Do Women Perceive Hate Differently: Examining the Relationship Between Hate Speech, Gender, and Agreement Judgments

Michael Wojatzki Tobias Horsmann Darina Gold Torsten Zesch
Language Technology Lab
University of Duisburg-Essen, Germany
{michael.wojatzki,tobias.horsmann,darina.gold,torsten.zesch}@uni-due.de

Abstract

Understanding hate speech remains a significant challenge for both creating reliable datasets and automated hate speech detection. We hypothesize that being part of the targeted group or personally agreeing with an assertion substantially effects hate speech perception. To test these hypotheses, we create FEMHATE – a dataset containing 400 assertions that target women. These assertion are judged by female and male subjects for (i) how hateful these assertions are and (ii) for whether they agree with the assertions. We find that women and men consistently evaluate extreme cases of hate speech. We also find a strong relationship between hate speech and agreement judgments, showing that a low agreement score is a prerequisite for hate speech. We show how this relationship can be used for automatic hate speech detection. Our best system based on agreement judgments outperforms a baseline SVM classifier (equipped with ngrams) by a wide margin.

1 Introduction

Hate speech can be defined as hateful or threatening communication targeted towards people deserving protection. For instance *A disciplinary slap in the face never hurts, even for the (own) wife* is hate speech against women.[1] The effects of this unpleasant form of communication range from poisoning the atmosphere in social media to psychic or physical violence in the real world (Mantilla, 2013).

To counteract the massive scale to which hate speech can occur in social media, automatic methods (pre-)identifying potentially hateful or threatening utterances are required. However, even for humans, the decision whether an utterance is hate speech or not is often difficult (Ross et al., 2016; Benikova et al., 2017).

We hypothesize that hate speech perception is substantially influenced by whether one belongs to the targeted group or by whether one agrees or disagrees with the statement to evaluate. For instance, it is likely that females perceive the statement *women have lower IQ than men* as more hateful than men. Furthermore, if anyone should strangely have the supposition that women really have a lower IQ than men, then this person will likely not attribute much hate to this statement.

To study these hypotheses, we create the FEMHATE data set containing 400 German assertions about women which have been collected through a web survey.[2] As a first step, we limit ourself to self-contained, explicit statements to which we will refer to as *assertions*. Subsequently, we let 40 females and 40 males annotate (i) how hateful the assertions are and (ii) whether the subjects agree or disagree with the assertions. However, indicating the amount of hatefulness on a numerical scale is a hard task which is associated with inter-rater inconsistencies (Ross et al., 2016; Benikova et al., 2017). Thus, we use the Best–Worst-Scaling (BWS) approach by Louviere (1991) for the latter, which results in more reliable scores for other opinion-related tasks (Kiritchenko and Mohammad, 2017). The intuition underlying BWS is that although humans do not share a common absolute scale for a topic, they still tend to agree when picking the worst and best from a tuple of choices. We make all data publicly available.[3]

We find that in extreme cases of hate speech against women, it does not matter if women or men

[1] Some of the examples in this paper are insulting, distressing, or offensive to some readers. These examples do in no way corresponds to the opinion of the authors.

[2] All examples shown in this paper have been freely translated from German by the authors.

[3] github.com/muchafel/femhate

Proceedings of the 14th Conference on Natural Language Processing (KONVENS 2018)
Vienna, Austria – September 19-21, 2018

judge the assertions. For less extreme instances, however, there are clear differences e.g. when evaluating female quotas. In addition, we find a strong correlation between agreement and hate speech scores, and that an agreement score of an assertion is a necessary prerequisite for the assertion containing hate speech. This means that an assertion with a high agreement score value is most likely not hate speech. Furthermore, we show how the relationship agree/disagree judgments and hate speech score can be exploited by automatic systems for predicting hate speech. We develop an automatic approach that is based on how similarly assertions are judged by all of our subjects. We demonstrate that this approach outperforms a baseline system (SVM equipped with ngrams) by a large margin.

2 Related Work

We now shed light on hate speech research and related methods as well as the various facets that make a formalization of hate speech difficult. Furthermore, we motivate our focus on misogyny in this study.

2.1 Hate Speech in General

In recent years, the increasing number of potentially offensive, hurtful or abusive utterances in the Internet shifted into the research focus. Such utterances can be roughly summarized under the umbrella term of *hate speech* (Warner and Hirschberg, 2012; Silva et al., 2016), which is itself only vaguely defined (Schmidt and Wiegand, 2017).

In earlier work on hate speech, hate speech has been framed as abusive or hostile messages or flames (Spertus, 1997). Other commonly used terms are *abusive language* (Waseem et al., 2017) or *offensive language* (Razavi et al., 2010), as well as sub-issues such as *cyberbullying* (Xu et al., 2012) or *trolling* (Mantilla, 2013). Research on this topic focused on an analysis of such utterances and methods for an automatic detection.

