply a binary annotation scheme, as well as a rating scale scheme and a best-worst scale scheme, to a corpus of HS. We set up experiments in order to assess whether such schemes help achieve a lower disagreement and, ultimately, a higher quality dataset for benchmarking and for supervised learning.

The experiment we set up involves two stages: after having the same dataset annotated with three different schemes on the crowdsourcing platform Figure Eight[2], we first compare their agreement

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

rates and label distributions, then we map all schemes to a "yes/no" structure to perform a cross-validation test with a SVM classifier. We launched three separate tasks on the platform: Task 1 with a binary scheme, Task 2 with an asymmetric rating scale, and Task 3 with a best-worst scale. For each task, a subset has been previously annotated by experts within the research team, to be used as gold standard against which to evaluate contributors' trustworthiness on Figure Eight.

2 Related Work

Several frameworks have been proposed and tested so far for HS annotation, ranging from straightforward binary schemes to complex, multi-layered ones and including a variety of linguistic features. Dichotomous schemes are used, for example, by Alfina et al. (2017), Ross et al. (2017) and Gao et al. (2017) for HS, by Nithyanand et al. (2017) for offensiveness and by Hammer (2016) for violent threats. Slightly more nuanced frameworks try to highlight particular features. Davidson et al. (2017) distinguish between *hateful*, *offensive but not hateful* and *not offensive*, as do Mathur et al. (2018) who for the second type use the label *abusive* instead; similarly, Mubarak et al. (2017) use the labels *obscene*, *offensive* and *clean*. Waseem (2016) differentiate hate according to its target, using the labels *sexism*, *racism*, *both* and *none*. Nobata et al. (2016) uses a two-layer scheme, where a content can be first labeled either as *abusive* or *clean* and, if abusive, as *hate speech*, *derogatory* or *profanity*. Del Vigna et al. (2017) uses a simple scale that distinguishes between *no hate*, *weak hate* and *strong hate*.

Where to draw the line between weak and strong hate is still highly subjective but, if nothing else, the scheme avoids feebly hateful comments to be classified as not hateful (thus potentially neutral or positive) just because, strictly speaking, they can not be called HS. Other authors, such as Olteanu et al. (2018) and Fišer et al. (2017), use heavier and more elaborated schemes. Olteanu et al. (2018), in particular, experimented with a rating-based annotation scheme, reporting low agreement. Sanguinetti et al. (2018) also uses a complex scheme in which HS is annotated both for its presence (binary value) and for its intensity (1–4 rating scale). Such frameworks potentially provide valuable insights into the investigated issue, but as a downside they make the whole annotation process very time-consuming. More recently, a ranking scheme has been applied to the annotation of a small dataset of German hate speech messages (Wojatzki et al., 2018).

3 Annotation Schemes

In this section, we introduce the three annotation schemes tested in our study.

Binary. Binary annotation implies assigning a binary label to each instance. Beside HS, binary classification is common in a variety of NLP tasks and beyond. Its simplicity allows a quick manual annotation and an easy computational data processing. As a downside, such a dichotomous choice presupposes that is always possible to clearly and objectively determine what answer is true. This may be acceptable in some tasks, but it is not always the case with human language, especially for more subjective and nuanced phenomena.

Rating Scales. Rating Scales (RS) are widely used for annotation and evaluation in a variety of tasks. Likert scale is the best known (Likert, 1932): values are arranged at regular intervals on a symmetric scale, from the most to the least typical of a given concept. It is suitable for measuring subjective opinion or perception about a given topic with a variable number of options. Compared to binary scheme, scales are better for managing subjectivity and intermediate nuances of a concept. On the other hand, as pointed out by (Kiritchenko and Mohammad, 2017), they present some flaws: high inter-annotator disagreement (the more fine-grained the scale, the higher the chance of disagreement), individual inconsistencies (judges may express different values for similar items, or the same value for different items), scale region bias (judges may tend to prefer values in one part of the scale, often the middle) and fixed granularity (which may not represent the actual nuances of a concept).

