Research Group Social Computing
Department of Informatics
Technical University of Munich

Hate Speech in Telegram Subnetworks

Daniel Bartmann[1] ✉ **and Burak Enes Cakici**[2] ✉

1Department of Informatics, Technical University of Munich

2Department of Informatics, Technical University of Munich

✉ bartmann@in.tum.de
✉ burak-enes.cakici@tum.de

August 3, 2021

Abstract — The Telegram communication platform is becoming increasingly popular with extremist groups and conspiracy theorists, making hate speech in Telegram a growing problem. To combat hate speech, neural networks are usually used to detect it automatically. In order to train these machine learning models, however, large amounts of labeled data are required, which are not yet sufficiently available for German Telegram messages. Therefore, we investigate in this paper to what extent models trained on Twitter data can be applied to Telegram messages. We train a total of eleven models on different Twitter datasets, introduce a methodology to classify messages without character restriction, and evaluate the models on a newly created Telegram test set consisting of 1149 messages.

1 Introduction

Online Hate Speech has become one of the biggest problems on the Internet today, especially on social media platforms (Siegel, 2020). Insults, anti-Semitic, racist, and sexist posts are ubiquitous, fueling extremism and often leading to offline violence. Williams et al have shown in their work that hate speech, which is directed against race or religion, and physical violence are directly linked (Williams, Burnap, Javed, Liu, & Ozalp, 2020). Now politicians have also realized the threat posed by hate on the Internet and are focusing their efforts on the operators of social media platforms, who are supposed to detect and block hate comments. Since millions of new comments are being posted every day, it is impossible to check them by hand. Machine learning is therefore increasingly being used to automatically detect hate speech.

But the quality of the classifier is significantly related to the training data, it can only learn the patterns of hate speech that are present in the training data. Due to the diversity of hate speech and ever emerging hate-related topics, such as the new Covid-19 topic, the generalizability of the classifiers is limited. However, not only the different topics, but also the different platforms can potentially reduce the quality of the classifiers. For example, tweets with the typical use of numerous hashtags and the 280-character limit often have a different structure than non-limited messages on other platforms.

Especially in recent years, the Telegram platform has been gaining popularity in the extremist scene due to its strict privacy policy (Walther & McCoy, 2021) (Semenzin & Bainotti, 2020). It has been shown that Telegram played a central role in the spread of ISIS propaganda (Hughes & Meleagrou-Hitchens, 2017). And in Germany, too, more and more channels of the right-wing movement or conspiracy theorists are emerging. To combat the spread of hate in these channels, you need a lot of such labeled messages. However, there are only very few German hate speech datasets and even fewer datasets based on Telegram messages. Therefore, we want to investigate to what extent machine learning models trained on different Twitter datasets are able to detect hate speech in Telegram messages.

2 Related Work

Within the field of hate speech detection on social media, there has been a great focus on Twitter due to its popularity and suitability to express opinions which can sometimes be offensive or hateful. For German, the shared task series Germeval (Wiegand, Siegel, & Ruppenhofer, 2018; Struß, Siegel, Ruppenhofer, Wiegand, & Klenner, 2019) and HASOC (Mandl et al., 2019; Mandl, Modha, Kumar M, & Chakravarthi, 2020) provided the literature with annotated tweet collections which became the base of our training process in this study.

On the other hand, Telegram has gained a widespread use only recently. In English, there have been little research that investigates the presence of the hateful language and that attempts to create a corpus (Scheffler, Solopova, & Popa-Wyatt, 2021). However, there is no significant work in German.

Regarding our research question, i.e. applicability of classifiers across platforms, there are several stud-

ies on other languages. Fortuna, Bonavita, and Nunes (2018) investigate the issue in Italian on Facebook and Twitter where they use a simple dense neural network instead of state-of-the-art models. For the classifiers with training and test sets coming from the same social media platform, macro F1 scores of 0.64 and 0.72 were achieved, respectively for Facebook and Twitter. However, there is a significant decrease in cross-platform tests. When the classifier is trained on one of the platforms and tested on the other, the macro F1 score drops to 0.44 (for both versions). A similar experiment is done for English by Markov and Daelemans (2021) across the same social media platforms. In-platform tests yielded F1 scores of 0.80 for Facebook and 0.83 for Twitter. F1 scores in cross-platform tests dropped to 0.70 for Twitter to Facebook and 0.74 for Facebook to Twitter. Their study also shows that an ensemble of deep learning models with SVM improves gives the best results in both in- and cross-platform cases.

3 Methodology

The process can be roughly divided into five parts. First, we analyzed the existing data. Then we had to train the different classifiers on the Twitter datasets so that we could classify all Telegram messages with them. Because the Telegram messages do not have a gold label, we had to annotate some messages by hand so that the classifiers could then be evaluated on these annotated messages.

3.1 Exploring the Data

As a first step, we will have an overview of the dataset of Telegram messages that was created in a previous work. To do this, we train a topic model on a subset of the messages. This subset of messages is selected as follows. We start with a set of 1500 German channels, from which all those containing less than 300 German messages are discarded in a first step. To weight each channel equally, the same number of messages is sampled from each of the remaining channels. The exact number is determined by the minimum number of messages of the remaining channels.

The messages selected in this way are then prepared in a preprocessing step. Here, all user names are masked, URLs are deleted, and if available, the link preview title and the link preview are appended to the end of the message. Especially the masking of the usernames and the links is important, because they have nothing to do with the actual topic of the message and often do not even consist of meaningful words. Appending the link previews, on the other hand, helps to identify the topic of the message, since the topic of the linked article is usually strongly related to the topic of the message.

