HS-EMO: Analyzing Emotions in Hate Speech

Johannes Schäfer and **Elina Kistner**
Institute for Information Science and Natural Language Processing
University of Hildesheim
Hildesheim, Germany
{johannes.schaefer,kistner}@uni-hildesheim.de

Abstract

This paper investigates the interplay between hate speech and emotions in social media postings with the goal of modeling both phenomena jointly. We present a bottom-up analysis and introduce an English text corpus with fine-grained annotations for both phenomena, in which we analyze possible correlations. Our results show that only some of the categories representing negative emotions correlate with hate speech classes, while others, such as sadness, do not. With our dataset, we explore methods for using partially annotated data to learn both classifications jointly in an experiment with a transformer-based neural network model. Our results suggest that using a hate speech dataset with emotion labels is more useful than standard multi-task learning with multiple separate datasets. We make our annotation and the code of our experiments publicly available.[1]

1 Introduction

Hate speech[2] remains a persisting issue in social media. This includes offensive language (Wiegand et al., 2018; Struß et al., 2019; Mandl et al., 2021) and toxicity (Borkan et al., 2019) which are of interest for regulation and thus motivate automatic detection approaches. Methods have to consider a variety of features to capture the complex phenomenon. A survey on general approaches for hate speech detection is given by Schmidt and Wiegand (2017). Recently, mainly transformer-based pre-trained language models (e.g. BERT by Devlin et al., 2019) have shown the most promising results (e.g. in Caselli et al., 2021; Magnossão de Paula et al., 2022). Typically, the focus in natural language processing research is on fine-tuning using a specialized dataset and on model development. In this paper, we propose to broaden the scope of analysis to include knowledge from the related field of emotions.

The underlying emotions in posts on social media are often studied based on datasets, e.g. Bostan and Klinger (2018) discuss several corpora annotated for emotion categories. These include emotions such as anger, disgust, sadness, joy, fear and surprise. Although hate can be considered a type of emotion, the interplay of the different emotions with hate speech content has not been precisely identified. Alorainy et al. (2018) find hate speech messages from suspended user accounts are often associated with negative emotions such as disgust, fear and sadness. This motivates using emotion analysis as features for hate speech detection, e.g. as shown by Martins et al. (2018), Markov et al. (2021), Chiril et al. (2022) and Rana and Jha (2022). Madukwe et al. (2021) use an emotion lexicon to generate a weighted emotion embedding vector as additional features that prove beneficial for hate speech classification.

To take emotions in hate speech even more into account, both phenomena can be learned in a joint model (Rajamanickam et al., 2020; Awal et al., 2021). Plaza-del Arco et al. (2021) present a multi-task learning system which includes a classifier for emotion detection as well as a classifier for hate speech and offensive language detection. They use a shared encoder which is trained sequentially with batches from a different dataset for each classification task.

In this paper, we investigate whether such a multi-task learning approach benefits further from using a single dataset that contains annotations for both phenomena. To this end, we perform a bottom-up analysis of emotions in hate speech posts and create an annotated dataset that can be used for joint classification. Our contributions include (i) a corpus annotated both for four hate speech and offensive language categories as well as for six

[1] https://github.com/Johannes-Schaefer/HS-EMO

[2] Warning: This paper contains examples of hate speech and offensive language. These examples are taken from social media corpora and do not represent the opinion of the authors.

Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023), pages 165–173
September 18–22, 2023.

	anger	*disgust*	*sadness*	*joy*	*fear*	*surprise*	*?*	_	total
TEC	1,555 (7 %)	761 (4 %)	3,830 (18 %)	8,239 (39 %)	2,814 (13 %)	3,848 (18 %)	-	-	21,047 (100 %)
HS-EMO	352 (35 %)	172 (17 %)	158 (16 %)	113 (11 %)	79 (8 %)	62 (6 %)	37 (4 %)	27 (3 %)	1,000 (100 %)

Table 1: Emotion label distribution in our HS-EMO corpus in comparison to the TEC corpus. The percentages refer to the proportions in each data set, i.e. they are relative values for the respective row of the table.

emotion categories, and (ii) a preliminary experiment to explore methods for learning the phenomena jointly by leveraging emotion analysis in hate speech detection.

The remainder of this paper is structured as follows. In Section 2, we outline our annotation procedure that we use for our dataset, which is presented in Section 3, where we also discuss salient observations. Section 4 presents the experiments on our dataset for joint modeling of hate speech and emotions. Finally, we conclude in Section 5.

2 Annotation

Since the phenomenon of hate speech is rarer than individual emotion categories, we begin our analysis with a dataset that has already been annotated for fine-grained categories relevant to the detection of hate speech. Here we use HASOC in the version from 2021 (Mandl et al., 2021) which contains Hate and Offensive (HOF) content collected from Twitter during the Covid-19 pandemic. This dataset comprises 3,843 English text messages of hate speech (HATE, 683 cases), offensive language (OFFN, 622 cases) and profane content (PRFN, 1,196 cases) as well as other/neutral content (NONE, 1,342 cases).