Best-Worst Scaling. The Best-Worst Scaling model (BWS) is a comparative annotation process developed by Louviere and Woodworth (1991). In a nutshell, a BWS model presents annotators with *n* items at a time (where $n > 1$ and normally $n = 4$) and asks them to pick the best and worst ones with regard to a given property. The model has been used in particular by Kiritchenko

ethnic group	religion	Roma
immigrat*, immigrazione	terrorismo	rom
migrant*, profug*	terrorist*, islam	nomad*
stranier*	mus[s]ulman*	
	corano	

Table 1: List of keywords used to filter our dataset.

and Mohammad (2017) and Mohammad and Kiritchenko (2018), who proved it to be particularly effective for subjective tasks such as sentiment intensity annotation, which are intrinsically nuanced and hardly fit in any fixed scale.

4 Dataset and task description

For our experiment, we employ a dataset of 4,000 Italian tweets, extracted from a larger corpus collected within the project *Contro l'odio*[3]. For the purpose of this research, we filtered all the tweets written between November 1st and December 31st with a list of keywords. This list, reported in Table 1, is the same proposed in Poletto et al. (2017) for collecting a dataset focused on three typical targets of discrimination — namely Immigrants, Muslims and Roma.

The concept of HS underlying all three annotation tasks includes any expression based on intolerance and promoting or justifying hatred towards a given target. For each task we explicitly asked the annotators to consider only HS directed towards one of the three above-mentioned targets, ignoring other targets if present. Each message is annotated by at least three contributors. Figure Eight also report a measure of agreement computed as a Fleiss' κ weighted by a score indicating the trustworthiness of each contributor on the platform. We note, however, that the agreement measured on the three tasks is not directly comparable, since they follow different annotation schemes.

4.1 Task 1: Binary Scheme.

The first scheme is very straightforward and simply asks judges to tell whether a tweet contains HS or not. Each line will thus receive the label *HS_yes* or *HS_no*. The definition of HS is drawn by (Poletto et al., 2017). In order to be labeled as hateful, a tweet must:

- address one of above-mentioned targets;
- either incite, promote or justify hatred, violence or intolerance towards the target, or demean, dehumanise or threaten it.

[3] https://controlodio.it/.

label	tweet
yes	*Allora dobbiamo stringere la corda: pena capitale per tutti i musulmani in Europa immediatamente!* Then we have to adopt stricter measures: death penalty for all Muslims in Europe now!
no	*I migranti hanno sempre il posto e non pagano.* Migrants always get a seat and never pay.

Table 2: Annotation examples for Task 1 (gold labels).

We also provided a list of expressions that are not to be considered HS although they may seem so: for example, these include slurs and offensive expressions, slanders, and blasphemy. An example of annotation for this task is presented in Table 2.

4.2 Task 2: Unbalanced Rating Scale

This task requires judges to assign a label to each tweet on a 5-degree asymmetric scale (from 1 to -3) that encompasses the content and tone of the message as well as the writer's intention. Again, the target of the message must be one of three mentioned above. The scheme structure is reported in Table 3, while Table 4 shows an example for each label.

label	meaning
+1	positive
0	neutral, ambiguous or unclear
-1	negative and polite, dialogue-oriented attitude
-2	negative and insulting/abusive, aggressive attitude
-3	strongly negative with overt incitement to hatred, violence or discrimination, attitude oriented at attacking or demeaning the target

Table 3: Annotation scheme for Task 2: evaluate the stance or opinion expressed in each tweet.

This scale was designed with a twofold aim: to avoid a binary choice that could leave too many doubtful cases, and to split up negative contents in more precise categories, in order to distinguish different degrees of "hatefulness".