These messages were now used to create a topic model using the BERTopic [1] library. We have limited the number of topics to 200. The process of calculating the topic model can be divided into three major steps, the computation of the embeddings, the clustering, and the topic extraction and reduction (see Figure 1). First, the messages are converted into BERT embeddings using the default sentence transformer for non-English messages "paraphrase-multilingual-MiniLM-L12-v2". Then a dimensionality reduction is performed using UMAP to cluster the messages with HDBSCAN. The dimensionality reduction step is necessary because HDBSCAN is very prone to the curse of dimensionality. In the final step, a class-based variant of TF-IDF is used to extract the most relevant words per cluster. These words then act as a description for the topic. The class-based TF-IDF vectors are also used to reduce the topics to the desired number. To do this, the two topics with the most similar TF-IDF vectors are simply merged. To further improve the quality of the words that serve as representatives for a topic, Maximal Marginal Relevance is used to remove those words that do not contribute much to the topic.

To get an overview over the topics in the offensive Telegram messages, we have build another hate-specific topic model, which was trained on 100.000 randomly sampled messages that were labeled as offensive by at least one classifier. Again the link preview title and the link preview were appended to the end of the messages and both URLs and usernames were masked. As this time we concentrate only on the hate topics, we reduced the number of topics to 30.

3.2 Training the Classifiers

In total, we trained eleven classifiers based on different training data, six of which are binary classifiers and the other five are multiclass classifiers. The datasets used were Germeval 18, Germeval 19, HASOC 2019, HASOC 2020, and a dataset created in a previous work regarding Covid-19 specific hate Wich, Räther, and Groh, 2021. The exact partitioning is discussed in more detail in the corresponding subsections. As German pre-trained BERT base models we used

[1] https://github.com/MaartenGr/BERTopic

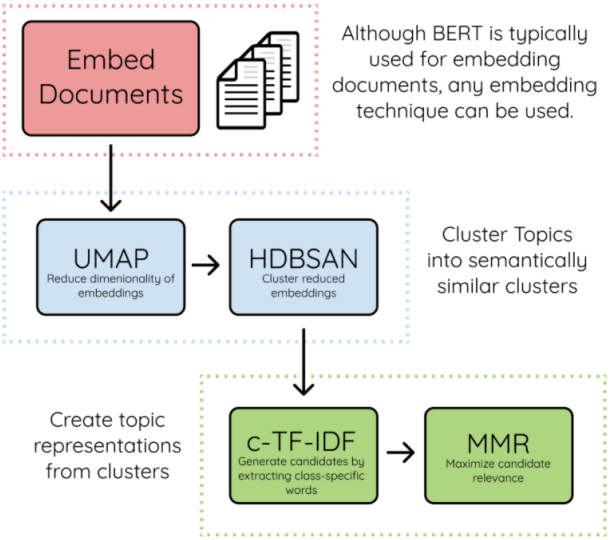


Figure 1 BERTopic pipeline

the three models *'german-nlp-group/electra-base-german-uncased'*, *'dbmdz/bert-base-german-cased'*, and *'deepset/gbert-base'* and for each classifier we took the one with the best performance. Since some datasets are highly imbalanced, we also tried different class weights for each model. Before training, the messages are passed through the ekphrasis [2] preprocessing pipeline, where among other things URLs and usernames are masked and emojis are replaced. Overall, we trained for a maximum of 8 epochs, with early-stopping implemented with a patience of four. In the end, the model with the highest evaluation macro f1 score was selected.

3.2.1 Germeval

In total, we trained three binary classifiers and three fine-grained classifiers using the Germeval 2018 and Germeval 2019 datasets (Wiegand et al., 2018). One pair of classifiers uses only the Germeval 2018 data, a second pair uses only the data from the Germeval 2019 dataset, and the last one uses a combination of both datasets. For the classifiers that use only the Germeval 2018 and Germeval 2019 data, respectively, we adopted the default training and test sets. For the classifiers that use the combination of both Germeval datasets, the training set consists of the Germeval 2018 training set, the Germeval 2018 test set, and the Germeval 2019 training set. The test set is the same as the Germeval 2019 testset. Then, duplicates were removed from all training sets and a 90%/10% train-validation split was performed. (see Figure 2)

[2]https://github.com/cbaziotis/ekphrasis

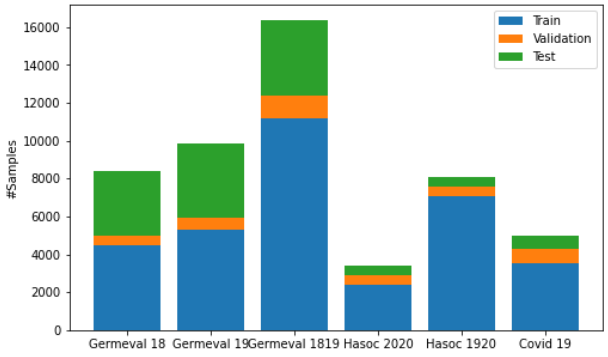


Figure 2 Train-Validation-Test Splits

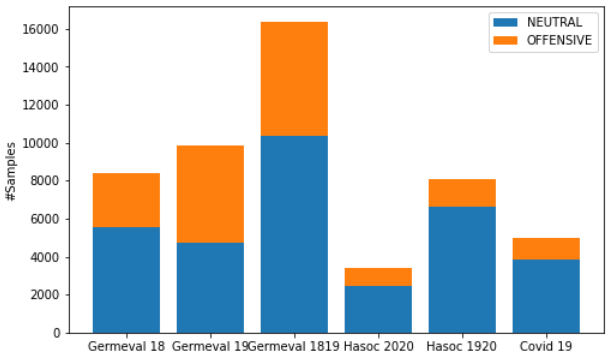


Figure 3 Class Distributions Binary Task

In the binary classification task we distinguish between the following two classes

- OFFENSE: abusive language, insults, as well as merely profane statements
- OTHER: everything that is not in the OFFENSE class

Figure 3 shows the class distribution for the different datasets. The label NEUTRAL corresponds to the class OTHER, while the label OFFENSIVE corresponds to the class OFFENSE.