2.2 Misogyny as a Form of hate speech

As noted by Mondal et al. (2017), hate speech is existent on many social media channels, resulting in many efforts to detect or eliminate hate speech and hate speech related phenomenons (Agarwal and Sureka, 2015; Bartlett et al., 2014; Gitari et al., 2015; Ting et al., 2013), often focusing on one specific form of hate speech, e.g. racism (Chaudhry, 2015; Waseem and Hovy, 2016a).

In this study, we decided to focus on another target group of hate speech - namely women.

Although there are some works on misogyny as a subform of hate speech (Mantilla, 2013; Bartlett et al., 2014; Cole, 2015), there is no dataset that serves as a gold standard for hate speech detection against women. The misogynist variant of hate speech was coined *Gendertrolling* by Mantilla (2013), which according to Mantilla (2013), is even more dangerous and destructive than regular trolling, often containing credible threats of physical and psychic violence. Bartlett et al. (2014) collected a corpus of Tweets containing terms such as *rape* and *slut* in order to analyze their usage and origin. While this is a fruitful approach in analyzing misogynist behavior, it is also limited to Tweets containing these terms.

2.3 Hate Speech and Opinions

Benikova et al. (2017) defined hate speech as expressing a very negative stance against a given target. Following this definition, we position hate speech amongst studies using opinion expressions to predict or rationalize stance (including implicit statements towards the given target) (Boltužić and Šnajder, 2014; Sobhani et al., 2015; Wojatzki and Zesch, 2016).

We hypothesize that there is a direct connection between the perception of hate speech and whether one agrees or disagrees with an assertion. For instance, whether the statement *Women cannot live up to the demands of the male olympic hundred meter run* is hate speech or not depends heavily on whether one thinks that women do not meet the requirements for the male olympic hundred meter run. Consider how this perception might change if we change *male olympic hundred meter run* to *sitting in the parliament*. If we insert *sitting in the parliament*, most people would likely disagree with the statement and the statement is probably more likely to be labeled as hate speech.

Furthermore, we hypothesize that being part of the target group influences the perception of hate speech and whether one agrees or disagrees on notions concerning this group. In a small survey with three participants, Kwok and Wang (2013) indicated that in racist targeted hate speech, race has an influence on the perception of hate speech.

2.4 Annotating Hate Speech

In the construction of a hate speech corpus, there are basically two steps: 1) collection of potential

Proceedings of the 14th Conference on Natural Language Processing (KONVENS 2018)
Vienna, Austria – September 19-21, 2018

hate speech 2) rating of these instances.

Most current studies rely on lists of offensive words and phrases for collecting potential hate speech (Mantilla, 2013; Njagi et al., 2015; Waseem and Hovy, 2016b). However, such collection inevitably brings in biases due to the limited number of query terms. For example, if one collects tweets by searching for the term *bitch*, it is not surprising that if there are hate speech annotations in this collection, it is strongly associated to this term.

As shown by Benikova et al. (2017), Waseem et al. (2017), and Ross et al. (2016), annotating hate speech using a numeric or binary scale on such data is a challenging task which is associated with low inter-annotator-agreement. Hypothesized reasons for these inconsistencies include differing thresholds from which a utterance should be classified as hateful, differing valuation of freedom of speech, and implicitness.

3 Dataset

Following the approach of Wojatzki et al. (2018a), we conduct the data collection in two steps: In the first step, we asked subjects in a web survey to generate utterances about women to which they agree and disagree, including utterances they would not make in public (as they are highly controversial or provocative). This led to a new set of assertions about women, related to women's rights, and their role in the society. In the second step, we asked 40 female and 40 male subjects in a laboratory setting to indicate (ii) how hateful the assertions are and (i) whether the subjects agree or disagree with them. For the latter we use a technique known as BWS (Louviere et al., 2015; Kiritchenko and Mohammad, 2016), which we discuss in more detail further below. The subjects received 15€ or subject hour certificates[4] as compensation for the participation. The experimental design was reviewed and approved by the ethics committee of our institution.[5] Table 1 gives an overview on the collected dataset.

3.1 Collecting Diverse Assertions

To generate a large variety of different assertions, we designed an online survey in which we directly asked participants to come up with assertions that are relevant to our topic.

	Number
Assertions	400
Agreement Judgments	32,000
BWS Judgments	4,800

Table 1: Overview on the collected dataset.