To analyze these data for the underlying emotions, we annotate a stratified sample of 1,000 instances with six different emotions (joy, anger, disgust, fear, sadness, surprise) on the basis of the categories by Ekman (1988). Additionally, we annotate the label *"?"* in cases where the classification is not clear and the label "_" in cases where no emotion is apparent from the message content. Emotions were classified according to the presumed emotional state of the author of the analyzed message. The annotation was performed by one annotator. To gain a better understanding of the annotation of Twitter data for emotions, the annotator trained on the Hashtag Emotion Corpus (TEC, Mohammad, 2012).

Challenges were presented by cases in which multiple emotions could be detected in a tweet, i.e., when the author presumably felt two different emotions. In such cases, the stronger emotion was determined by guessing which emotion triggered the writing of the message. While in total we annotate six different emotions, we also consider subclasses to ease the annotation. These include, for example:

- Joy: affection, goodwill, zest, pride, hope, acceptance, excitement, relief, passion, caring.
- Anger: irritability, jealousy, rage, frustration.
- Disgust: torment, shame, contempt.
- Fear: nervousness, threat, uncertainty, anxiety, panic, shock.
- Sadness: suffering, regret, displeasure, embarrassment, sympathy, depression.
- Surprise: unexpectedness, astonishment, confusion, unpreparedness.

We provide examples for the annotation of different emotions found in this dataset in Appendix A.

3 HS-EMO Corpus

Our corpus is a sample of instances from the HASOC corpus which we annotate for emotion categories as described above. In total our annotated dataset HS-EMO comprises 1,000 messages where approximately 65% are to be considered hateful or offensive. Table 1 illustrates the distribution of emotions which we identified in this data in comparison to the distribution in the TEC dataset. We observe a more skewed distribution in our data towards negative emotions (especially *anger* and *disgust*) while more positive emotions are less frequent.

We now analyze the correlation of the different emotions with the annotated hate speech categories (see Table 2 and Table 3). Table 2 shows the distributions for the binary categories HOF vs. NONE. Here we observe an even stronger skewed distribution for the HOF class towards the negative emotions *anger* and *disgust*. Out of the instances annotated as HOF, approximately 64% (278 and

	anger	*disgust*	*sadness*	*joy*	*fear*	*surprise*	*?*	_	total
HOF	278 (43 %)	139 (21 %)	47 (7 %)	78 (12 %)	30 (5 %)	40 (6 %)	20 (3 %)	16 (2 %)	648 (100 %)
NONE	74 (21 %)	33 (9 %)	111 (32 %)	35 (10 %)	49 (14 %)	22 (6 %)	17 (5 %)	11 (3 %)	352 (100 %)

Table 2: Emotion and coarse-grained HOF/NONE label correlation in our corpus HS-EMO. The percentages refer to the proportions for each of the labels HOF/NONE, i.e. they are relative values for the respective row of the table. The total counts for each emotion are displayed in Table 1 (row HS-EMO).

	anger	*disgust*	*sadness*	*joy*	*fear*	*surprise*	*?*	_	total
PRFN	136 (44 %)	33 (11 %)	14 (5 %)	65 (21 %)	8 (3 %)	26 (8 %)	16 (5 %)	11 (4 %)	309 (100 %)
OFFN	75 (46 %)	48 (29 %)	9 (5 %)	11 (7 %)	9 (5 %)	7 (4 %)	1 (1 %)	4 (2 %)	164 (100 %)
HATE	67 (38 %)	58 (33 %)	24 (14 %)	2 (1 %)	13 (7 %)	7 (4 %)	3 (2 %)	1 (1 %)	175 (100 %)

Table 3: Emotion and fine-grained HOF label correlation in our corpus HS-EMO. The percentages refer to the proportions for each label PRFN/OFFN/HATE, i.e. they are relative values for the respective row of the table. The total counts for each emotion are displayed in Table 2 (row HOF).

139 instances) fall into one of these two emotion categories. Interestingly, the other negative category *sadness* does not correlate with HOF. Only 7% (47 instances) of HOF cases occur with the emotion *sadness*, while *sadness* was annotated for 32% (111 instances) of non-HOF cases. We find this to be the case, since such examples often contain only sad sympathy for the misfortunes of others and tend not to be offensive or hateful. We support these findings by discussing the HOF content for these emotion categories using selected examples displayed in Table 4. Examples #1 through #4 are HOF cases with the emotions *anger* or *disgust*. These texts mostly report negative feelings on the government or political situation. Here we find expressions in which the blame is assigned to someone. Actions of certain people or groups are despised and they are attacked for it. This blaming is rarely found in examples with *sadness*. For example, consider examples #5 through #7, in which the authors are more reflective. The expressions are not necessarily directed towards a person, but rather refer to an event or the general situation, which is not expressed as hate speech.

