Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)

2019 GermEval Shared Task on Offensive Tweet Detection h_da submission

Isabell Börner* **Midhad Blazevic*** **Maximilian Komander*** **Margot Mieskes†**
University of Applied Sciences, Darmstadt, Germany
*firstname.lastname@stud.h-da.de
†firstname.lastname@h-da.de

Abstract

This paper presents the models submitted to the 2019 GermEval Shared Task on Offensive Language Detection in Tweets. Our system is based on a lexicon of swear words and several rules. These rules were developed after a thorough data and error analysis. This also revealed that the detection of offensive language is far from trivial and in a lot of cases requires more than just a Tweet in isolation, but rather would require more knowledge about the context and/or the topic the Tweet is related to, which was not available in this data set.

1 Introduction

"Offensive language is commonly defined as hurtful, derogatory or obscene comments made by one person to another person. This type of language can be increasingly found on the web." With these words Wiegand et al. (2018) introduced the 2018 edition of the GermEval 2018 Shared Task on the identfication of offensive language. While this indicates an academic interest in the topic, the German Netzdurchsetzungsgesetz (NetzDG) requires social networks to remove illegal content (Smedt and Jaki, 2018) which might overlap with offensive language in general. Recent events surrounding the murder of a German politician in June 2019, police forces look into social media containing hate speech (German "Hasskommentare") related to this event.[1] Additionally, there is very little work on German hate speech, as opposed to English hate speech and/or offensive language.

This paper presents the description of the system submitted by the University of Applied Sciences, Darmstadt (h_da) to the GermEval 2019 edition of the shared task on detecting offensive language in Tweets. While most systems in the 2018 edition used machine or deep learning, we created a rule-based system after performing a thorough data analysis.[2] While a range of our observations could be translated into features for machine learning, this was not the main focus of this work. Similar to (Klenner, 2018) we observe that the annotations are not as clear, as the annotations suggest. Accordingly, we feel (similar to (Smedt and Jaki, 2018)) that releasing AI without a proper verification is ethically critical. Therefore, we suggest to use confidence scores, rather than absolute annotations to indicate the potential label and to also have a closer look at the manual annotations, which are not always as clear-cut as they might seem. Especially in isolation not all annotations are comprehensible and might need some further discussion.[3]

2 Data Analysis

Initially, we thoroughly looked at the 2018 and 2019 data sets in order to gain a better intuition for the material we are dealing with. It became obvious, that many offensive tweets have one common ground: they use offensive language to offend certain people, institutions, countries or companies. Our idea was, that a script could classify tweets by looking for "bad" language inside the tweets and thus categorize them as either OFFENSIVE or OTHER. The basis for what we consider bad language, is a list of words found at: http://www.insult.wiki/wiki/Schimpfwort-Liste. Our error analysis revealed that the classification contained too many mistakes. We therefore removed words such as "Ameise" (ant), "Vielflieger" (frequent flyer) or "Bär" (bear) which do not have negative connota-

[1] https://www.zeit.de/gesellschaft/zeitgeschehen/2019-08/walter-luebcke-hasskommentare-internet

[2] Details of our system are available at https://github.com/mieskes/germEval2019

[3] We present examples taken from the data in German and provide a rough translation into English.

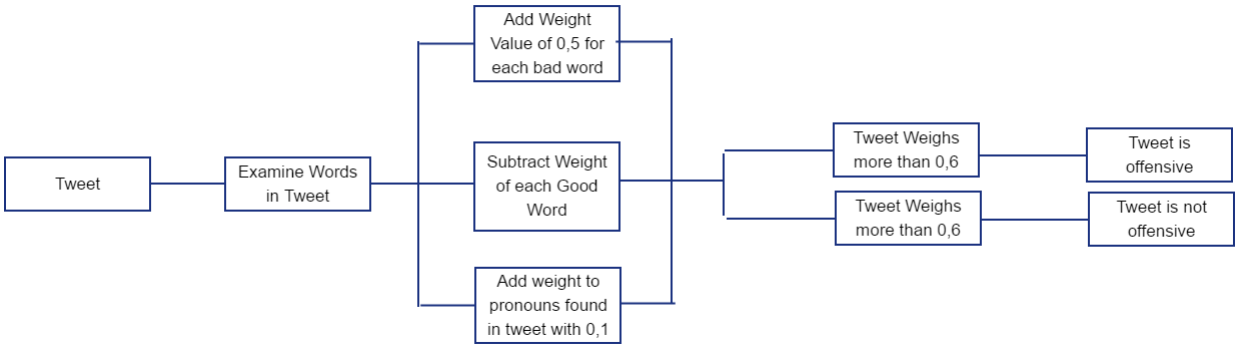

Figure 1: Architecture of the h_da lexicon- and rule-based system.

tions.