Assertion Generation To narrow the topic down for the subjects, we presented them with a list of (sub)-topics. The participants were explicitly instructed that these topics may be used as a source of inspiration for generating the assertions but that they are not limited to them. These topics include: gendered language (e.g. *waitresses* vs. *wait staff*), legal differences between men and women (laws for divorce and custody), professional life (e.g. differences in salary, leadership positions, women in the army), social roles (e.g. 'typical women's interests', women and family, 'typical women's jobs'), biological differences, and gender identity. As we wanted to generate assertions that differ in how controversial they are, we asked the subjects to provide us at least three assertions with which they personally agree and three assertions with which they disagree. On a voluntary basis we also asked the subjects to generate at least three assertions with which they personally agree, but which they would not express in public. In order to clarify the task, for each option we provided one example which takes a pro woman stance and one example which takes the opposite position. In this phase of the data collection, we do not control for any possible bias, as we aim for collecting a diverse stimulus for the subsequent rating phase. However, due to the free generation of the utterances, we are less prone to artifacts that occur in a key word based data collection (c.f. 2.4). Subjects were additionally instructed not to use expressions that indicate subjectivity (e.g. *I tend to think*), co-reference or references to other statements, and hedged statements (e.g. indicated by *maybe*, *perhaps*, or *possibly*). We removed assertions which were duplicates, not self-contained and understandable without further context, or formulated in a way that a third person cannot agree or disagree with it.

Subjects We posted the link to our survey in various online forums to ensure a wide range of opinions including communities with a thematic

[4] as needed by their study program

[5] Computer Science and Applied Cognitive Sciences at the Faculty of Engineering of the University of Duisburg-Essen (uni-due.de/kognitionspsychologie/ethikkommission_eng)

connection to the topic (e.g. the German subreddit *from women for women* `r/Weibsvolk/`) or that are expected to have a critical attitude on the subject (e.g. the Facebook group *gender mich nicht voll (don't gender me)*). Furthermore, we posted the link to topically unrelated communities such as the public Facebook group of the University of Duisburg-Essen to capture less extreme opinions.

We obtained 810 assertions from 81 participants, which means that on average each subject generated ten assertions, although only a minimum of six was required. After clean up 627 assertions remained of which we randomly subsampled 400 assertions with which we will continue to work hereinafter.

3.2 Collecting Judgments on Assertions

In a laboratory study, we let voluntary participants annotate if they agree or disagree with an assertion and which assertions in a tuple of four is the most and least hateful one.

Annotating Hate Speech To link the agreement and disagreement to the degree of hatefulness, we measure the hatefulness of each assertion. We provide the subjects with a definition of hate speech following the definition made by the Council of Europe (McGonagle, 2013)[6]: *Hate speech is when people are attacked, devalued or when hate or violence is called for against them.* It has been shown that human subjects have difficulties when annotating hate speech on a numerical scale (Ross et al., 2016). We use BWS – a comparative approach, in which each subject selects the most and least hateful assertion from a 4-tuples of assertions, which allows to rank the assertions with considerably lower effort.[7] We create 600 4-tuples using the script provided by Kiritchenko and Mohammad (2016). This script ensures that the created tuples satisfy the following constraints: (i) each 4-tuple occurs only once, (ii) each assertion occurs only once within a tuple (no duplicates) and that (iii) each assertion appears approximately in the same number as tuples as other assertions. Each tuple is answered by four female and four male subjects.

Given the comparative annotations we calculate a real value score of hatefulness for each assertion using the formulae by Orme (2009):

$$hss(a) = \%most\ hs(a) - \%least\ hs(a) \quad (1)$$

Consequently, the score ranges from -1(least hate speech) to 1 (most hate speech).

Annotating Agreement For each of the 400 assertions, all subjects had to indicate whether they personally agree or disagree with it. We do not provide an 'undecided' option to encourage subjects to take a stance. In order to make the decision as efficient as possible, we choose to let the subjects judge the assertions via arrow keys (left arrow: disagree, right arrow: agree). Thereby, we use a principle that is popular in modern applications that are considered with the evaluation of people, goods or other things (e.g. Tinder, Stylect, Jobr, or Blynk). As rating too many items in a row can be exhaustive, we split the assertions into five units containing 80 assertions each. Between each unit the subjects had to take a 60 seconds break. To prevent any effects of ordering, the assertions were presented in a random order. From the ratings, we use the percentage of times subjects agreed to an assertion as agreement score. Thus, the score lies in the range $[0..1]$ with 0 meaning everyone disagrees and 1 everyone agrees with an assertion. Assertions with a score around .5 are the most controversial in our dataset.

Subjects As we hypothesize that identification with the target group is potentially an important factor in the perception of hate speech, we systematically control for this dimension in our target group. Therefore, we collect an equally sized sample of female and male subjects. In detail, we had 40 female and 40 male subjects in the second phase of our study.[8] In order to minimize the effect of other potentially influential variables such as age or education, we tried to obtain data from a otherwise rather homogeneous group by recruiting students of the University of Duisburg-Essen. The mean age of our subjects was 23.4 years (standard deviation: 4.3). 78% of our subjects were undergraduate students, 21% graduate students and 1% had a different educational level or did not provide this information.