As a deeper analysis, we further consider the distribution of emotion categories in the fine-grained hate speech classes. In Table 3, the tweets annotated with emotions are divided into the three HOF categories (PRFN, OFFN and HATE). Out of the 352 tweets annotated with *anger* (cf. Table 1), 278 contain HOF (cf. Table 2) and of these 136 are PRFN, i.e. almost half of the HOF tweets labeled as *anger* contain just vulgar language without targeting a particular person or group. However, for the emotion label *disgust* we observe a correlation with the more severe hate speech categories (HATE and OFFN). For the emotion label *surprise*, 40 out of a total of 62 tweets are marked with HOF and of these only seven examples are considered to be severe HATE (most of them instead belong to the PRFN class). Similarly, for the emotion label *joy* with a total of 113 examples, 78 are marked as HOF with most of them belonging to the PRFN class. Interestingly, for this emotion label *joy* we even find two cases which involve HATE. We now take a closer look at the texts that contain some of these surprising findings.

The most unexpected cases are probably the two examples which are both annotated for *joy* und HATE. These texts of these messages are as follows:

- "@USER I don't think so I am a stupid and never tell others stupid bcoz it is their Ignorance. But still I stand with #Resign_PM_Modi #ResignModi #resign_modi"
- "This time I am with you! Bloody #China spreading #chinesevirus! URL"

Both examples can be seen as instances where the

#	Text	Emotion	HOF
1	"#CommunistVirus is wreaking havoc in india. Not a single liberal is blaming their Beijing Masters. Hypocrites. #ChineseVirus"	*anger*	yes
2	"Wow. Massive asshole timing. Fuck this guy forever. He must be popular with the Trumpers. URL"	*anger*	yes
3	"What a bunch of absolute fucking idiots in #india #IndiaCovidCrisis. Brainless morons wonder why they have a "crisis" (this is goa, sent by an Armenian living there for months) @USER @USER @USER"	*disgust*	yes
4	"Such a pathetic government who keeps denying that there is no shortage of oxygen....shameless characters to go immediately #AndhBhakt #BjpDestroyedIndia @USER @USER #ResignModi"	*disgust*	yes
5	"I have coworkers whose family and friends are sick and dying in India. Other offshore coworkers are sick themselves. Praying the international community does the right thing to help India. Yes, India's Covid crisis hurts everyone. #PrayForIndia #IndiaCovidCrisis URL"	*sadness*	no
6	"#COVID19 After 70 years of independence we failed to deliver Oxigen, medical facilities and vaccination to us. #IndiaCovidCrisis"	*sadness*	no
7	"We aren't opposing BJP, we're only criticizing them because we don't want to loss lives of Hindus in Bengal violence .. #SpinelessBJP #isupportmodi #Modi #BJP #Shamemamatabannerjee #tmcgoons #ShameOnMamata #ArrestMamata #BengalBurning #BengalViolence #TMCTerror URL"	*sadness*	no

Table 4: Examples of anger/disgust HOF cases in comparison to sadness non-HOF cases as text instances from our corpus (HS-EMO). Username mentions and URLs have been anonymized.

author seems to be joyful out of an enthusiastic group sentiment, but that collectively fuels hatred.

In addition, we report another example from the HATE category where, unexpectedly, no emotion was detected:

- "Now, the "poorly paid, but professional, criminals" i.e. "gutter worms" from BJP IT Cell - which is India's No. 1 #FakeNews factory - have got another picture to trend by using these following hashtags: #BengalBurning #BengalViolence #ShameOnMamata #ArrestMamata URL"

4 Experiments

We now use our data sample annotated for both emotion and hate speech to assess whether this joint annotation can be beneficial for modeling both phenomena jointly. To test different methods for learning to recognize hate speech while possibly considering emotion analysis, we implement a neural network approach. We encode text messages using the transformer-based pre-trained language model BERT (Devlin et al., 2019) and perform the classification for each task in a separate linear layer on top of the pooled encoder output. Further details and hyperparameters are described in Appendix B.

4.1 Experimental Setups

We train the shared encoder in any setup and the classifiers only for the respective tasks available given the used dataset. All our models are trained on the HASOC data to optimize the hate speech classification component (training data *HS*). Additionally, we implement optional training steps to incorporate emotion analysis in different ways as follows. We use the TEC corpus as additional source material to train the emotion classifier alternately with the hate speech classifier (standard multi-task learning (MTL) on separate datasets, training data *HS&Emo*). We also allow for training on our dataset to train both classifiers simultaneously via joint classification (training data *HSEmo*). The combination of these training steps results in four overall approaches which we investigate:

- *HS* as a first baseline for hate speech detection without emotion analysis.
- *HS & Emo* as a second baseline with standard MTL on two separate datasets.
- *HS & Emo & HSEmo* as extension of the second baseline including joint MTL on our dataset.
- *HS & HSEmo* as extension of the first baseline including joint MTL on our dataset.

For each of those we investigate coarse-grained (binary) as well as fine-grained (four classes) hate speech detection.