We tried not to influence annotators by matching the grades of our scale in Task 2 to widespread concepts such as stereotypes, abusive language or hateful language, which people might tend to apply by intuition rather then by following strict rules. Instead, we provided definitions as neutral and objective as possible, in order to differentiate this task from the others and avoid biases. An asymmetric scale, although unusual, fits our purpose of an in-depth investigation of negative language very well. A possible downturn of this

label	tweet
+1	*Gorino Alla fine questi profughi l'hanno scampata bella. Vi immaginate avere tali soggetti come vicini di casa?* These asylum-seekers had a narrow escape. Can you imagine having such folks (TN: racist Gorino inhabitants) as neighbours?
0	*Bellissimo post sulle cause e conseguenze dell'immigrazione, da leggere!* Great post on causes and consequences of immmigration, recommended!
-1	*I migranti hanno sempre il posto e non pagano.*
-2	*Con tutti i soldi elargiti ai rom,vedere il degrado nel quali si crogiolano,non meritano di rimanere in un paese civile!* Seeing the decay Roma people wallow in, despite all the money lavished on them, they don't deserve to stay in a civilized country!
-3	*Allora dobbiamo stringere la corda: pena capitale per tutti i musulmani in Europa immediatamente!*

Table 4: Examples of annotation for Task 2 (gold labels).

label	tweet
least	*Roma, ondata di controlli anti-borseggio in centro: arrestati 8 nomadi, 6 sono minorenni.* Rome, anti-pickpocketing patrolling in the centre: 8 nomads arrested, 6 of them are minor. *Tutti i muslims presenti in Europa rappresentano un pericolo mortale latente. L'islam è incompatibile con i valori occidentali.* All Muslims in Europe are a dormant deadly danger. Islam is incompatible with Western values. *Trieste, profughi cacciano disabile dal bus: arrivano le pattuglie di Forza Nuova sui mezzi pubblici.* Trieste, asylum-seekers throw disabled person off the bus: Forza Nuova (TN: far-right, nationalist fringe party) to patrol public transport.
most	*Unica soluzione è cacciare TUTTI i musulmani NON integrati fino alla 3a gen che si ammazzassero nei loro paesi come fanno da secoli MALATI!* Only way is to oust EVERY NON-integrated Muslim down to 3rd generation let them kill each other in their own countries as they've done for centuries INSANE!

Table 5: Examples of annotation for Task 3: 4-tuple with marks for the least hateful and the most hateful tweets.

scheme is that grades in the scale are supposed to be evenly spread, while the real phenomena they represent may not be so.

4.3 Task 3: Best-Worst Scaling

The structure of this task differs from the previous two. We created a set of tuples made up by four tweets (4-tuples), grouped so that each tweet is repeated four times in the dataset, combined with three different tweet each time. Then we provided contributors with a set of 4-tuples: for each 4-tuple they were asked to point out the most hateful and the least hateful of the four. Judges have thus seen a given tweet four times, but have had to compare it with different tweets every time[4]. This method avoids assigning a discrete value to each tweet and gathers information on their "hatefulness" by comparing them to other tweets. An example of annotation, with the least and most hateful tweets marked in a set of four, is provided in Table 5.

5 Task annotation results

In Task 1, the distribution of the labels *yes* and *no*, referred to the presence of HS, conforms to that of other similar annotated HS datasets, such as Burnap and Williams (2015) in English and Sanguinetti et al. (2018) in Italian. After applying a majority criterion to non-unanimous cases, tweets labeled as HS are around 16% of the dataset (see Figure 1). Figure Eight measures the agreement in terms of *confidence*, with a κ-like function weighted by the *trust* of each contributor, i.e., a measure of their reliability across their history on the platform. On task 1, about 70% of the tweets were associated with a confidence score of 1, while the remaining 30% follow a low-variance normal distribution around .66.