In the fine-grained classification task, there are the following four classes

- PROFANITY: profane words are used without the intention to insult anyone
- INSULT: unlike PROFANITY there is a clear intention to offend someone
- ABUSE: unlike INSULT, the tweet does not just insult a person but represents the stronger form of abusive language
- OTHER: everything that does not belong to the three previous classes

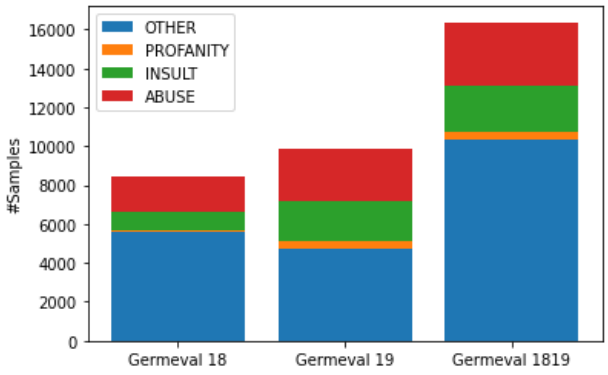


Figure 4 Germeval Class Distribution Fine-grained Task

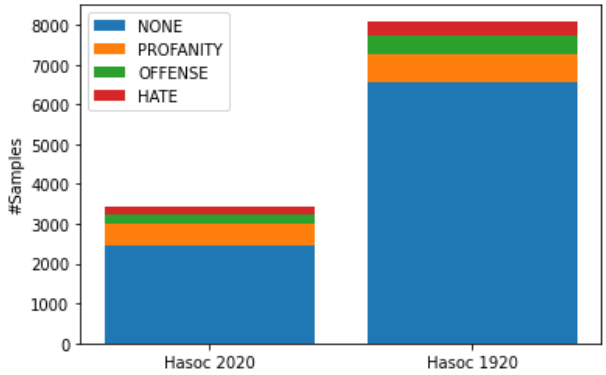


Figure 5 HASOC Class Distribution Fine-grained Task

In Figure 4 one can see the distribution of the four classes in the different datasets.

3.2.2 HASOC

We used the two datasets HASOC 2019 and HASOC 2020 to train a total of four different classifiers, two for a binary classification and two for the fine-grained task (Mandl et al., 2020). Two of these classifiers were trained solely on the Hasoc 2020 data, while the other two classifiers were trained on a combination of the HASOC 2019 and HASOC 2020 datasets. Since the HASOC 2020 dataset was already divided into training validation and testset, we simply adopted this division which results in a 70%/15%/15% train-validation-test ratio. For the classifiers using the HASOC 2019 and HASOC 2020 data, we added the entire HASOC 2019 data to the training set. The resulting ratio of training validation and testset is 87%/6.5%/6.5% which is also shown in Figure 2.

In the binary classification task we consider the two classes

- HOF: the post contains hate speech, profane or offensive content
- NOT: the post does not contain hate speech, profane or offensive content

The distribution of those classes can be seen in Figure 3. The NEUTRAL label corresponds to the class NOT and the OFFENSIVE label corresponds to the HOF class.

In the fine-grained classification task we differentiate between four classes

- PROFANITY: the post contains profane words
- OFFENSE: unlike PROFANITY the post contains offensive content
- HATE: unlike OFFENSE, the post contains hate speech content
- NONE: everything that does not belong to the three previous classes

The exact class distribution is shown in Figure 5.

3.2.3 Covid-19

Over the past year, Covid-19 has become a significant topic in extremist communities, so we also trained a binary classifier on the Covid-19 specific hate speech dataset (Wich et al., 2021). The following two classes are distinguished

- ABUSIVE: The tweet contains any form of insult, harassment, hate, degradation, identity attack, and the threat of violence targeting an individual or a group
- NOT: everything that does not belong to the ABUSIVE class

The class distribution of those labels can be seen in Figure 3.(OFFENSIVE $\hat{=}$ ABUSIVE, NEUTRAL $\hat{=}$ NOT) In order to obtain training-, validation-, and testsets, we performed a random 70%/15%/15% train-validation-test split.(see Figure 2)

3.3 Classifying the Telegram Messages

After training the eleven different classifiers, we used them to classify all German Telegram messages. One problem here is that Telegram messages, unlike Twitter posts, have no character limit. However, the underlying models are only designed for messages consisting of a maximum of 512 different tokens. Therefore, long Telegram messages have to be split into smaller parts and at the end the labels of these parts have to be

reassembled into one global label. This process can be seen in Figure 6.