4.2 Results

The performance results of our optimized models are displayed in Table 6 for the coarse-grained (binary) hate speech detection and in Table 5 for the fine-grained (four classes) hate speech detection. For the different models, we respectively report the class-based F1 score values as well as the macro-averaged F1 score value for hate speech

Training Data	F1$_{NONE}$	F1$_{PRFN}$	F1$_{OFFN}$	F1$_{HATE}$	macro-avg F1
HS	.7018	**.7827**	**.5121**	.5438	**.6351**
HS & Emo	.7100	.7143	.4054	.5436	.5933
HS & Emo & HSEmo	**.7169**	.7522	.4823	.5620	.6283
HS & HSEmo	.7154	.7315	.4509	**.5655**	.6158

Table 5: Fine grained hate speech detection performance of best models trained on different data.

Training Data	F1$_{HOF}$	F1$_{NONE}$	macro-avg F1
HS	.7277	**.8535**	**.7906**
HS & Emo	.7293	.8326	.7810
HS & Emo & HSEmo	**.7340**	.8459	.7900
HS & HSEmo	.7147	.8299	.7723

Table 6: Coarse grained hate speech detection performance of best models trained on different data.

detection on the HASOC 2021 test dataset (Mandl et al., 2021). Detailed results of different runs including hyperparameter optimization are given in Appendix C.

We briefly compare our best results to the performances of the top systems according to the leaderboards from the HASOC 2021 shared task which are available online.[3] The best observed performance of all our models on test data is 0.8187 macro-average F1 for coarse-grained hate speech detection (see Appendix C, Table 8) and 0.6486 macro-average F1 for fine-grained hate speech detection (see Appendix C, Table 9). These runs would place us fourth for coarse-grained detection and third for fine-grained detection, only about 1% and 2% behind the top systems. Thus, we assume that our general approach is competitive, while the hyperparameter optimization of our basic model remains quite simple.

When incorporating emotion classification, the overall results (i.e. the macro-average F1 scores of the optimized models displayed in Table 6 and in Table 5) show that this does not improve the hate speech detection performance (the *HS* approach performs best). However, in the MTL setups, the approaches including joint multi-task learning (*HS & Emo & HSEmo* and *HS & HSEmo*) mostly outperform the standard MTL approach (*HS & Emo*).

5 Conclusion

In total, we present a corpus of 1,000 messages with emotion labels containing also hate speech and offensive language. Our bottom-up analysis of the occurrence of emotions in hate speech shows that, as expected, there is a correlation between certain negative emotions such as disgust and severe hate speech classes. However, we also identified other negative emotions that mostly do not correlate with hate speech, such as sadness. In some cases, we even found that the authors presumably felt positive emotions such as joy in hateful messages.

Our experiments with this preliminary dataset show the benefit of a joint annotation in comparison to standard multi-task learning with multiple datasets. However, since we have only annotated a sample of the hate speech data so far, further research is needed to use such data to improve hate speech detection. Future work has to consider a fair comparison with a fully annotated dataset for joint learning. Further attempts for optimization should consider assigning variable weights to the auxiliary task when the main goal is to improve hate speech detection.

Ethical Considerations

Limitations. Our analysis of the correlation of hate speech and emotions is based on an emotion annotation by only one single annotator. While we extensively discussed difficult cases beforehand and the annotation was carefully done, we currently cannot evaluate the quality of this annotation. In addition, the annotator was indecisive about the emotion in about 4% of the instances. Future plans are to include a second annotation by another annotator.

The dataset used for our analysis and experiment is rather small and contains a topic bias towards Covid-19 in India in particular. This limits the generalizability of our results.

Reproducibility. We use datasets with annotations for hate speech and emotions. All of these datasets are freely available for research use. We use these data for their intended use, to develop detection systems. Since we research hate speech, the datasets have not been filtered or anonymized for offensive language.

[3] https://hasocfire.github.io/hasoc/2021/results.html

We publish our program code for maximum transparency. The described models and predictions of labels can be reproduced with this code. For training we randomly split the dataset into specific portions. Additionally, we provide a script to reproduce the random split used in our experiments to benefit future research. We report relevant information for the used artifacts and refer to the original publications for further documentation. We believe that these descriptions make our approach reproducible.

References

Wafa Alorainy, Pete Burnap, Han Liu, Amir Javed, and Matthew L Williams. 2018. Suspended accounts: A source of tweets with disgust and anger emotions for augmenting hate speech data sample. In *2018 International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 2, pages 581–586. IEEE.

Md Rabiul Awal, Rui Cao, Roy Ka-Wei Lee, and Sandra Mitrović. 2021. AngryBERT: Joint Learning Target and Emotion for Hate Speech Detection. In *Advances in Knowledge Discovery and Data Mining*, pages 701–713, Cham. Springer International Publishing.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500.