[6] accessible at `https://no-hate-speech.de/en/knowledge/`

[7] Each best–worst annotation consists of only two decisions, which is the most and least hateful utterances, compared to making a binary decision between each pair that could be created from a 4-tuple, i.e. for the assertions A, B, C, and D, if A is selected as best, and D is selected as worst, then we know that A $>$ B, A $>$ C, A $>$ D, B $>$ D, and C $>$ D.

[8] We also provided an option for those who do not identify with either gender. This was never chosen in our survey.

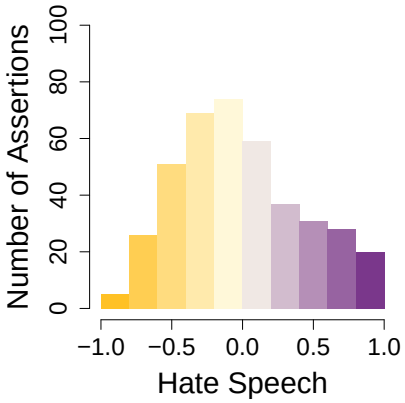


Figure 1: Distribution of hate speech score obtained using BWS. Scores are discretized into ten bins. Positive scores (gold) indicate a high degree of hatefulness and negative scores (violet) indicate a low degree of hatefulness.

4 Dataset Analysis

Now, we take a closer look at the generated scores, their distribution and the relationship between agreement and hate speech score.

4.1 Hate Speech Scores

In Figure 1, we show the histogram of the calculated hate speech score. We observe that the distribution of the scores resembles a normal distribution. The top- and bottom-ranked assertions are particularly interesting to study the phenomenon of hate speech. The three assertions ranked as most hateful are the following:

- *There are women who enjoy being raped. This kind of women form the non-lesbian portion of the goody-goody parties such as the Greens.* (hate speech score: 1.0)
- *Women belong in the kitchen.* (hate speech score: .98)
- *Women are incapable and too stupid.* (hate speech score: .98)

For people at large, these are hateful, unfounded, and invalid assertions about women fulfilling the requirement of hate speech, by belittling violence against women as well as their abilities. For both the female and the male subgroup, these three assertions are amongst the ten assertions which were rated most hateful.

The following three assertions are the three least hateful ones, and among the ten least hateful assertions for both gender subgroups:

- *Women aren't weak and shouldn't feel that way.* (hate speech score: -.84)
- *Women and men are equally intelligent.* (hate speech score: -.84)
- *Women are strong and to have them in a team makes every economic as well as social company stronger.* (hate speech score: -.89)

For most readers, these assertions do not indicate any negative sentiment towards women. Overall, assertions with a high hate speech score are easily perceived as being misogynistic, whereas low rated assertions often do not display negative sentiment towards women, which both men and women agree upon.

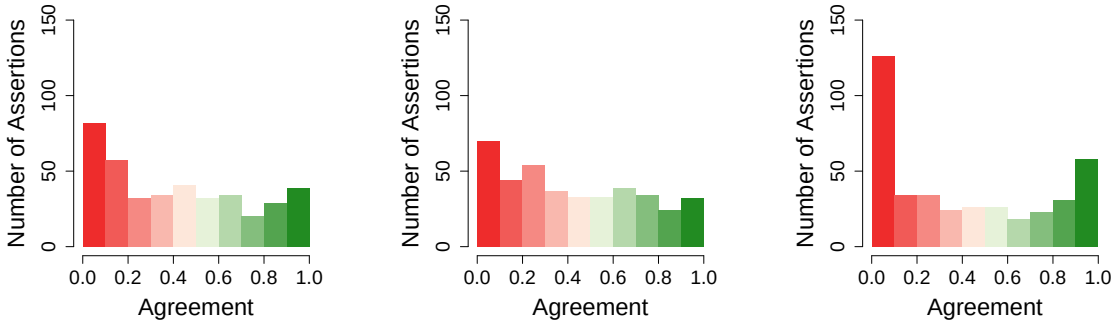
The three assertions with the biggest score difference between the female and male group are all concerned with socio-political topics – army requirements for women, gender-paygap, and structural disadvantages for men. All of these are topics that are usually discussed in political context and could be raised as such in official debates. However, there are also some assertions that were rated as more hateful by women, which are not just political, but seem to be more personal: *Feminists are man-hating women, who found no happiness in life.* (women hate speech score: .83; men hate speech score: .42), and *Gender equality actually just means "favoring women"*. (women hate speech score: .67; men hate speech score: .25). Assertions with a big score difference between men and women may contain belittlement of female activists or gender equality, but often do not display explicit hate towards women as a group.