In a first step, a message consisting of more than 412 words is split into smaller parts. We chose the limit of 412 words to ensure that no more than 512 tokens are generated, since a word can be split into multiple tokens. Furthermore, we made sure that messages are only split at the ends of sentences. Compared to other token limits, 412 has proven to be a good choice. Further details on the different token limits are described in the appendix. In the second step, all parts are classified individually by the respective model. In the third step, the label of the entire message is selected based on the labels of its individual parts. This is done by simply picking the most hateful label. The idea is that if one sentence in a message is offensive, it is enough to classify the whole message as offensive, even if the rest of the message is not. For the Germeval classes the relation in terms of hate is OTHER < PROFANITY < INSULT < ABUSE, for the HASOC classes it is NONE < PROFANITY < OFFENSE < HATE. In the fourth step, the global confidence value has to be determined. For this purpose, the confidence values of the most hateful message parts are examined and, similar to the label, the maximum value is selected.

However, before a message passed through this classification pipeline, some preprocessing was applied. First, the link preview title and the link preview were appended to the end of the message, if available. Then the message was sent through the same ekphrasis preprocessing pipeline as the training data. This means that, for example, URLs and usernames were masked and emojis were replaced.

In addition to the eleven classifiers, we also classified the German Telegram messages using the Perspective API [3]. This is an API developed by Google and Jigsaw that returns a toxicity score for any text, where toxicity is defined as a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion. Since there are no character, word, or token restrictions, even long messages can be classified as a whole and do not need to be split up. Furthermore, no special preprocessing is required. Nevertheless, we deleted URLs and appended both the link preview title and the link preview to the end of the message before calculating the toxicity value.

[3]https://www.perspectiveapi.com

3.4 Annotating the Data

Since the Telegram data is unlabeled, the quality of the classifiers cannot be evaluated. Therefore, we annotated the data ourselves. Due to a lack of time and resources, we could not annotate the entire Telegram messages by hand. Therefore, we selected a set of 1150 messages to be annotated and on which we could then evaluate the different classifiers. Out of the 1150 messages, 700 were selected using the seven binary classifiers and the remaining 450 were selected based on a topic model.

The 700 messages were selected as follows. For each of the seven binary classifiers (three Germeval, two HASOC, one Covid-19, one Perspective API) 100 messages were randomly selected, with the respective classifier labeling 50 of these messages as NEUTRAL and 50 as OFFENSIVE. It was ensured that no message was duplicated.

The remaining 450 messages were provided by an external source. However, they were selected using a topic model such that 30 messages were randomly sampled for each of the 15 most prominent topics. In this way we got a set of messages that contains a balanced mix of NEUTRAL and OFFENSIVE messages as well as messages from the most common topics. It was again checked that there were no duplicates.

The annotation schema for the sampled messages contains the following two classes

- OFFENSIVE: The tweet contains any form of insult, harassment, hate, degradation, identity attack, and the threat of violence targeting an individual or a group
- NEUTRAL: everything that does not belong to the OFFENSIVE class

The messages were annotated by five non-experts, all male and in their twenties. In order to obtain consistent annotations, all annotators were prepared by presenting the annotation guidelines and discussing some examples. Since the annotators were non-experts, there was an possibility to skip messages, although everyone was encouraged to do so only in case of complete uncertainty. Due to limited resources, only a set of 50 messages was labeled by all annotators in a first round. To ensure a balance of hate speech and normal messages, the 50 messages were selected to contain 25 NEUTRAL and 25 OFFENSIVE messages according to the Perspective API. Then, inter-rater reliability was measured using Krippendorff's alpha (Krippendorff, 2018). In addition, all messages that had a consensus

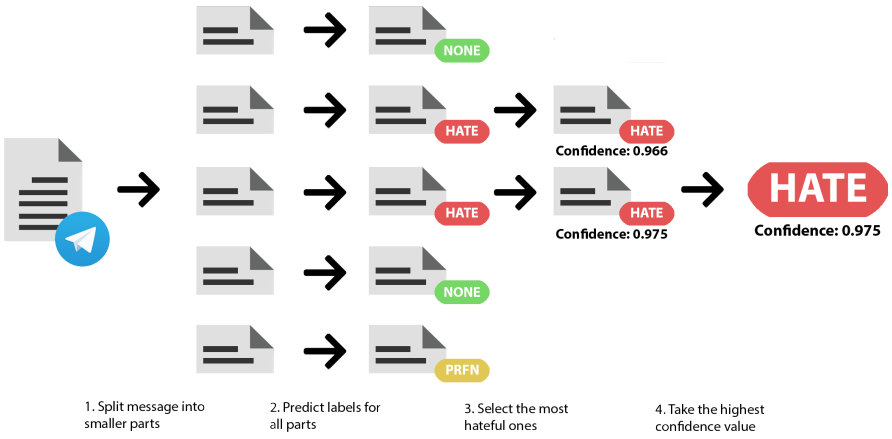


Figure 6 Telegram Message Classification Pipeline

below 80% were revisited with all annotators to further increase consistency among the annotators.The remaining 1150 messages then received labels from only two of the five annotators. If these two labels matched, then it was finally adopted as a gold label, otherwise the message was annotated in a third round by another three persons who decided on the final label. We used Kili Technology [4] as an annotation tool.

3.5 Evaluating the Classifiers

To test the generalizability of the classifiers, we evaluated each classifier on the others' test sets. This is possible for the binary classifiers because the labels of all datasets have a strong semantic overlap. However, it is not possible with the classifiers for the fine-grained task, since here the labels have larger differences for the Germeval datasets and the HASOC datasets.