Laura-Ana-Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Patricia Chiril, Endang Wahyu Pamungkas, Farah Benamara, Véronique Moriceau, and Viviana Patti. 2022. Emotionally informed hate speech detection: a multi-target perspective. *Cognitive Computation 14*, pages 322–352.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Paul Ekman. 1988. Gesichtsausdruck und Gefühl: 20 Jahre Forschung von Paul Ekman. *Innovative Psychotherapie und Humanwissenschaften*, 38.

Kosisochukwu Judith Madukwe, Xiaoying Gao, and Bing Xue. 2021. What emotion is hate? incorporating emotion information into the hate speech detection task. In *PRICAI 2021: Trends in Artificial Intelligence*, pages 273–286, Cham. Springer International Publishing.

Angel Felipe Magnossão de Paula, Paolo Rosso, Imene Bensalem, and Wajdi Zaghouani. 2022. UPV at the Arabic hate speech 2022 shared task: Offensive language and hate speech detection using transformers and ensemble models. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 181–185, Marseille, France. European Language Resources Association.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Prasenjit Majumder, Johannes Schäfer, Tharindu Ranasinghe, Marcos Zampieri, Durgesh Nandini, and Amit Kumar Jaiswal. 2021. Overview of the HASOC Subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages. In *Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation*, pages 1–19, India. CEUR Workshop Proceedings.

Ilia Markov, Nikola Ljubešić, Darja Fišer, and Walter Daelemans. 2021. Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 149–159, Online. Association for Computational Linguistics.

Ricardo Martins, Marco Gomes, José João Almeida, Paulo Novais, and Pedro Henriques. 2018. Hate speech classification in social media using emotional analysis. In *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 61–66.

Saif Mohammad. 2012. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada. Association for Computational Linguistics.

Flor Miriam Plaza-del Arco, Sercan Halat, Sebastian Padó, and Roman Klinger. 2021. Multi-Task Learning with Sentiment, Emotion, and Target Detection to Recognize Hate Speech and Offensive Language. *In Working Notes of FIRE 2021 – Forum for Information Retrieval Evaluation, December 13-17, 2021, India*.

Santhosh Rajamanickam, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2020. Joint modelling of emotion and abusive language detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4270–4279, Online. Association for Computational Linguistics.

Aneri Rana and Sonali Jha. 2022. Emotion based hate speech detection using multimodal learning. *arXiv preprint arXiv:2202.06218*.

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 (SocialNLP@EACL 2017)*, pages 1–10, Valencia, Spanien.

Julia Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of GermEval Task 2, 2019 Shared Task on the Identification of Offensive Language. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 354–365, Erlangen, Deutschland. German Society for Computational Linguistics & Language Technology.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*, pages 1–10, Wien, Österreich. Österreichische Akademie der Wissenschaften.

A Examples from HS-EMO

Table 7 enumerates examples of Twitter text messages from the HASOC 2021 dataset (Mandl et al., 2021) which we annotate for emotion labels and collect in the HS-EMO corpus.

General Examples for the Emotion Categories. #1 is an example for the emotion category *joy*. The writer of this tweet was presumably happy about something. In the tweet, the person was waiting for something and now they are excited that it is the time of "Babs".

#2 is an example for the emotion category *anger*. The author of this message is probably frustrated, angry and annoyed. The writer compares his actions to the actions of someone else. He uses a swear word calling the other person a "shitbag" in an expression of anger presumably over an unfair treatment.

#3 is an example for the emotion category *disgust*. Here the writer probably feels shame for his country and its people. They are disgusted by the behavior of the people. They feel shame for something and are contemptuous towards someone.

#4 is an example for the emotion category *fear*. This person may be afraid as they see only cruelty around them and no help. They list everything that frightens them and ask for help. The words "dying", "begging", "clueless", "lies" in this context may be alarming and are an indicator for the emotion fear.

#5 is an example for the emotion category *surprise*. Here the writer is confused by the actions of a certain group.

#6 is an example for the emotion category *sadness*. The writer of this message is saddened by the situation in India. They suffer with their fellow human beings. Words such as "hurting me", "hope", "god save us" may be indicators of sadness.

#7 and #8 are examples where no emotion could be detected. From these tweets it is impossible to tell the emotional state of the author without additional context. Are they pleased, disappointed or angry? This distinction is not evident from the content of the message alone.

Borderline Cases. Potentially ambiguous cases were mostly between the emotion categories *disgust* vs. *anger*, *anger* vs. *sadness* and *disgust* vs. *sadness*. The stronger emotion was selected when multiple emotions could be detected. The main goal was to identify which one was the guiding motivation for writing the tweet.

Example #9 is annotated it with *sadness*, however, it could also be *disgust*. The writer is presumably sad and at the same time ashamed of his government. However, they probably wrote this tweet out of sadness. Thus, this emotion is stronger in this example.

Example #10 is annotated with *anger* while at first glance it could also be *disgust*. However, the anger the person feels seems to be stronger and ultimately the reason for writing the message.

B Hyperparameters

We pad/truncate instances to the length of 103 tokens. We determined this value by the 99th percentile of instance lengths in the HASOC 2021 dataset.

