eevvgg at SemEval-2023 Task 11: Offensive Language Classification with Rater-based Information

Ewelina Gajewska
Faculty of Psychology and Cognitive Sciences
Adam Mickiewicz University
ewegaj@st.amu.edu.pl

Abstract

A standard majority-based approach to text classification is challenged with an individualised approach in the Semeval-2023 Task 11. Here, disagreements are treated as a useful source of information that could be utilised in the training pipeline. The team proposal makes use of partially disaggregated data and additional information about annotators provided by the organisers to train a BERT-based model for offensive text classification. The approach extends previous studies examining the impact of using raters' demographic features on classification performance (Hovy, 2015) or training machine learning models on disaggregated data (Davani et al., 2022). The proposed approach was ranked 11 across all 4 datasets, scoring best for cases with a large pool of annotators (6th place in the MD-Agreement dataset) utilising features based on raters' annotation behaviour.

1 Introduction

Semeval-2023 Task on Learning With Disagreements (Le-Wi-Di) challenges the standard approach in natural language processing (NLP) that there is a single interpretation of language (Leonardellli et al., 2023). At the same time, it follows a recent trend in computational linguistics – a perspectivist approach (Basile et al., 2021; Abercrombie et al., 2022). Different opinions and views are taken into account here, when studying language phenomena.

This change from a majority-based to individualised approach is particularly important in the case of studying subjective phenomena such as abusive/offensive/hateful language. Studies show that experts and amateur raters have different strategies for hate speech annotation (Waseem, 2016). Sap et al. (2022) in turn found that some individuals are more likely to mark African American English dialect as toxic language.

The approach proposed by the eevvgg team follows previous studies in this area (Kocoń et al., 2021). In the case of all four datasets (sub-tasks) a model is trained on partially disaggregated data and additional features created with rater-based information. All four detection tasks fall into a broad definition of offensive language detection, however the *HS-Brexit* dataset (Akhtar et al., 2021) is annotated with hate speech, a task for the *ConvAbuse* dataset (Curry et al., 2021) regards abusiveness detection, the *ArMIS* dataset (Almanea and Poesio, 2022) focuses on misogyny and sexism detection, and the annotation task in the *MD-Agreement* dataset (Leonardelli et al., 2021) involves offensiveness detection.

The ArMIS dataset comprises of Arabic texts, and the other three contain English data. In regard to the MD-Agreement dataset, a pool of annotators is larger than in other datasets – the corpus was annotated by over 800 individuals compared with 3 to 8 raters in the case of other datasets. Therefore, the task could be regarded as more challenging in this case.

The proposed system utilises additional information provided in the datasets (besides text) that are employed to model disagreements between annotators. Therefore, information about raters and their annotation behaviour is utilised to train a BERT-based model.

2 Background

Hovy (2015) found classification performance could be improved with the use of demographic factors in machine learning models. Davani et al. (2022) propose to approach the detection of emotions on disaggregated data as a multi-task classification with an individual classification layer for each annotator on the one hand, and an ensemble model on the other hand.

Regarding state-of-the-art performance on offensive language detection, BERT-based models systematically achieved best results in the previous SemEval editions (Zampieri et al., 2020). Regard-

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 171–176

ing additional data, the use of hate speech lexicons was a popular option.

3 System Overview

BERT (Devlin et al., 2019) is used as the classification model. Specifically, BERTweet[1] version for the English language data and BERT-base for Arabic[2] available in the Transformers library (Wolf et al., 2020).

Pre-trained models are fine-tuned separately on each dataset. Specifically, BERT encodings from the CLS token are fed to a first fully-concatenated network, followed by a dropout layer. Then, additional features are concatenated and fed to a second fully-concatenated network. Lastly, there is a classification layer for a binary prediction with softmax activation. Architecture of a model employed for the task is presented in Figure 1.

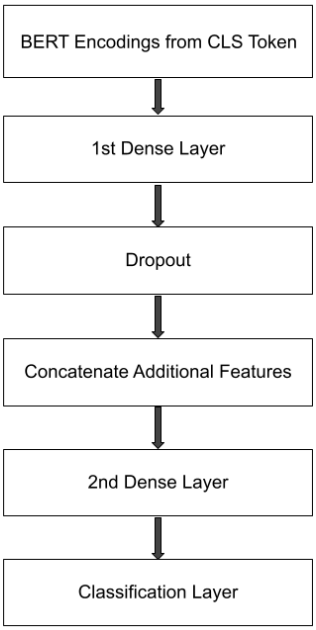

Figure 1: Proposed model.