Finally, we evaluated all classifiers on the hand-labeled Telegram messages. Since only binary labels were available due to lack of time and resources, we also interpreted the multi-class classifiers as binary classifiers. For the Germeval data, the labels PROFANITY, INSULT, and ABUSE were combined into the label OFFENSIVE, and analogously for the HASOC data, the labels PROFANITY, OFFENSE, and HATE were combined. For the Perspective API, we chose a threshold of 0.5, such that a message with a toxicity score ≤ 0.5 is labeled as NEUTRAL, and one with a toxicity score > 0.5 is labeled as OFFENSIVE.

[4] https://kili-technology.com

Table 1 Binary Classifiers Test Set Results

	Acc	Prec	Rec	Macro F1	Model
GE 18	78.4	71.1	61.0	75.0	dbmdz
GE 19	77.1	72.2	85.1	77.1	dbmdz
GE 1819	84.0	87.6	77.6	83.8	dbmdz
HS 20	85.0	69.0	73.7	80.6	deepset
HS 1920	85.2	71.0	69.9	80.3	dbmdz
CO 19	88.5	73.9	69.9	82.3	deepset

4 Results

Our dataset includes a total of 13.862.930 unlabeled Telegram messages from a total of 48.104 different channels. Of these, 5.439.691 messages and 46.275 channels are in German language. All messages stem from a period from the first of January 2019 to the fifteenth of March 2021. The analysis of the most frequent topics in the German Telegram messages according to the topic model has revealed that Corona, especially vaccinations, cryptocurrencies, esoterics, and politics seem to be the most common topics.

The results of the six binary classifiers on their respective testset can be seen in Table 1, the precision and the recall refer to the offensive class. In the 'Model' column one can also see which of the three base models *'german-nlp-group/electra-base-german-uncased'*, *'dbmdz/bert-base-german-cased'*, and *'deepset/gbert-base'* was used to achieve these metrics. Analogously, Table 2 shows the results of the five classifiers trained on the fine-grained classification task, where the columns 'Prec' and 'Rec' indicate the Macro Precision and Macro Recall, respectively.

The cross-dataset evaluation of the binary classifiers yielded the results presented in the following three tables showing the Macro F1 score (see Table 3), the

Table 2 Fine-grained Classifiers Test Set Results

	Acc	Prec	Rec	Macro F1	Model
GE 18	74.7	57.1	46.6	49.7	deepset
GE 19	69.3	58.7	57.7	57.9	deepset
GE 1819	78.2	68.8	65.3	66.8	dbmdz
HS 20	80.2	60.3	63.5	61.3	deepset
HS 1920	79.7	59.7	65.1	61.1	deepset

Table 3 Cross Dataset Evaluation Macro F1

	CO 19	GE 18	GE 19	HS 20
GE 18	71.5	75.0	76.3	64.9
GE 19	71.6	74.1	77.1	49.4
GE 1819	75.7	—	83.2	66.7
HS 20	60.0	64.2	58.5	80.6
HS 1920	59.2	59.9	54.6	80.3
CO 19	82.3	74.0	75.9	66.0

precision (see Table 4), and the recall (see Table 5). In all three tables, the rows correspond to the classifiers and the columns to the testsets. Since the Germeval 1819 classifier and the Germeval 19 classifier both have the same testset, the column for the Germeval 1819 testset was omitted. The same holds true for the HASOC 2020 and the HASOC 1920 classifiers. Moreover, the Germeval 18 testset is part of the Germeval 1819 classifier's trainingset, which is why it could not be evaluated on the Germeval 18 testset.

In total, we annotated 1149 Telegram messages by hand, as one message had to be discarded during the annotation process due to missing text. The Krippendorff's alpha value after the sample round was 63.15%, which is acceptable in the context of hate speech. Out of the 1099 messages in the main annotation round, 126 had to be labeled again in another round due to a tie. In the end, 968 messages were labeled NEUTRAL and 181 messages were labeled OFFENSIVE. This leads to a test set with 84.2% NEUTRAL messages and 15.8% OFFENSIVE ones. After all 1149 messages were annotated, we recalculated the Krippendorff's alpha value for the whole annotated data. The final Krippendorff's alpha value is 73.87%, which is much better than the value after the sample round.

Table 4 Cross Dataset Evaluation Precision

	CO 19	GE 18	GE 19	HS 20
GE 18	59.5	71.1	82.8	42.4
GE 19	48.2	57.5	72.2	32.7
GE 1819	68.7	—	87.6	44.3
HS 20	92.8	85.1	93.4	69.0
HS 1920	76.5	78.3	88.4	71.0
CO 19	73.9	68.9	79.0	43.4