As for Task 2, label distribution tells a different story. When measuring inter-annotator agreement, the mean value between all annotations has been computed instead of using the majority criterion. Therefore, results are grouped in intervals rather than in discrete values, but we can still easily map these intervals to the original labels. As shown in Figure 1, tweets labeled as having a neutral or positive content (in green) are only around 27%, less than one third of the tweets labeled as non-hateful in Task 1. Exactly half of the whole dataset is labeled as negative but oriented to dialogue (in yellow), while 20% is labeled as negative and somewhat abusive (orange) and only less than 3% is labeled as an open incitement to hatred, violence or discrimination (red). With respect to the inter-annotator agreement, only 25% of the instances are associated with the maximum confidence score of 1, while the distribution of confidence presents a high peak around .66 and a minor peak around 0.5. Note that this confidence distribution is not directly comparable to Task 1, since the schemes are different.

In Task 3, similarly to Task 2, the result of the annotation is a real value. More precisely, we

[4]The details of the tuple generation process are explained in this blog post: `http://valeriobasile.github.io/Best-worst-scaling-and-the-clock-of-Gauss/`

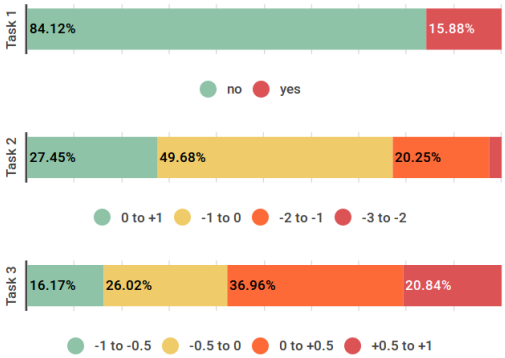

Figure 1: Label distribution for Tasks 1, 2 and 3 (red portion of Task 2 bar corresponds to 2.63%).

compute for each tweet the percentage of times it has been indicated as *best* (more indicative of HS in its tuple) and *worst* (least indicative of HS in its tuple), and compute the difference between these two values, resulting in a value between -1 (non-hateful end of the spectrum) and 1 (hateful end of the spectrum). The bottom chart in Figure 1 shows that the distribution of values given by the BWS annotation has a higher variance than the scalar case, and is skewed slightly towards the hateful side. The confidence score for Task 3 follows a similar pattern to Task 2, while being slightly higher on average, with about 40% of the tweets having confidence 1.

A last consideration concerns the cost of annotation tasks in terms of time and resources. We measured the cost of our three tasks: T1 and T2 had almost the same cost in terms of contributors retribution, but T2 required about twice the time to be completed; T3 resulted the most expensive in terms of both money and time. With nearly equal results, a strategy could be chosen instead of others for being quicker or cheaper: therefore, when designing a research strategy, we deem important not to forget this factor.

6 Classification tests with different schemes at comparison

Having described the process and results for each task, we will now observe how they affect the quality of resulting datasets. Our running hypothesis is that a better quality dataset provides better training material for a supervised classifier, thus leading to higher predictive capabilities.

Assuming that the final goal is to develop an effective system for recognizing HS, we opted to test the three schemes against the same binary classifier. In order to do so, it was necessary to make our schemes comparable without losing the information each of them gives: we mapped Task 2 and Task 3 schemes down to a binary structure, directly comparable to Task 1 scheme. For Task 2, this was done by drawing an arbitrary line that would split the scale in two. We tested different thresholds, mapping the judgements above each threshold to the label *HS_no* from Task 1 and all judgements below the threshold to the label *HS_yes*. We experimented with three values: *-0.5*, *-1.0* and *-1.5*. For Task 3, similarly, we tried setting different thresholds along the hateful end of the answers distribution spectrum (see Section 5), respectively at *0*, *0.25*, *0.5* and *0.75*. We mapped all judgements below each threshold to the label *HS_no* from Task 1 and all judgements above the threshold to the label *HS_yes*.