In all our experiments we use a batch size of 8 and apply a dropout (probability 0.2) to the output of the encoder. For optimization we use the Adam optimizer with default parameters.

We reserve 10% of the HASOC 2021 data as validation dataset to determine an optimal early

#	Text	Emotion
1	"@USER @USER Because karma is a bitch. Babs' time had finally come. The Wanker. #LGRW"	*joy*
2	"@USER Here's hoping. That little shitbag gets arrested. I got arrested for "threatening" when I didn't even make a direct threat, and our law system is furbar if crap like this keeps sliding."	*anger*
3	"@USER @USER @USER You guys make me sick to the core!!!! Is that really your concern right now, formation of alliance, when the entire country is on its knees!!! I guess news of dead bodies pilling up, people dropping dead on the st"	*disgust*
4	"People are dying, left to go begging for basic medical resources. No ventilators. No O2. Delayed/No Response from Centre. State govts clueless due lack of aid from Centre. Lies. Cover ups. Please, for the countrys sake, #ResginModi & let someone more competent do the job. URL"	*fear*
5	"What the heck the bjp is doing... destroying people life? #ResignPMmodi #BjpDestroyedIndia #BJP #prayaraj"	*surprise*
6	"@USER Just tired of all these deaths hurting me from inside hope good days will come back :(may god save us all #COVIDSecond #COVIDSecondWAVE #COVID119India #COVID19 #OxygenEmergency #IndiaFightsCorona #IndiaFightsCOVID19 #CovidVaccine #Covid19IndiaHelp #COVIDSecondWaveInIndia #indianeedoxygen"	*sadness*
7	"@USER Did you get the old bastard 1 or the young gun 1"	_
8	"@USER @USER We need Ethan Winters to say it too"	_
9	"I feel devastated for India and deeply ashamed of our Government'a attitude and actions. #IndiaCovidCrisis URL"	*sadness*
10	"I am ashamed that I was blind supporter of @USER Your People are dying , Gang Raped and You are doing this Shit ? #SpinelessBJP #spinelessmodi #MamtaisTerrorist #BengalViolence #BengalBurning"	*anger*

Table 7: Examples of annotated text instances from our corpus (HS-EMO). Username mentions and URLs have been anonymized.

stopping epoch (patience 3, minimum delta 0.005) with a maximum of 10 training epochs. To be able to use the same data split when training on our dataset (which is a sample from the HASOC 2021 data), we ensure that the validation data is sampled from the HASOC 2021 data instances which are not included in our dataset. The remaining 90% of the HASOC 2021 data is used for training (training data *HS*).

We run hyperparameter optimization by selecting the best learning rate based on validation dataset performance. We test the following ten different values for the learning rate: $1e{-}7$, $2.5e{-}7$, $5e{-}7$, $7.5e{-}7$, $1e{-}6$, $2.5e{-}6$, $5e{-}6$, $7.5e{-}6$, $1e{-}5$, $2.5e{-}5$.

C Detailed Experimental Results

Table 8 shows the performance of the different approaches for different learning rate values at coarse grained hate speech and emotion classification. Table 9 shows the performance of the different approaches for different learning rates at fine grained hate speech and emotion classification. In both tables we report the performance of the different models on the validation dataset which has been used for early stopping and learning rate optimization (test data: Val HS) as well as the performance on the HASOC 2021 (Mandl et al., 2021) test dataset (test data: Test HS). The best macro-averaged F1 scores for hate speech detection on the validation dataset are underlined for each training data setup (best learning rate value). The last column in each of the two tables shows the macro-averaged F1 score for emotion classification on our dataset (test data: HS-EMO). Note that for some runs (training data: HSEmo) this dataset is also used during training.