Table 5 Cross Dataset Evaluation Recall

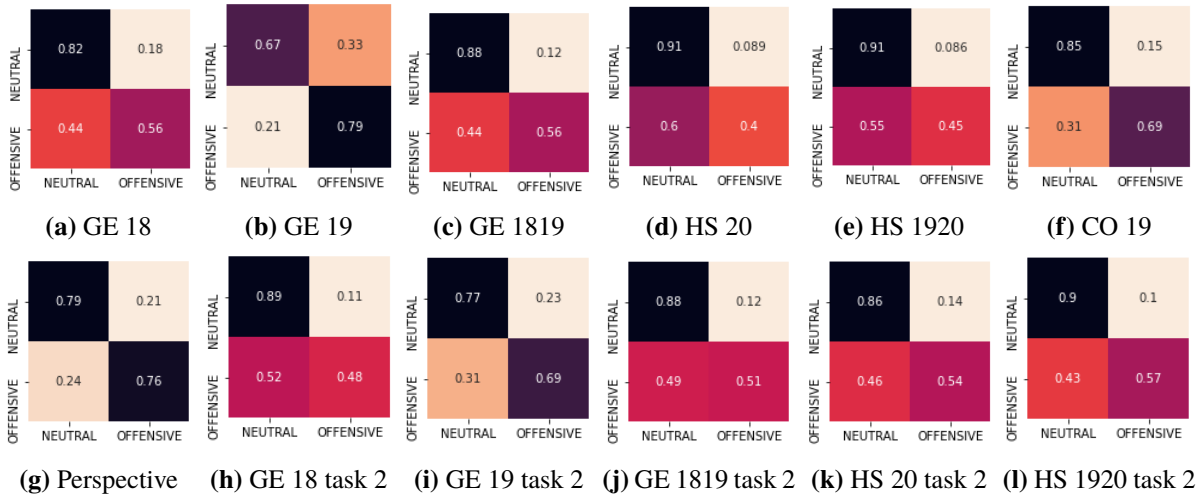
	CO 19	GE 18	GE 19	HS 20
GE 18	49.3	61.0	65.1	94.0
GE 19	73.3	93.3	85.1	92.5
GE 1819	54.1	—	77.6	79.7
HS 20	17.8	30.4	27.7	73.7
HS 1920	17.8	24.8	23.2	69.9
CO 19	69.9	60.5	68.4	81.9

Table 6 Evaluation on Hand-labeled Data

	Acc	Prec	Rec	Macro F1
GE 18	77.6	36.4	56.4	65.1
GE 19	69.2	31.2	79.0	61.7
GE 1819	82.7	45.9	56.4	70.1
HS 20	83.0	45.6	39.8	66.3
HS 1920	84.1	49.4	44.8	68.8
CO 19	82.2	45.6	69.1	71.9
Perspective	78.4	40.2	76.2	69.3
Fine-grained				
GE 18	82.9	45.7	47.5	68.2
GE 19	76.2	36.4	69.1	66.1
GE 1819	82.2	44.2	50.8	68.3
HS 20	80.7	41.4	54.1	67.5
HS 1920	84.8	51.5	56.9	72.5

Finally, we evaluated all classifiers on the 1149 hand-labeled Telegram messages. Table 6 shows the results of this evaluation. The upper half shows the results of the binary classifiers and the Perspective API with a threshold of 0.5, while the lower half shows the results of the multi-class classifiers that were considered as binary classifiers for the purpose of this evaluation. For a more detailed analysis, Figure 7 shows the resulting confusion matrices.

Table 7 shows the eleven most meaningful topics identified by the hate-specific topic model. It can be seen that in addition to the typical topics such as racism, religion, and refugees, the Corona pandemic also appears to be a common topic. With the topics 'vaccination', 'corona', 'china', and 'corona related' there are four topics that are directly linked to the Corona pandemic. So we can expect the newly emerging corona-related hate speech to play an important role in our Telegram dataset.



(a) GE 18 **(b)** GE 19 **(c)** GE 1819 **(d)** HS 20 **(e)** HS 1920 **(f)** CO 19

(g) Perspective **(h)** GE 18 task 2 **(i)** GE 19 task 2 **(j)** GE 1819 task 2 **(k)** HS 20 task 2 **(l)** HS 1920 task 2

Figure 7 Confusion Matrices

Table 7 Hate Topics in Telegram Data

Topics	Keywords
vaccination	impfstoff, impfung, covid, impfpflicht, virus
religion	israel, juden, muslime, islamisten, zionisten, israelis
refugees	migranten, flüchtlinge, einwanderer, griechenland
US politics	trump, wahl, wahlbetrug, demokraten, präsident
corona	virus, pandemie, quarantäne, impfung, infizierte
racism	rassismus, nazi, schwarzen, deutschland, rassisten
climate	klima, co2, grüne, klimawandel
economics	wirtschaft, milliarden, korruption, geld, regierung
pedophilia	kindesmissbrauch, kinderpornographie, pädophile
china	china, chinesen, virus, wuhan, kommunisten
corona related	ffp2, pcr, test, 5g

5 Discussion

Looking again at the testset results of the binary and fine-grained classifiers, one can see that none of the classifiers uses the *'german-nlp-group/electra-base-german-uncased'* model as a basis. This base model had problems especially with the very imbalanced datasets, and even the class weighting could not improve the results. However, the *'dbmdz/bert-base-german-cased'* and the *'deepset/gbert-base'* models performed both very well. All classifiers have a rather balanced precision and recall and yield good results on their own testsets.

We have also eveluated the binary classifiers on the other classifier's testsets in order to investigate to cross-dataset performance. In Table 3 one can see that the Germeval classifiers and the one trained on the Covid-19 dataset perform rather well on each others testsets, whereas their F1 score significally drops on the HASOC datasets. The HASOC classifiers on the other hand only perform well on their own testset. This could means that the HASOC classifiers focus on a different kind of hateful language than the rest of the classifiers. Knowing that there are also many Telegram messages containing some Covid-19 topic, it is good to see that the Germeval classifiers seem to recognize also the Covid-19 specific hate at least to some extent.