When considering as *HS_yes* all tweets whose average value for Task 2 is above *0.5*, the number of hateful tweets increases (25.35%); when the value is set at *-1.0*, slightly decreases (10.22%); but as soon as the threshold is moved up to *-1.5*, the number drops dramatically. A possible explanation for this is that a binary scheme is not adequate to depict the complexity of HS and forces judges to squeeze contents into a narrow black-or-white frame. Conversely, thresholds for Task 3 return different results (however partial). The threshold *0.5* is the closest to the Task 1 partition, with a similar percentage of HS (16.90%), while lower thresholds allow for much higher percentages of tweets classified as hateful — setting the value at *0*, for example, results in 40.52% of tweets classified as HS.

To better understand the impact of the different annotation strategies on the quality of the resulting datasets, we performed a cross-validation experiment. We implemented a SVM classifier using n-grams ($1 \leq N \leq 4$) as features and measuring its precision, recall and F1 score in a stratified 10-fold fashion. Results are shown in Table 6.

From the results of this cross-validation experiment, we draw some observations. When mapping the non-binary classification to a binary one, choosing an appropriate threshold has a key impact on the classifier performance. For both RS and BWS, the strictness of the threshold (i.e., how close it is to the hateful end of the spectrum) is directly proportional to the performance on the negative class (0) and inversely proportional to the

Dataset	Threshold	support (0)	support (1)	P (0)	R (0)	F1 (0)	P (1)	R (1)	F1 (1)	F1 (macro)
binary		3365	635	.878	.923	.899	.450	.316	.354	.627
RS	-0.5	2976	1014	.785	.841	.812	.408	.322	.359	.585
RS	-1.0	3581	409	.912	.966	.938	.391	.186	.250	.594
RS	-1.5	3845	145	.964	.991	.978	.200	.028	.047	.512
BWS	0.0	2206	1782	.677	.703	.690	.614	.585	.599	.644
BWS	0.25	2968	1020	.806	.860	.832	.492	.398	.439	.635
BWS	0.5	3480	508	.893	.949	.920	.390	.222	.281	.601
BWS	0.75	3835	153	.963	.992	.977	.147	.039	.060	.518

Table 6: Result of 10-fold cross-validation on datasets obtained with different annotation strategies.

performance on the positive class (1). This may be explained by different amounts of training data available: as we set a stricter threshold, we will have fewer examples for the positive class, resulting in a poorer performance, but more examples for the negative class, resulting in a more accurate classification. Yet, looking at the rightmost column, we observe how permissive thresholds return a higher overall F1-score for both RS and BWS.

Regardless of the threshold, RS appears to produce the worst performance, suggesting that reducing continuous values to crisp labels is not the best way to model the phenomenon, however accurate and pondered the labels are. Conversely, compared to the binary annotation, BWS returns higher F1-scores with permissive threshold (0.0 and 0.25), thus resulting in the best method to obtain a stable dataset. Furthermore, performances with BWS are consistently higher for the positive class (HS): considering that the task is typically framed as a *detection* task (as opposed to a *classification* task, this result confirms the potential of ranking annotation (as opposed to rating) to generate better training material for HS detection.

7 Conclusion and Future Work

We performed annotation tasks with three annotation schemes on a HS corpus, and computed inter-annotator agreement rate and label distribution for each task. We also performed cross-validation tests with the three annotated datasets, to verify the impact of the annotation schemes on the quality of the produced data.

We observed that the RS we designed seems easier to use for contributors, but its results are more complex to understand, and it returns the worst overall performance in a cross-validation test. It is especially difficult to compare it with a binary scheme, since merging labels together and mapping them down to a dichotomous choice is in contrast with the nature of the scheme itself. Furthermore, such scale necessarily oversimplifies a complex natural phenomenon, because it uses equidistant points to represent shades of meaning that may not be as evenly arranged.