		Test Data:	Val HS			Test HS			HS-EMO
Training Data	Epochs	lr	$F1_{HOF}$	$F1_{NONE}$	macro-avg $F1_{HS}$	$F1_{HOF}$	$F1_{NONE}$	macro-avg $F1_{HS}$	$F1_{Emo}$
HS	10	1e-07	.6200	.7120	.6660	.6416	.7109	.6762	–
HS	10	2.5e-07	.6800	.7990	.7400	.6788	.7967	.7377	–
HS	8	5e-07	.7200	.8540	.7870	.7119	.8443	.7781	–
HS	7	7.5e-07	.7340	.8680	.8010	.7221	.8551	.7886	–
HS	7	1e-06	.7290	.8660	.7980	.7213	.8517	.7865	–
HS	2	2.5e-06	.7090	.8570	.7830	.6987	.8443	.7715	–
HS	2	5e-06	.7540	.8750	.8140	.7277	.8535	.7906	–
HS	1	7.5e-06	.7370	.8610	.7990	.7329	.8462	.7896	–
HS	1	1e-05	.7430	.8590	.8010	.7707	.8667	.8187	–
HS	1	2.5e-05	.7540	.8560	.8050	.7623	.8486	.8054	–
HS & Emo	3	1e-07	.4950	.6350	.5650	.5019	.6416	.5717	.0649
HS & Emo	10	2.5e-07	.6140	.6870	.6510	.6341	.6939	.6640	.0759
HS & Emo	8	5e-07	.6560	.8260	.7410	.6565	.8095	.7330	.1163
HS & Emo	9	7.5e-07	.7170	.8510	.7840	.7078	.8352	.7715	.1742
HS & Emo	8	1e-06	.7090	.8570	.7830	.6861	.8323	.7592	.1802
HS & Emo	3	2.5e-06	.6800	.8490	.7650	.6818	.8459	.7639	.1449
HS & Emo	4	5e-06	.7140	.8580	.7860	.7435	.8646	.8041	.2104
HS & Emo	3	7.5e-06	.7100	.8650	.7870	.7315	.8634	.7974	.1916
HS & Emo	2	1e-05	.7330	.8470	.7900	.7293	.8326	.7810	.1923
HS & Emo	3	2.5e-05	.7130	.8640	.7880	.7146	.8517	.7832	.2279
HS & Emo & HSEmo	3	1e-07	.3790	.6480	.5140	.3671	.6399	.5035	.0606
HS & Emo & HSEmo	10	2.5e-07	.6490	.7810	.7150	.6524	.7674	.7099	.3482
HS & Emo & HSEmo	10	5e-07	.7210	.8500	.7860	.6881	.8214	.7547	.4280
HS & Emo & HSEmo	5	7.5e-07	.7070	.8330	.7700	.7226	.8290	.7758	.4483
HS & Emo & HSEmo	5	1e-06	.7130	.8430	.7780	.7164	.8283	.7723	.5602
HS & Emo & HSEmo	2	2.5e-06	.7090	.8320	.7710	.7238	.8223	.7731	.4256
HS & Emo & HSEmo	3	5e-06	.7700	.8760	.8230	.7340	.8459	.7900	.8291
HS & Emo & HSEmo	4	7.5e-06	.7290	.8540	.7910	.7227	.8330	.7779	.9824
HS & Emo & HSEmo	2	1e-05	.7390	.8610	.8000	.7360	.8310	.7835	.8824
HS & Emo & HSEmo	3	2.5e-05	.7490	.8670	.8080	.7109	.8380	.7744	.9871
HS & HSEmo	9	1e-07	.5880	.5960	.5920	.5991	.5726	.5858	.1359
HS & HSEmo	10	2.5e-07	.6300	.7780	.7040	.6387	.7679	.7033	.2889
HS & HSEmo	9	5e-07	.7460	.8520	.7990	.7044	.8196	.7620	.5469
HS & HSEmo	9	7.5e-07	.7570	.8670	.8120	.7147	.8299	.7723	.6352
HS & HSEmo	8	1e-06	.7380	.8480	.7930	.7209	.8225	.7717	.7001
HS & HSEmo	2	2.5e-06	.7440	.8560	.8000	.7193	.8313	.7753	.6710
HS & HSEmo	3	5e-06	.7380	.8670	.8030	.7194	.8486	.7840	.8977
HS & HSEmo	1	7.5e-06	.7280	.8690	.7990	.7208	.8555	.7881	.6214
HS & HSEmo	1	1e-05	.7360	.8580	.7970	.7458	.8517	.7987	.6431
HS & HSEmo	1	2.5e-05	.7400	.8510	.7950	.7325	.8282	.7804	.8524

Table 8: Coarse grained (binary) hate speech and emotion classification performance.