Train and dev data splits were combined and used to train a BERT model. Individual annotations were retrieved from data files instead of majority voted labels. Then, instances of text that raters did not unanimously agree on the label were used in the train data twice – first with the label "1", and second with the label "0". Otherwise one instance of text with the agreed label was used in the training set. Thus, data fed to a model is partially disaggregated, i.e. if not all annotators agreed on the label, a text sample with both labels "0" and "1" is used for training purposes.

In regard to data pre-processing, minor changes were applied to the text. In ConvAbuse "prev_agent", "prev_user", "agent", and "user" tokens were removed, so that only text content was extracted and fed to a BERT model. In HS-Brexit and MD-Agreement multiple occurrences of "<user>" token were replaced with a single instance of "user" token.

4 Experimental Setup

Besides text, rater-based features were engineered by making use of the additional information ("other info") provided by the organisers in the data files. Thus, in regard to each of the four dataset, different features were fed to the model besides text encodings from BERT.

Hyperparameters set for BERT-based models are presented in Tab. 1. Model architectures are based on similar systems developed for classification tasks (Awal et al., 2021; Plaza-Del-Arco et al., 2021) as well as recommendations provided by the authors of the Transformers library used for model training (Wolf et al., 2020). In the Transformers library an additional layer added for the classification purposes comprises 768 nodes, and similar works propose to lower the size for the consecutive layers of a classification model. Pre-experimental sessions were conducted on train and dev splits (a train split used for training and a dev split used for evaluation purposes) in order to validate the settings. Different hyperparameters were established for the layers of the model in the case of the ArMIS dataset because of the underperformance of the system on the initial settings compared to the other three datasets (other systems approached or reached .80 micro-F_1). Initial number of nodes were lowered by half, and then gradually added to the first and the second fully-connected layers until the system approach .80 micro-F_1. In addition, all systems were trained for 2 or 3 epochs and the better performing alternative was used in the final settings. The final BERT-based models were developed on a combined train and dev splits.

In the HS-Brexit dataset additional data comprises annotations for aggressive and offensive language, and raters group identity (target vs. control group). The former is utilised in the proposed

[1] https://huggingface.co/vinai/bertweet-covid19-base-uncased

[2] https://huggingface.co/asafaya/bert-base-arabic

system – for each text sample a number of positive labels for aggressive and offensive language annotation is calculated. In addition, specific keywords for the hate speech class are extracted. Here, words that appear exclusively in the positive ("1") class with a minimum document occurrence of 3 in a combined train and dev splits are considered as keywords. Then, this feature is binarised, i.e. marked as "1" in cases with at least 1 keyword in a text sample, and "0" otherwise. In order to extract those keywords, text was lemmatised with the use of spaCy library[3].

In the ArMIS dataset there is no variability in the case of annotators pool and every sample was rated by the same group of individuals. Therefore, a different approach for feature engineering was employed. In regard to the ArMIS dataset, keywords for the misogyny and sexism class were extracted as an additional feature to text encodings from BERT. In particular, words with a high precision score for the positive class were retrieved. Here, a precision score is calculated as a number of occurrences of a word in the positive class divided by the overall number of occurrences of this word in data samples. Then, words above a selected threshold are extracted. Here, 70% threshold was chosen based on performance in experimental trials. As a result, 1429 of such words were retrieved.

HParameter	HS-Brexit	ArMIS	ConvAbuse	MD-Agree
1st Dense	768	500	768	768
2nd Dense	246	128	246	246
Dropout			0.3	
No. epochs	3	2	2	2
Learning rate			4e-5	
Batch size			10	

Table 1: Experimental setup of the proposed models.

Similar approach is employed in the case of the ConvAbuse dataset, where such keywords are extracted for the abusive class, as well as a selected set of additionally annotated labels (homophobic, intellectual, racist, sexist, sex harassment, target.generalised, target.individual, target.system, explicit, implicit; other categories were annotated very infrequently). Regarding the abusive class 55% threshold and minimum document occurrence of 2 was chosen (55 keywords were extracted here); Threshold of 51% was chosen for the additional categories[4]. Then, keywords extracted for each text instance were transformed into a binary feature separately for each category (marked as "1" if at least 1 keyword was extracted and "0" otherwise). As a result, 11 binary features were obtained.

The MD-Agreement dataset is annotated by a large set of raters (over 800) which on the one hand, is challenging, and on the other hand, suits the perspectivist approach very well. Here, annotation on offensiveness of each individual rater is compared against the majority-voted labels. Four different metrics (Cohen's kappa, accuracy, precision, recall) are employed for this purpose, i.e. calculation of agreement between individual raters and the majority[5]. In a sense it allows to model reliability of individual raters against the opinion of the majority. Finally, individual scores were averaged over a text instance as each sample was annotated by several raters.

5 Results

The team is officially ranked 11 across 4 datasets, scoring best for the MD-Agreement dataset (6th place) in terms of cross entropy. Official results from the test sets are presented in Table 2.