Conversely, our experiment with BWS applied to HS annotation gave encouraging results. Unlike Wojatzki et al. (2018), we find that a ranking scheme is slightly better than a rating scheme, be it binary or scalar, in terms of prediction performance. As future work, we plan to investigate the extent to which such variations depend on circumstantial factors, such as how the annotation process is designed and carried out, as opposed to intrinsic properties of the annotation procedure.

The fact that similar distributions are observed when the dividing line for RS and BWS is drawn in a permissive fashion suggests that annotators tend to overuse the label *HS_yes* when they work with a binary scheme, probably because they have no milder choice. This confirms that, whatever framework is used, the issue of hateful language requires a nuanced approach that goes beyond the binary classification, being aware that an increase in complexity and resources will likely pay off in terms of more accurate and stable performances.

Acknowledgments

The work of V. Basile, C. Bosco, V. Patti is partially funded by Progetto di Ateneo/CSP 2016 *Immigrants, Hate and Prejudice in Social Media* (S1618_L2_BOSC_01) and by Italian Ministry of Labor (*Contro l'odio: tecnologie informatiche, percorsi formativi e storytelling partecipativo per combattere l'intolleranza*, avviso n.1/2017 per il finanziamento di iniziative e progetti di rilevanza nazionale ai sensi dell'art. 72 del decreto legislativo 3 luglio 2017, n. 117 - anno 2017). The work of F. Poletto is funded by Fondazione Giovanni Goria and Fondazione CRT (*Bando Talenti della Società Civile 2018*).

References

Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017. Hate speech detection in the indonesian language: A dataset and preliminary study. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 233–238. IEEE.

Pete Burnap and Matthew L. Williams. 2015. Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. *Policy & Internet*, 7(2):223–242.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*, pages 368 – 371.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate Me, Hate Me Not: Hate Speech Detection on Facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95.

Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. 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*, pages 46–51.

Lei Gao, Alexis Kuppersmith, and Ruihong Huang. 2017. Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. *arXiv preprint arXiv:1710.07394*.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Hugo Lewi Hammer. 2016. Automatic detection of hateful comments in online discussion. In *International Conference on Industrial Networks and Intelligent Systems*, pages 164–173. Springer.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470. ACL.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*, 22(140).

Jordan J Louviere and George G Woodworth. 1991. Best-worst scaling: A model for the largest difference judgments. *University of Alberta: Working Paper*.

Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018. Detecting offensive tweets in hindi-english code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26.

Saif Mohammad and Svetlana Kiritchenko. 2018. Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, pages 198–209.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56.

Rishab Nithyanand, Brian Schaffner, and Phillipa Gill. 2017. Measuring offensive speech in online political discourse. In *7th {USENIX} Workshop on Free and Open Communications on the Internet ({FOCI} 17)*.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153. International World Wide Web Conferences Steering Committee.

Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush R Varshney. 2018. The effect of extremist violence on hateful speech online. In *Twelfth International AAAI Conference on Web and Social Media*, pages 221–230.

Fabio Poletto, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, and Cristina Bosco. 2017. Hate Speech Annotation: Analysis of an Italian Twitter Corpus. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*. CEUR.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the European refugee crisis. *arXiv preprint arXiv:1701.08118*.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter Corpus of Hate Speech against Immigrants. In *Proceedings of the 11th Language Resources and Evaluation Conference 2018*, pages 2798–2805.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Michael Wojatzki, Tobias Horsmann, Darina Gold, and Torsten Zesch. 2018. Do Women Perceive Hate Differently: Examining the Relationship Between Hate Speech, Gender, and Agreement Judgments. In *Proceedings of the Conference on Natural Language Processing (KONVENS)*, pages 110–120, Vienna, Austria.

Georgios Yannakakis, Roddy Cowie, and Carlos Busso. 2018. The ordinal nature of emotions: An emerging approach. *IEEE Transactions on Affective Computing*, pages 1–20, 11. Early Access.