		Test Data:	Val HS					Test HS					HS-EMO
Training Data	Epochs	lr	$F1_{NONE}$	$F1_{PRFN}$	$F1_{OFFN}$	$F1_{HATE}$	macro-avg $F1_{HS}$	$F1_{NONE}$	$F1_{PRFN}$	$F1_{OFFN}$	$F1_{HATE}$	macro-avg $F1_{HS}$	$F1_{Emo}$
HS	9	1e-07	.1930	.7040	.0430	.4640	.3510	.1856	.6965	.0153	.4190	.3291	–
HS	10	2.5e-07	.5590	.7480	.0770	.5230	.4770	.4986	.7566	.0957	.5114	.4656	–
HS	10	5e-07	.5410	.7470	.3690	.5060	.5410	.5544	.7553	.2073	.4955	.5031	–
HS	10	7.5e-07	.6940	.7750	.5540	.5920	.6530	.6473	.7532	.3973	.5325	.5826	–
HS	10	1e-06	.7040	.7760	.5360	.5520	.6420	.7079	.7798	.5012	.5455	.6336	–
HS	8	2.5e-06	.7370	.7680	.5470	.6420	.6730	.7182	.7539	.4822	.5343	.6221	–
HS	4	5e-06	.7690	.7750	.5690	.5830	.6740	.7091	.7598	.4849	.5475	.6253	–
HS	2	7.5e-06	.7510	.7700	.5670	.6470	.6840	.7018	.7827	.5121	.5438	.6351	–
HS	3	1e-05	.7360	.7410	.5560	.6170	.6620	.7307	.7582	.5246	.5808	.6486	–
HS	4	2.5e-05	.7450	.7660	.5650	.6460	.6800	.7119	.7503	.4452	.5285	.6090	–
HS & Emo	10	1e-07	.5630	.7160	.3970	.0900	.4420	.5685	.7193	.2373	.1254	.4126	.1063
HS & Emo	9	2.5e-07	.5480	.7610	.0450	.4460	.4500	.5072	.7625	.0199	.4603	.4375	.0637
HS & Emo	10	5e-07	.5780	.7670	.1430	.5510	.5100	.5452	.7634	.1185	.5131	.4851	.1416
HS & Emo	9	7.5e-07	.6470	.7660	.3260	.5470	.5710	.6062	.7668	.2890	.4970	.5397	.1114
HS & Emo	7	1e-06	.6580	.7700	.4690	.5690	.6170	.6131	.7664	.3021	.4960	.5444	.0694
HS & Emo	9	2.5e-06	.7420	.7570	.5100	.5940	.6510	.7086	.7378	.4467	.5666	.6149	.1574
HS & Emo	4	5e-06	.7200	.7920	.5690	.6480	.6820	.7036	.7665	.5149	.5750	.6400	.0958
HS & Emo	3	7.5e-06	.7270	.7720	.5760	.5690	.6610	.7256	.7748	.4692	.5639	.6334	.1511
HS & Emo	2	1e-05	.7000	.7810	.5270	.6220	.6580	.7057	.7690	.4847	.5726	.6330	.1363
HS & Emo	3	2.5e-05	.7840	.7500	.5790	.6330	.6870	.7100	.7143	.4054	.5436	.5933	.1555
HS & Emo & HSEmo	9	1e-07	.2210	.7190	.0100	.4530	.3510	.2927	.7238	.0591	.4600	.3839	.1069
HS & Emo & HSEmo	10	2.5e-07	.5790	.7670	.1020	.5030	.4880	.5419	.7342	.0648	.4925	.4584	.1813
HS & Emo & HSEmo	10	5e-07	.7130	.7650	.3690	.5810	.6070	.6553	.7577	.3446	.5171	.5687	.5139
HS & Emo & HSEmo	10	7.5e-07	.6940	.7470	.4300	.5920	.6160	.6943	.7665	.3765	.5521	.5974	.6320
HS & Emo & HSEmo	5	1e-06	.6840	.7570	.5040	.5890	.6330	.6561	.7537	.3259	.5348	.5676	.3449
HS & Emo & HSEmo	8	2.5e-06	.7360	.7320	.5300	.5920	.6470	.7053	.7116	.3187	.5191	.5637	.7907
HS & Emo & HSEmo	4	5e-06	.7110	.7280	.5710	.6540	.6660	.7174	.7000	.3910	.5420	.5876	.8134
HS & Emo & HSEmo	2	7.5e-06	.7460	.7430	.5470	.5930	.6570	.7027	.7306	.4346	.5223	.5975	.7409
HS & Emo & HSEmo	2	1e-05	.7500	.7710	.5670	.6270	.6790	.7169	.7522	.4823	.5620	.6283	.8000
HS & Emo & HSEmo	2	2.5e-05	.7420	.7630	.5550	.5830	.6610	.7242	.7412	.4256	.5407	.6079	.9574
HS & HSEmo	10	1e-07	.4780	.7320	.2190	.4950	.4810	.4394	.7023	.1885	.4098	.4350	.1426
HS & HSEmo	10	2.5e-07	.6320	.7560	.3060	.5380	.5580	.6065	.7525	.2582	.4861	.5259	.2588
HS & HSEmo	10	5e-07	.6060	.7690	.4300	.5700	.5940	.5977	.7735	.2856	.4745	.5328	.3308
HS & HSEmo	6	7.5e-07	.7040	.7630	.5170	.6110	.6490	.6667	.7384	.3404	.5230	.5671	.3567
HS & HSEmo	10	1e-06	.7130	.7690	.4660	.5770	.6310	.6777	.7320	.4072	.5125	.5824	.5114
HS & HSEmo	3	2.5e-06	.7350	.7520	.5260	.5870	.6500	.6987	.7413	.4319	.5501	.6055	.4416
HS & HSEmo	4	5e-06	.7560	.7440	.5840	.6280	.6780	.7154	.7315	.4509	.5655	.6158	.8005
HS & HSEmo	2	7.5e-06	.7230	.7080	.5670	.6620	.6650	.7022	.7322	.3357	.5444	.5787	.6371
HS & HSEmo	3	1e-05	.7620	.7210	.5860	.6280	.6740	.7308	.7241	.3772	.5751	.6018	.9246
HS & HSEmo	2	2.5e-05	.7050	.7540	.5980	.5970	.6640	.7093	.7256	.3492	.5547	.5847	.9521

Table 9: Fine grained hate speech and emotion classification performance.