In addition to the official metrics, macro-averaged F_1 is reported on the test split as it weights equally both labels and is especially informative about the system performance in the case of imbalanced distribution of categories. Results indicate that the proposed system achieves good performance for both classes, particularly in the case of ConvAbuse and MD-Agreement datasets (macro-F_1 approaching 0.9 and 0.8, respectively).

In Figure 2 confusion matrices for each of 4 datasets are presented. Regarding incorrect classification, all 4 models underperform for the positive category (predict "0" when the true label is "1"). Regarding soft evaluation, 75th percentile falls between 0.11 and 0.36 in the ConvAbuse and ArMIS datasets, respectively, measured as the mere difference between the true distribution and predicted probability of two categories.

Regarding correct classification by the proposed system, examples from the test data split include the following cases of (almost) perfect classification in terms of distribution of labels:

[3] https://spacy.io; version 3.4.1

[4] For example, "moron" and "imbecile" were extracted as keywords for intellectual category, "histeric" and "b*tch" for sexist category, and "faggit" and "homo" for homophobic category

[5] Scikit was used to calculate these metrics https://scikit-learn.org/

Metric / Approach	Average	HS-Brexit	ArMIS	ConvAbuse	MD-Agree
Micro-F_1					
Best per dataset	.89	.93	.85	.94	85
Baseline	.63	.84	.42	.74	.53
eevvgg	.84 (+33%)	.86 (+2%)	.74 (+76%)	.92 (+24%)	.82 (+55%)
Macro-F_1					
eevvgg	.77	.68	.73	.87	.79
CE					
Best per dataset	.34	.24	.47	.19	.47
Baseline	5.62	2.72	8.91	3.48	7.39
eevvgg	.41	.33	.56	.25	.50

Table 2: Micro-averaged F_1-score (hard eval) and cross entropy (soft eval) of team eevvgg per test dataset. Values in parentheses indicate percentage over baseline score in terms of micro-F_1. Approaches in grey are shown for comparison: organiser's baseline and the best team approach. In addition, macro-F_1 score is depicted for the team submission.

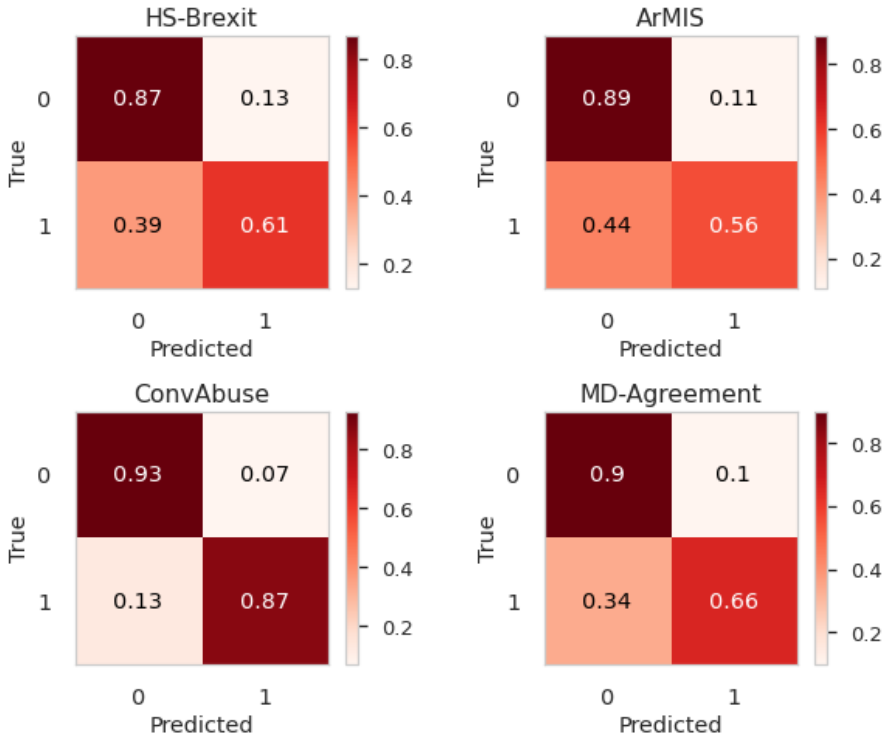

Figure 2: Normalised confusion matrices.