Reliability The reliability of BWS is usually evaluated with split-half reliability (Louviere, 1993; Kiritchenko and Mohammad, 2016). Split-half reliability relies on the idea that in a consistent assessment procedure two subsets of annotators have highly correlating results. Split-half reliability is calculated by 1) randomly splitting the participants' results in two halves, 2) computing the scores for each half separately and 3) computing the Pearson correlation of the two halves' scores.

To avoid random effects, we repeat this procedure 100 times and compute the average correlation.[9] We compute the split-half reliability for the

[9] As Pearson's r is defined in a probabilistic space it can-

Proceedings of the 14th Conference on Natural Language Processing (KONVENS 2018)
Vienna, Austria – September 19-21, 2018



(a) all subjects (b) male (c) female

Figure 2: Distribution of agreement scores. Scores are discretisized into ten bins. We use a color scheme to encode how positive (green) or negative (red) the scores are.

whole group, females, and males. For the whole group we obtain a quite strong correlation coefficient of $r = .90$. The correlations of the female ($r = .82$) subjects and male ($r = .81$) subjects are significantly lower, however still substantial. Interestingly, the sexes do not differ in their consistency.

To examine the consistency of the scores of the two genders, we also compute the split-half reliability with one half being the group of males and one half being the group of females. This comparison results in a correlation coefficient of $r = .93$. This means that male and female subjects largely agree on the ranking of hate speech.

4.2 Agreement Scores

Figure 2 shows the histograms of the calculated agreement scores for the full set of subjects, for females, and for males. For the whole set of subjects, we observe that the scores are rather evenly distributed across the range of possible agreement scores. The mass of the distribution is slightly bigger in the negative range of scores. This means that there are more assertions to which the majority disagrees than assertions to which the majority agrees. Accordingly, the mean agreement score over all assertions for the whole group is .42.

For the male subjects the scores are distributed even more evenly and the mean agreement score is slightly higher (.44). As this score is closer to .5, we conclude that the assertions are more controversial for male subjects. In contrast, for female subjects we observe a higher concentration at scores 1.0 and 0.0. However, the number of assertions with a 0.0 score is clearly bigger than the number of assertions with a 1.0 score. The mean agreement score is slightly lower than the score of the whole group (.41). It is possible that women, being the target of the potential hate speech, are more affected and thus more extreme in their judgment.

Reliability Analogous to the comparative hate speech annotations, we calculate the split half reliability for the agreement scores. For the whole dataset, we obtain a correlation coefficient of $r = .96$. Therefore, the agreement scores can be regarded as being robust. The coefficients of the male ($r = .92$) and female ($r = .95$) group are only slightly lower and thus also quite robust. The high correlations could be regarded as an indication of the homogeneity of the sample collected.

Furthermore, we compute the correlation between the scores of the female and male participants, and obtain coefficient of $r = .83$. This means that although there is a clear correlation of the resulting scores, there are also substantial differences between the genders. 34% of the items that the female and male participant group had a judgment disagreement on contained the term "female quota", while only 2% of the ones which did not cause disagreement between the groups contained the term. The largest difference in the agreement score was between the assertions *Female quotas are nonsense* (women agreement score: 0.10 hate speech score: 0.38; men agreement score: 0.53 hate speech score: 0.21) and *Female quotas are cosmetic restraints and constitute a form of discrimination* (women agreement score: 0.22 hate speech score: 0.25; men agreement score: 0.65 hate speech score: 0.42). In contrast to the general trend of assertions

not be averaged directly. Therefore, we first z-transform the scores, average them and then transform them back into the original range of values.

with a high hate speech score having a low agreement score, both assertions have relatively high hate speech scores for both subgroups while the male group disagrees with the assertions.

4.3 Relationship between Agreement and Hate Speech

Now, we analyze the relationship between agreement and hate speech score. Figure 3 shows the comparison of agreement and hate speech score for the whole sample, the female and the male subjects. We observe the clearest correlation for the scores of the female subjects and and the least clear correlation for the male group. This is also reflected in the correlation coefficients: $r = -.79$ for females, $r = -.67$ for males, and $r = -.76$ for the whole group.

Interestingly, there are no cases with a high agreement score and also high hate speech score. As mentioned in the previous section, there are some assertions with a moderately high hate speech score that were agreed with. In contrast, there are cases with low agreement and low hate speech score, which seem to be too much of a cliché, for instance: *Men have to like football*.