In another step, we performed an error analysis, looking in detail at the mis-classified Tweets. Errors came in two forms: One set of errors is based on the mis-classification of our method. Another set of errors can be attributed to annotations that are less clear-cut. For example, there was a tweet in which one person stated, that he wishes the old german anthem back and also the time (WW2-era), too. This tweet has not been officially classified as OFFENSIVE, while we came to the conclusion that it was indeed, offensive. This naturally leads to problems, as our script classified some tweets as OFFENSIVE because it contained those offensive words and insults, while the official file marked them as OTHER. To increase the accuracy, we looked up the tweets and gathered offensive words that our own list did not contain at this time.

As the accuracy did not incease significantly, we decided to add weights to the words in our lists. In addition, we observed that Tweets contained words which might not be offensive as such (i.e. "Hund" (english: dog), but changes to being offensive if combined with a pronoun and/or an (offensive) adjective. A sentence such as "Ein hässlicher Hund" (An ugly dog) becomes offensive in the case of "Du hässlicher Hund" (You ugly dog). We therefore added weights based on a word being in the word list, occurring with an pronoun and with an adjective.

3 Experimental Setup

The first phase consisted of using a "badword list" to identify tweets that are offensive. Our system compares the words in a tweet to the words that can be found in the "badword list", and if a tweet has one of these badwords then is considered offensive. This simple comparison provided mixed results due to the "badword list" not being optimal, due to words within the list that might or might not be considered as bad or offensive words, depending on context. The second phase (shown in Figure 1 above) involved adding weights to identify offensive tweets, by analyzing as many words of the tweet as possible and using the end weight to identify if a tweet is offensive or not. The "badword list" from the first phase is being used, and all of the words that can be found in the list are weighted as +0,5. Pronouns are also weighted due to the fact that many hate speech tweets consist of, for example, a person being attacked directly by using the word "du" (you). Pronouns are currently in this stage weighted at +0,1.[4] In order to further enhance the analysis, we used SentiWS_v2.0 (Remus et al., 2010) lists to optimize the analysis by also using positive words, along with SentiWS' value of these positive words to minimize the weights of the tweets. Furthermore, we thought of using the negative word list as well, but it misses swear words. Our own "badword list" is also being further developed. In version 2.0 the list will be newly created by real people via Google Survey, which has been sent to different people from all ages and sexes. The overall weighting system will also be fitted later on, as we proceed.

4 Results

Our system is primarily based on the list of insults as described in Section 3 above. The model looks for every bad word in the selected tweet and thus makes an assumption about its polarity. Evaluating the first runs, we notice that the classification was

[4] An experimental analysis of the weights was not possible due to time constraints.

	Average			Offense			Other		
system	p	r	f	p	r	f	p	r	f
test 2018 weighted	50.12	50.13	50.12	34.16	**43.34**	38.21	**66.07**	56.91	61.15
test 2018 unweighted	48.22	48.45	48.33	**31.42**	24.54	27.56	65.02	**72.36**	68.49
train 2019 weighted	54.97	54.63	54.35	36.64	**51.98**	42.98	**71.51**	57.27	63.60
train 2019 unweighted	57.55	55.87	56.70	**44.05**	29.60	35.41	71.05	**82.13**	76.19
final	50.12	50.13	50.12	34.16	**43.34**	38.21	**66.07**	56.91	61.15
test 2019 (official; run 1)	59.60	58.24	58.91	**46.42**	36.08	40.60	72.77	**80.40**	76.39
test 2019 (official; run 2)	54.55	58.24	54.87	36.95	**52.68**	43.43	**72.15**	57.69	64.11
post-evaluation	58.09	56.39	57.23	44.81	30.85	36.54	71.37	81.94	76.29

Table 1: Results for various variants of our system.

prone to mistakes, as our badwordlist contained too many insults and slurs that on the other hand were used in non-offensive tweets and thus resulting in false positives, with an accuracy under 50 %. We therefore reduced the amount of bad words in our list from about 2000 to 1520 to increase accuracy.

Based on our error anlysis (described in Section 2) we add weights to the bad words and pronouns. A bad word receives a weight of 0.5 and selected pronouns a weight of 0.1. If the tweet has a weight of at least 0.6 it is considered offensive. The pronouns we include are ”ihr”, ”du”, ”sie”, ”dich”, ”euer”, ”ihrer”, ”deren” and ”dein”.

Table 1 shows the results of our systems on various data sets including the official test evaluation results. We observe that the weighted system consistently has higher Recall results when labelling a Tweet as OFFENSE, whereas it achieves higher Precision when labelling a Tweet as OTHER. The unweighted model shows higher Precision for OFFENSE and higher Recall for OTHER. As recognizing an offensive Tweet is a critical task, from several points of view, it is desirable to achieve a higher Recall in order to ensure that a Tweet labelled as offensive is actually offensive.

We also combined the two models during the post-evaluation analysis, which increased the performance on average and also in both categories. The combined model takes the output of both models. In case the models agreed the decision was used. For non-animous decisions the weighted model decided for the OFFENSE category and the unweighted model for the OTHER category.