HS-Brexit:

- C'mon Lymies, if you don't #Brexit now then no one will be left to team up with America when Europe falls to Islam <url> (hate speech)
- <user> <user> It seems our Govt + of the Nom for POTUS has gotten major funding from terrorist countries. Wonder not on #Brexit GB (non hate speech)

ConvAbuse:

- "prev_agent": "_", "prev_user": "hi my name is solay", "agent": "I have told you before, I do not care about names.", "user": "you bitch" (abusive)
- "prev_agent": "How long have you been looping?", "prev_user": "I don't loop. You, the retarded chat bot do.", "agent": "Why do you not loop i the retarded chat bot do?", "user": "That makes no sense" (non abusive)

MD-Agreement:

- This is not a game <user> Moron! You worry

more about your stupid hats than 170,000 American lives lost from covid...go away you imbecile... (offensive)

- <user> <user> Biden needs to look in the mirror talk about ugly after all he has done which was to help his family get rich. (non offensive)

Most of the correctly classified positive cases contain sensitive or slur words which could be easily separated from cases that do not contain them. However, one can find potentially offensive words such as "terrorist" and retarded also in texts classified as non-offensive/abusive/hate speech which is more challenging for machine learning classifiers. The proposed system was able to correctly detect those cases.

On the other hand, there are several cases when the system was completely wrong – predicted (almost) the opposite distribution of labels:

HS-Brexit:

- The ppl of the U.K. are cleary fed up with Muslim Immigration that refuse to assimulate. How many more #refugees? <url> (true: non hate speech (1); predicted: hate speech (0.65))[6]

- More on how some migrant men, posing as children, behave i public schools in Canada <url> <user> <user> #CCOT #Brexit (true: hate speech (0.67); predicted: non hate speech (0.93))

ConvAbuse:

- "prev_agent": "_", "prev_user": "_", "agent": "_", "user": "F off George Soros puppets" (true: non abusive (1); predicted: abusive (0.87))

- "prev_agent": "Why do you say your bad?", "prev_user": "i am sorry,.", "agent": "How long have you been sorry?", "user": "im not to fond of blacks" (true: abusive (1); predicted: non abusive (0.91))

MD-Agreement:

- Supporting #blacklivesmatter is supporting a racist hate group. Sorry not sorry (true: offensive (1); predicted: non offensive (0.85))

- Video to video, would u rather serve under Gen CQ Brown Jr or this clown? #BunkerBoy #DotardTrump #TrumpBullShit #ResignNowTrump #TrumpVirus #TrumpBullShit <url> (true: non offensive (1); predicted: offensive (0.76))

Some of the cases incorrectly classified by the system are tricky – for example, put a person or a group in a negative frame and use "disguised" swear words. On the other hand, some texts with a negative label put the mentioned entity in a negative frame as well as in the case of the last example from the MD-Agreement dataset. The system performed worse on those challenging cases.

6 Conclusion

Although the proposed system is simple in its architecture and employed features, it achieves good performance in terms of both F_1 metrics as well as soft evaluation in the official ranking (see Table 2). In future experiments some form of pre-training on larger datasets could be utilised to further adapt the system to the domain of offensive/abusive/hate speech. Although in pre-experimental sessions conducted with the use of HateBERT instead of BERTweet, it did not show improvement in the system performance.

The proposed approach could be further tested in the prediction of labels for individual raters which could be a step towards personalised hate speech detection and filtering systems.

References

Gavin Abercrombie, Valerio Basile, Sara Tonelli, Verena Rieser, and Alexandra Uma. 2022. Proceedings of the 1st workshop on perspectivist approaches to nlp@ lrec2022. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*.

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. *arXiv preprint arXiv:2106.15896*.

Dina Almanea and Massimo Poesio. 2022. Armis-the arabic misogyny and sexism corpus with annotator subjective disagreements. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2282–2291.

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*

[6]Values in parentheses indicate true distribution of soft label and probability of the label predicted by the proposed system, respectively

Knowledge Discovery and Data Mining: 25th Pacific-Asia Conference, PAKDD 2021, Virtual Event, May 11–14, 2021, Proceedings, Part I, pages 701–713. Springer.

Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. Toward a perspectivist turn in ground truthing for predictive computing. *arXiv preprint arXiv:2109.04270*.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. Convabuse: Data, analysis, and benchmarks for nuanced detection in conversational ai. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (Volume 1: Long papers)*, pages 752–762.

Jan Kocoń, Alicja Figas, Marcin Gruza, Daria Puchalska, Tomasz Kajdanowicz, and Przemysław Kazienko. 2021. Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing & Management*, 58(5):102643.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539.

Elisa Leonardellli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Massimo Poesio, Verena Rieser, and Alexandra Uma. 2023. SemEval-2023 Task 11: Learning With Disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Flor Miriam Plaza-Del-Arco, M Dolores Molina-González, L Alfonso Ureña-López, and María Teresa Martín-Valdivia. 2021. A multi-task learning approach to hate speech detection leveraging sentiment analysis. *IEEE Access*, 9:112478–112489.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Cagri Coltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447. Association for Computational Linguistics.