5 Predicting Hate Speech Using Judgment Similarity

In this section, we investigate if we can automatically predict hate speech scores. So far, we showed a strong relationship between the (dis-)agreement to an assertion and the resulting hate speech score. Now, we investigate if this relationship is strong enough for an automatic prediction of the hate speech score. We hypothesize that it is possible to approximate the hate speech score by predicting if two assertions are judged similarly by a number of people. One finds numerous agree and disagree judgements in social media[10] indicating how different people judge a posting (assertion). If the similarity between agree and disagreement judgements enables an approximation of the hate speech score, the judgements found in social media offer an inexpensive source of new judgements for an easy identification of hate speech. Consequently, for predicting the assertion score, we (1) calculate how similarly an assertion is judged compared to other assertions and (2) base our prediction on the most similar assertions.

[10]e.g. thumb-ups and thumb-downs, or likes and dislikes

The first step is the calculation of judgement similarity between two assertions. We use the agreement matrix in which rows represent the subjects and columns represent the assertions to compute the cosine as similarity measure. Hence, judgement similarity (JS) is defined as:

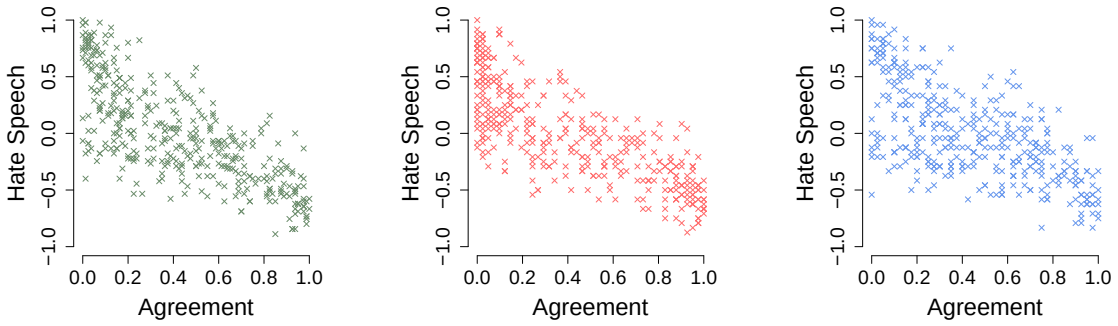
$$JS(a1,a2) = \frac{\vec{a_1} \cdot \vec{a_2}}{|\vec{a_1}| \cdot |\vec{a_2}|} \quad (2)$$

where $\vec{a_1}$ is the vector representing all judgments on a_1 and $\vec{a_2}$ representing all judgments on a_2. Similarly judgment of two assertions a_1 and a_2 can have several reasons. These include that a_1 and a_2 have a high semantic text similarity (Agirre et al., 2012), are paraphrases (Bhagat and Hovy, 2013), are in any entailment relationship (Dagan et al., 2009), or that underlying social, personal or other reasons result in a correlation of judgments.

In the second step, we try to predict the hate speech score of a new assertion a_1 based on the assertions that have the highest judgment similarity with a_1. Therefore, we implement an SVM-based approach that uses the hate speech scores of the most similar assertion as features. As a reference approach, we implement an SVM-regression equipped with 1-3 gram features – a system that yields highly competitive performance for hate speech detection (Waseem and Hovy, 2016a; Benikova et al., 2017). To generalize to unseen assertions, we approximate judgement similarity by using a Siamese Neural Network (SNN) that solely relies on text features.

5.1 Automatically Estimating Judgment Similarity

Our approach requires a large number of agreement and disagreement judgments on assertions for which we want to predict a hate speech score. However, such knowledge is – even in rich social media data scenarios – not always present, e.g. if we want to process a completely new assertion that has not been evaluated by a sufficient number of people. To overcome this limitation, we follow the approach of Wojatzki et al. (2018b) and train a system that is able to estimate the judgement similarity of two assertions from their texts. Therefore, we implement an SNN – a neural network architecture that is well suited to learn text similarity (Mueller and Thyagarajan, 2016; Neculoiu et al., 2016) or the connection of pairs of sentences (e.g. replies to tweets) (Hu et al., 2014).



(a) all subjects ($r = -.76$) (b) female ($r = -.79$) (c) male ($r = -.67$)

Figure 3: Comparison of agreement and hate speech scores according to gender.

SNNs consist of two identical subnetworks and a final merge-layer that merges them. Each subnet receives one assertion and tries to extract useful representations of the assertions. In our implementation, one subnet consists of an embedding layer, a convolution layer with a filter size of two, a max pooling over time layer, and a dense layer with 100 nodes. For merging the subnets, we calculate the cosine between the resulting representations. We created the architecture within the deep learning framework DeepTC (Horsmann and Zesch, 2018) with Tensorflow (Abadi et al., 2016) in the backend. In a 10-fold cross validation experiment, we obtain a Pearson correlation between the predicted and the gold similarity of $r = .72$.