To find the reason for the good or poor performance, we also looked at precision and recall. It is noticeable that the HASOC classifiers have a high precision but a really low recall on the other test sets, while precision and recall are balanced on their own test set. The weak cross-dataset performance of the HASOC models is therefore due to the fact that they do not

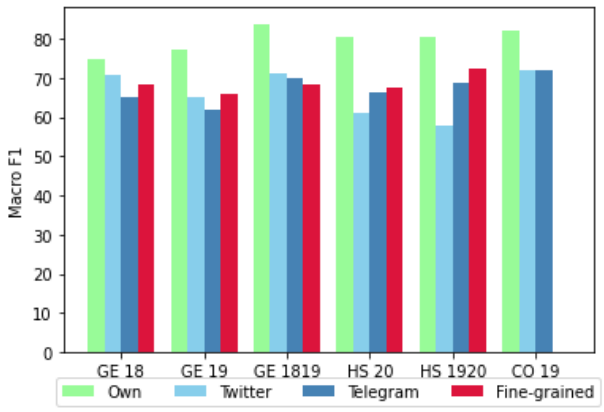


Figure 8 Macro F1 Comparison across different Testsets

recognize many offensive tweets as such and classify them as non-offensive. At the same time, it can be observed that all other classifiers have a very high recall and a low precision on the HASOC testset. This suggests that the HASOC data does not contain a different kind of hateful language, but that the label HOF denotes a stronger kind of hate than the labels OFFENSE and ABUSIVE in the Germeval and Covid-19 data, respectively. It is also noticeable that the Germeval 19 classifier has a very high recall and the lowest precision across all test sets. This means that the model is too strict, it also classifies many non hateful tweet as offensive. However, in total the results show that at least the Germeval and the Covid-19 classifiers seem to generalize sufficiently well over other Twitter datasets, which means that they might also be suitable for a classification task on Telegram messages.

Finally, we evaluated all classifiers on the annotated Telegram messages. As can be seen in the table, all classifiers have a similar Macro F1 score in a range from 0.61 to 0.73. In particular, they perform as well as the external Perspective API, which we used as a baseline in this work. If for now we consider only the binary classifiers, we can again observe that the two HASOC classifiers have the lowest recall. However, the difference to the other classifiers is by far not as large as in the cross-dataset evaluation. Interestingly, the fine-grained HASOC classifiers do not have the problem of low recall. With the Germeval 19 classifier, one can again observe the tendency towards high recall and low precision which was identified in the cross-dataset evaluation. Thus, we can say that the same patterns that were already found in the cross-dataset evaluation are emerging, although the messages stem from a different platform.

To better compare the results on the Telegram test set, Figure 8 shows the Macro F1 scores of all classifiers on the different test sets. The green bar correponds to the test set of the dataset that the classifiers were trained on, the light blue bar represents the mean of the F1 scores on the test sets of the other classifiers determined in the cross-dataset evaluation, the dark blue bar belongs to the result of the binary classifiers on the annotated Telegram data and the red bar shows the result of the fine-grained classifiers on the Telegram data. It can be seen that for all classifiers the performance is best on their own test set, which was expected since this test set should be the most similar to the training data. However, if we compare the results on the other Twitter data with the results on the Telegram data, we cannot see a big difference. The HASOC classifiers perform much better on the Telegram data, while the Germeval 18 classifier performs slightly worse on the Telegram messages. For the other Germeval classifiers and the Covid 19 classifier, the difference is only marginal. This means that the platform shift from Twitter to Telegram does not have a major impact on the performance of the classifiers. The lower F1 value compared to the own test set can be entirely attributed to the loss due to the cross-dataset evaluation, regardless of the platform. It is also interesting to note that all fine-grained classifiers, except Germeval 1819, perform better on the Telegram data than their binary counterpart.

To get a better overview of where the classifiers make mistakes, you can look at the confusion matrices in Figure 7. Here we can see that the Perspective API is by far the most balanced classifier with respect to the ratio of false positives and false negatives, all other classifiers have a clear tendency towards either false positives or false negatives. All classifiers, except the two Germeval 19 and the Covid 19 classifier, have massive difficulties with false negatives. This is not surprising for the Germeval 19 classifiers, since it could already be observed in the cross-dataset evaluation that the classifier tends to have a high recall at the cost of precision. A possible reason for the high false negative rates could be the Covid-19 specific hate present in the Telegram data (see Table 7), as the Germeval and HASOC classifiers already showed by far the lowest recall in the cross-dataset evaluation on the Covid-19 dataset.

It would be especially interesting to see how the classifiers perform on long messages that need to be split, as one could thereby evaluate how good our process of splitting the message into smaller parts and taking the

most hateful label is. Unfortunately, among the annotated Telegram messages there are only 39 messages that were split. This is because we did not consider the length of the messages in the selection process. Because it is impossible to draw reasonable conclusions with so little messages, this remains a research subject for future works.

6 Conclusion

We trained a total of eleven classifiers on different Twitter hate speech data sets and showed that they also perform well across platforms on Telegram messages. This can be seen in the comparison with the external Perspective API, which achieves a very similar F1 score on our Telegram test set as the other classifiers. Although for all eleven classifiers precision recall and Macro F1 score are significantly lower than on their own test sets, the comparison with their performance on other Twitter test sets showed that these differences are probably not due to the platform change. One point of criticism we are aware of is that the performance of most classifiers would probably drop with a more balanced test set, as the majority of the classifiers has significant problems with false negatives. However, we think that this is mainly because of the newly evolved Corona-specific hate that was not part of their training sets, which means that this issue is again platform independent and does not originate from the platform change itself.

Additionally, we have created a new annotated dataset consisting of 1149 German Telegram messages.15.8% of the messages are labeled as OFFENSIVE and the remaining 84.2% are NEUTRAL. It was also ensured that the dataset contains messages from all common Telegram topics that were identified via a topic model beforehand.