5 Error Analysis

After the gold labels for the test data were released, we performed a detailed round of error analysis on the actual test data. The tweets themselves prove to be a challenge. Many tweet labels are not clear, and thus even though a tweet is labeled as offensive or abusive, we do not consider every offensive labeled tweet to be offensive. We have found tweets in which a simple figure of speech such as “Ich glaub ich muss kotzen” (I think, I have to throw up) is considered offensive. Our system also found these tweets to not be offensive, and this is in our opinion correct. Another example: a tweet has been marked as OFFENSE INSULT with the content “Diese Studenten, die ihren Studienausweis zücken, bevor der Kontrolleur kommt” (Those students, who take out their student id before an inspector shows up), which in our opinion does not represent Hate-Speech at all. While for these cases contexts can be imagined, where such an utterance could be considered hate speech, others, such as “Seit wann magst du Kartoffeln?” (Since when do you like potatoes?) or “Bratkartoffeln aus rohen Kartoffeln best, aber verdammt immer eine Riesensauerei” (Hash Browns out of raw potatoes are the best, but that sure means a big mess) it is harder to imagine a context where these utterances could be considered offensive.

Even more challenging is how labeling occurred when looking closer at tweets that can be considered political statements, in which no person or entity is directly harmed. Some tweets can even be considered sarcastic with reference to the past. These sarcastic comments are not positive but also do not attack a person directly. Also, it is still an open question whether sarcastic or ironic comments are necessarily considered Hate Speech, as in the context of political comedy these methods are frequently used. But other tweets that are targeted towards specific groups, have not been labelled as offensive, while we came to the conclusion that they could probably be considered offensive, such

as “Klar. Danach kannste dir direkt umsonst auch noch Schläge abholen” (Sure. Afterwards you can get a beating for free).

The context behind the tweet, which is missing in the labeled data, is nonexistent for us, and for our system. This leads to problems such as incorrect labeling by the system. A tweet that is labeled by the annotators as offensive due to context, cannot always be labeled as offensive by the system. Since a simple application cannot dive deeper into context, many tweets could not be analyzed correctly.

During the manual inspection of our systems results, we decided to add words such as “Jude” (jude), “Moslem” (muslim), etc. to the first list, due to our system missing offensive labeled tweets which had these words. These words are regrettably misused for offensive purposes. Using these words in our badword list does create false positives but improves results. Examples, where we found that a tweet was not labelled as offensive, but could be considered offensive towards muslims is a tweet like: “Hey, das war ausschließlich gegen Muslime gerichtet, halb so wild!” (Hej, this was only targeted towards Muslims, no big deal!).

Using a badword list for comparison and identifying bad or offensive tweets has also proven to be difficult. We have tested our system with two different lists. The first list consists of 1.520 words. The second list consists of 11.303 words that can be used as offensive words. The overall results using the first list were better than when using the second list. The second list seemed to have falsely labeled too many tweets as offensive. Overall, we consider smaller lists that have good quality to be better than extensive list, thus quality goes over quantity.

6 Discussion & Conclusions

While our system does not outperform the others, we think that the analyses we carried out during the project are quite valuable. Additionally, these analyses indicate, that the classification of Tweets at least in most cases requires contextual and/or meta information. There are a range of cases, where it is easy to imagine, that a context might exist, which renders a Tweet harmless or harmful. Without information about previous Tweets, the topic, the Tweet under consideration refers to, it is hard to be absolutely sure.

Nevertheless, we see a range of options to improve our system. One of the first steps is, rather than relying on fixed sets of words, such as the list of pronouns, some more linguistic preprocessing, such as Part-of-Speech tagging might prove useful. Additionaly, our findings could be incorporated in a Machine Learning setup, which would benefit the overall precision/recall values.

Also, the definition of hate speech was in some cases quite strict. Several tweets have been officially classified as OFFENSE although no hate speech or offensive language could be detected by our group. We do not consider simple sarcasm or irony as hate-speech, which also results in lower accuracy rates.

On a more general note, the task of identifying offensive language has to walk a very fine line between targetting offensive language, which might also be illegal, as in the case of the German “Volksverhetzung” (incitement of the people) and censorship. Thus, from ethical point of view, we should be careful about how strict our definition of offensive language is and what has to be accepted under the freedom of speech.

References

Manfred Klenner. 2018. Offensive language without offensive words (olwow). In *Proceedings of the GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018) Vienna, Austria, September 21, 2018*, pages 11–15.

R. Remus, U. Quasthoff, and G. Heyer. 2010. SentiWS - a Publicly Available German-language Resource for Sentiment Analysis. In *Proceedings of the 7th International Language Ressources and Evaluation (LREC'10)*, pages 1168–1171.

Tom De Smedt and Sylvia Jaki. 2018. Challenges of automatically detecting offensive language online: Participation paper for the germeval shared task 2018 (haua). In *Proceedings of the GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018) Vienna, Austria, September 21, 2018*, pages 11–15.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. In *Proceedings of the GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018) Vienna, Austria, September 21, 2018*, pages 1–10.