5.2 Evaluation

We evaluate the systems based on gold and SNN judgement similarity again in a 10-fold cross-validation experiment using Pearson's r as a performance metric. To enable a fair comparison, the judgment similarity systems are also implemented as an SVM-regression, which uses the score of the n most similar assertions in the respective training set. We also compare the performance between using only the score of the most similar assertion in comparison to using the scores of n most similar assertions. Furthermore, we experiment with using both the hate speech score and the agreement score of the most similar assertion.

Table 2, shows the results of this experiment. As expected with 400 assertions, a direct prediction based on ngrams shows a low performance of $r = .35$. All approaches based on *gold* judgment similarity score significantly better. For the approaches based on *SNN* judgment similarity, we observe substantial differences depending on which n we use. We see substantial improvements between the $n = 1$ and $n = 25$ most similar assertions but see only moderate changes for even larger n. As expected, it is advantageous to use the hate speech score over the agreement score for both, the *SNN* and *gold* setup.

Feature Set	**Score Type**	***n* most similar**	***Pearson's r***
ngrams (1-3)	-	-	.35
judgment similarity *SNN*	*agreement score*	1	.25
		25	.39
		50	.48
		75	.49
	hate speech score	1	.25
		25	.52
		50	.54
		75	.53
judgment similarity *gold*	*agreement score*	1	.70
		25	.70
		50	.63
		75	.61
	hate speech score	1	.75
		25	.67
		50	.65
		75	.65

Table 2: Performance of different systems for predicting hate speech.

The result underline that for the automatic detection of hate speech it seems highly promising to look on assertions that are judged similarly by several people. Such judgments could e.g. easily be gained from the likes and dislikes on posts in social network sites. In practice, once one has identified a few statements that clearly contain hate speech, one can automatically identify other hate speech messages by exploiting judgment similarity.

Our evaluation also demonstrated that an approach based on judgement similarity can be approximated by the mere text of the assertions.

6 Conclusion & Future Work

In this paper we present the FEMHATE dataset which contains 400 assertions that have been collected via crowdsourcing and that have subsequently been judged by 80 subjects (40 female and 40 male). We collected 32,000 judgments indicating whether the subjects agree or disagree with the statements and 4,800 judgments that indicate the strength of contained hate speech. The ratings were shown to be reliable. We were able to show that people never agree with assertions they themselves consider to be hate speech.

Furthermore, we could show that the agreement with the assertions addressing gender debates as well as their comparative rating are relatively similar and robust throughout gender. Although there are cases of great disagreement, they are not cases of highly rated misogyny, neither by men nor women. In this way, we could provide evidence for the hypothesis that on both poles of the range of hate speech scores there is a high agreement between the male and female subjects. Hence, for cases of extreme misogyny, it is irrelevant whether men or women rate it. We also show that the strong relationship between agreement/disagreement and the amount of hate speech can successfully be exploited by automatic systems that try to predict hate speech. If we transfer the scores of assertions that have been similarly judged by by a large amount of people, we obtain dramatic gains over a baseline system. We envision that such a system could be used on real life social media data, which are rich of judgments on assertions (e.g. thumbs up/thumbs down on YouTube comments).

In future work, we plan to further examine the relationship between hate speech, agreement and judgment similarity by applying a typology to the relations (e.g. if a_1 is a more implicit form of a_2) and the assertions (e.g. if a_1 is a threat or insult). In addition, we plan to examine if the relation between hate speech and agreement can also be utilized to boos performance of automatic hate speech detection in real world social media data. Therefore, we suggest to make use of agreement and disagreement judgments that are readily available in large quantities (e.g. up-votes or down-votes of forum posts).

Acknowledgments

We would like to thank everyone who participated in our studies and Mara Ortmann for her support during the laboratory study. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media".

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. TensorFlow: A System for Large-scale Machine Learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI'16, pages 265–283, Savannah, USA.

Swati Agarwal and Ashish Sureka. 2015. Using KNN and SVM based one-class Classifier for Detecting Online Radicalization on Twitter. In *International Conference on Distributed Computing and Internet Technology*, pages 431–442. Springer.

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 Task 6: A pilot on Semantic Textual Similarity. In *Proceedings of the SemEval*, pages 385–393, Montreal, Canada.

Jamie Bartlett, Jeremy Reffin, Noelle Rumball, and Sarah Williamson. 2014. Anti-social media. *Demos*, pages 1–51.

Darina Benikova, Michael Wojatzki, and Torsten Zesch. 2017. What does this imply? Examining the Impact of Implicitness on the Perception of Hate Speech. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology*, pages 171 – 179, Berlin, Germany.

Rahul Bhagat and Eduard Hovy. 2013. What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Filip Boltužić and Jan Šnajder. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58, Baltimore, USA.