A Different Token Limits

As described in the previous chapters, Telegram messages have no character limit and thus have to be split up if they exceed a certain token limit, as the base models can only handle up to 512 different tokens. We have tried several different token limits, which we then evaluated on the 1149 annotated Telegram messages. The results can be seen in Table 8, Table 9 and Table 10. Those 1149 messages contain 39 messages that have more than 412 tokens, 80 messages that have more than 200 tokens, 265 ones that have more than 100 tokens, 687 ones that have more than 50 tokens, and 912 ones that have more than 25 tokens.

Table 8 Macro F1 Scores for different Token Limits

	412	200	100	50	25
GE 18	65.1	65.2	64.3	63.3	56.8
GE 19	61.7	61.8	60.6	56.7	51.1
GE 1819	70.1	70.0	69.3	67.5	60.6
HS 20	66.3	64.8	65.6	67.7	69.4
HS 1920	68.8	69.0	69.4	69.8	69.1
CO 19	71.9	71.8	71.6	69.1	66.8
Fine-grained					
GE 18	68.2	68.4	68.4	69.0	60.8
GE 19	66.1	66.4	66.5	62.3	53.5
GE 1819	68.3	68.3	67.2	67.2	64.6
HS 20	67.5	67.5	67.7	67.8	65.7
HS 1920	72.5	72.3	72.6	73.1	73.6

Table 9 Precision for different Token Limits

	412	200	100	50	25
GE 18	36.4	36.4	35.0	33.0	27.2
GE 19	31.2	31.3	30.2	27.2	24.4
GE 1819	45.9	45.7	43.2	38.5	30.1
HS 20	45.6	45.9	47.8	49.3	49.4
HS 1920	49.4	51.0	46.4	43.6	40.9
CO 19	45.6	45.2	44.6	40.7	36.9
Fine-grained					
GE 18	45.7	46.0	44.3	42.4	30.2
GE 19	36.4	36.7	36.4	31.8	25.4
GE 1819	44.2	43.7	40.8	38.2	34.2
HS 20	41.4	41.1	40.5	39.3	36.0
HS 1920	51.5	51.0	49.6	49.4	50.9

One can see that on average the classifiers perform the best regarding the Macro F1 score with a token limit of 412, although the differences to the 200 and 100 token limit is negligible. However, especially for the Germeval classifiers the macro F1 score drops significantly with a token limit of 25, which might be due to inability to recognize long term dependencies.

Also the precision and the recall are very similar for the token limits 412, 200 , and 100. For the smaller token limits 50 and 25, the recall starts to increase a lot for several classifiers, while the precision starts to drop.

Table 10 Recall for different Token Limits

	412	200	100	50	25
GE 18	56.4	56.9	56.9	64.6	81.8
GE 19	79.0	80.1	81.2	82.3	90.1
GE 1819	56.4	56.4	59.7	68.5	76.8
HS 20	39.8	34.3	35.4	40.9	47.0
HS 1920	44.8	43.6	46.4	43.6	40.9
CO 19	69.1	70.2	71.3	70.2	74.6
Fine-grained					
GE 18	47.5	48.1	51.4	60.2	72.4
GE 19	69.1	70.7	75.7	82.3	86.7
GE 1819	50.8	51.9	53.6	66.9	76.2
HS 20	54.1	54.7	58.6	65.7	66.9
HS 1920	56.9	56.9	61.3	64.6	63.5

References

Fortuna, P., Bonavita, I., & Nunes, S. (2018). Merging datasets for hate speech classification in italian. In *Evalita@clic-it*.

Hughes, S. & Meleagrou-Hitchens, A. (2017). The threat to the united states from the islamic state's virtual entrepreneurs. *CTC Sentinel*, *10*(3), 1–8.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Mandl, T., Modha, S., Kumar M, A., & Chakravarthi, B. R. (2020). Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In *Forum for information retrieval evaluation* (pp. 29–32).

Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., & Patel, A. (2019). Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. New York, NY, USA: Association for Computing Machinery. doi:10.1145/3368567.3368584

Markov, I. & Daelemans, W. (2021). Improving cross-domain hate speech detection by reducing the false positive rate. doi:10.18653/v1/2021.nlp4if-1.3

Scheffler, T., Solopova, V., & Popa-Wyatt, M. (2021). The telegram chronicles of online harm. *Journal of Open Humanities Data*, *7*. doi:10.5334/johd.31

Semenzin, S. & Bainotti, L. (2020). The use of telegram for non-consensual dissemination of intimate images: Gendered affordances and the construction of masculinities. *Social Media+ Society*, *6*(4), 2056305120984453.

Siegel, A. A. (2020). Online hate speech. *Social Media and Democracy: The State of the Field, Prospects for Reform*, 56–88.

Struß, J. M., Siegel, M., Ruppenhofer, J., Wiegand, M., & Klenner, M. (2019). Overview of germeval task 2, 2019 shared task on the identification of offensive language.

Walther, S. & McCoy, A. (2021). Us extremism on telegram. *Perspectives on Terrorism*, *15*(2), 100–124.

Wich, M., Räther, S., & Groh, G. (2021). German Abusive Language Dataset with Focus on COVID-19. In *Accepted at 17th konvens*.

Wiegand, M., Siegel, M., & Ruppenhofer, J. (2018). Overview of the germeval 2018 shared task on the identification of offensive language.

Williams, M. L., Burnap, P., Javed, A., Liu, H., & Ozalp, S. (2020). Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, *60*(1), 93–117.