Irfan Chaudhry. 2015. # hashtagging hate: Using twitter to track racism online. *First Monday*, 20(2).

Kirsti K Cole. 2015. It's like she's eager to be verbally abuseda: Twitter, trolls, and (en) gendering disciplinary rhetoric. *Feminist Media Studies*, 15(2):356–358.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 15(4):i–xvii.

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.

Tobias Horsmann and Torsten Zesch. 2018. DeepTC – An Extension of DKPro Text Classification for Fostering Reproducibility of Deep Learning Experiments. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 2539–2545, Miyazaki, Japan.

Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional Neural Network Architectures for Matching Natural Language Sentences. In *Advances in Neural Information Processing Systems (NIPS 27)*, pages 2042–2050, Montreal, Canada.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing Reliable Fine-Grained Sentiment Associations by Crowdsourcing and Best–Worst Scaling. In *Proceedings of The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 811–817, San Diego, California.

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 Annual Meeting of the Association for Computational Linguistics*, pages 465–470, Vancouver, Canada. ACL.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, pages 1621–1622, Washington, USA.

Jordan J. Louviere, Terry N. Flynn, and A. A. J. Marley. 2015. *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press.

Jordan J. Louviere. 1991. Best-worst scaling: A model for the largest difference judgments. Working Paper.

Jordan J. Louviere. 1993. The best-worst or maximum difference measurement model: Applications to behavioral research in marketing. In *The American Marketing Association's Behavioral Research Conference*, Phoenix, Arizona.

Karla Mantilla. 2013. Gendertrolling: Misogyny adapts to new media. *Feminist Studies*, 39(2):563–570.

Tarlach McGonagle. 2013. The council of europe against online hate speech: Conundrums and challenges. *Expert paper, doc. no*, 1900(2013):005.

Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. A Measurement Study of Hate Speech in Social Media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, HT '17, pages 85–94, New York, NY, USA. ACM.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese Recurrent Architectures for Learning Sentence Similarity. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 2786–2792, Phoenix, USA.

Paul Neculoiu, Maarten Versteegh, and Mihai Rotaru. 2016. Learning text similarity with siamese recurrent networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 148–157.

Dennis Njagi, Z Zuping, Damien Hanyurwimfura, and Jun Long. 2015. A Lexicon-based Approach for Hate Speech Detection. In *International Journal of Multimedia and Ubiquitous Engineering*, volume 10, pages 215–230.

Bryan Orme. 2009. Maxdiff analysis: Simple counting, individual-level logit, and HB. *Sawtooth Software*.

Amir H. Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In *Proceedings of the 23rd Canadian Conference on Advances in Artificial Intelligence*, pages 16–27, Berlin, Heidelberg. Springer-Verlag.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, pages 6–9.

Anna Schmidt and Michael Wiegand. 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*, pages 1–10. Association for Computational Linguistics.

Leandro Araújo Silva, Mainack Mondal, Correa Denzil, and Fabrício Benevenuto. 2016. Analyzing the Targets of Hate in Online Social Media. In *Proceedings of the Tenth International Conference on Web and Social Media*, pages 687–690.

Parinaz Sobhani, Diana Inkpen, and Stan Matwin. 2015. From argumentation mining to stance classification. In *Proceedings of the NAACL HLT 2015*, pages 67–77, Denver, USA.

Ellen Spertus. 1997. Smokey: Automatic recognition of hostile messages. In *AAAI/IAAI*, pages 1058–1065.

I-Hsien Ting, Hsing-Miao Chi, Jyun-Sing Wu, and Shyue-Liang Wang. 2013. An approach for hate groups detection in facebook. In *The 3rd International Workshop on Intelligent Data Analysis and Management*, pages 101–106. Springer.

William Warner and Julia Hirschberg. 2012. Detecting Hate Speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016a. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of NAACL-HLT*, pages 88–93.

Zeerak Waseem and Dirk Hovy. 2016b. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June. Association for Computational Linguistics.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84. Association for Computational Linguistics.

Michael Wojatzki and Torsten Zesch. 2016. Stance-based Argument Mining - Modeling Implicit Argumentation Using Stance. In *Proceedings of the KONVENS*, pages 313–322, Bochum, Germany.

Michael Wojatzki, Saif M. Mohammad, Torsten Zesch, and Svetlana Kiritchenko. 2018a. Quantifying Qualitative Data for Understanding Controversial Issues. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1405 – 1418, Miyazaki, Japan.

Michael Wojatzki, Torsten Zesch, Saif M. Mohammad, and Svetlana Kiritchenko. 2018b. Agree or disagree: Predicting judgments on nuanced assertions. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM)*, New Orleans, USA.

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from Bullying Traces in Social Media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 656–666, Stroudsburg, PA, USA. Association for Computational Linguistics.