Automated Hate Speech Target Identification

Andraž Pelicon*
andraz.pelicon@ijs.si
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia

Blaž Škrlj*
blaz.skrlj@ijs.si
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia

Petra Kralj Novak*
petra.kralj.novak@ijs.si
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia

ABSTRACT

We present a new human-labelled Slovenian Twitter dataset annotated for hate speech targets and attempts to automated hate speech target classification via different machine learning approaches. This work represents, to our knowledge, one of the first attempts to solve a Slovene-based text classification task with an autoML approach. Our results show that the classification task is a difficult one, both in terms of annotator agreement and in terms of classifier performance. The best performing classifier is SloBERTa-based, followed by AutoBOT-neurosymbolic-full.

KEYWORDS

hate speech targets, autoML, text features spaces

1 INTRODUCTION

Hate speech and offensive content has become pervasive in social media and has become a serious concern for government organizations, online communities, and social media platforms [13]. Due to the amount of user-generated content steadily increasing, the research community has been focusing on developing computational methods to moderate hate speech on online platforms [6, 1, 8]. While several of the proposed methods achieve good performance on distinguishing hateful and respectful content, several important challenges remain, some of them related to the data itself. Several studies report both low amounts of hate speech instances in the labelled datasets, as well as relatively low agreement scores between annotators [9]. The low agreement score between annotators indicates that recognizing hate speech is a hard task even for humans suggesting that this task requires a more broad semantic interpretation of the text and its context beyond simple pattern matching of linguistic features.

To test this assumption, we have gathered a new Slovenian dataset containing tweets annotated for hate speech targets [1]. This dataset builds on the dataset used for detecting hate speech communities [3] and topics [2] on Slovenian Twitter. The dataset is available in the clarin.si dataset repository with the handle: https://www.clarin.si/repository/xmlui/handle/11356/1398.

Next, we addressed the hate speech target classification task by the autoML approach autoBOT [10]. The key idea of autoBOT is that, instead of evolving at the learner level, evolution is conducted at the representation level. The proposed approach consists of an evolutionary algorithm that jointly optimizes various sparse representations of a given text (including word, subword, POS tag, keyword-based, knowledge graph-based and relational features) and two types of document embeddings (non-sparse representations). To our knowledge, this is one of the first attempts to solve a Slovene-based text classification task with an autoML approach. Finally, we trained a model based on the SloBERTa pre-trained language model [11], a state-of-the-art transformer-based language model pre-trained on a Slovenian corpus and a set of baselines.

Our results show that the context-aware SloBERTa model significantly outperforms all the other models. This result, together with the lower inter-annotator scores, confirms our initial assumption that hate speech target identification is a complex semantic task that requires a complex understanding of the text that goes beyond simple pattern matching. The SloBERTa model reaches annotator agreement in terms of classification accuracy, indicating a fair performance of the model.

2 DATA

We collected almost three years worth of all Slovenian Twitter data in the period from December 1, 2017, to October 1, 2020, in total 11,135,654 tweets. The period includes several government changes, elections and the first Covid-19-related lockdown. We used the TweetCat tool [5], which is developed for harvesting Twitter data of less frequent languages.

2.1 Annotation Schema

Our annotation schema is adapted from OLID [13] and FRENK [4]. It is a two-step annotation procedure. After reading a tweet, without any context, the annotator first selects the type of speech. We differentiate between the following **speech types**:

- **0 acceptable** - non hate speech type: speech that does not contain uncivil language;
- **1 inappropriate** - hate speech type: contains terms that are obscene, vulgar but the text is not directed at any person specifically;
- **2 offensive** - hate speech type: including offensive generalization, contempt, dehumanization, indirect offensive remarks;
- **3 violent** - hate speech type: author threatens, indulges, desires or calls for physical violence against a target; it also includes calling for, denying or glorifying war crimes and crimes against humanity.

If the annotator chooses either the offensive or violent hate speech type, they also include one of the twelve possible targets of hate speech:

- Racism (intolerance based on nationality, ethnicity, language, towards foreigners; and based on race, skin color)
- Migrants (intolerance of refugees or migrants, offensive generalization, call for their exclusion, restriction of rights, non-acceptance, denial of assistance ...)
- Islamophobia (intolerance towards Muslims)

*All authors contributed equally to this research.
[1]Slovenian Twitter dataset 2018-2020 1.0: http://hdl.handle.net/11356/1423

Information Society 2021, 4–8 October 2021, Ljubljana, Slovenia

- Antisemitism (intolerance of Jews; also includes conspiracy theories, Holocaust denial or glorification, offensive stereotypes …)
- Religion (other than above)
- Homophobia (intolerance based on sexual orientation and/or identity, calls for restrictions on the rights of LGBTQ persons
- Sexism (offensive gender-based generalization, misogynistic insults, unjustified gender discrimination)
- Ideology (intolerance based on political affiliation, political belief, ideology… e.g. "communists", "leftists", "home defenders", "socialists", "activists for…")
- Media (journalists and media, also includes allegations of unprofessional reporting, false news, bias)
- Politics (intolerance towards individual politicians, authorities, system, political parties)
- Individual (intolerance towards any other individual due to individual characteristics; like commentator, neighbor, acquaintance)
- Other (intolerance towards members of other groups due to belonging to this group; write in the blank column on the right which group it is)

2.2 Sampling for Training and Evaluation

The training set is sampled from data collected before February 2020. The sampling was intentionally biased to contain as much hate speech as possible in order to obtain enough organic examples to train the model successfully. A simple model was used to flag potential hate speech content, and additionally, filtering by users and by tweet length (number of characters) was applied. 50,000[2] tweets were selected for annotation.

The evaluation set is sampled from data collected between February 2020 and August 2020. Contrary to the training set, the evaluation set is an unbiased random sample. Since the evaluation set is from a later period compared to the training set, the possibility of data linkage is minimized. Furthermore, the estimates of model performance made on the evaluation set are realistic, or even pessimistic, since the model is tested on a real-world distribution of data where hate speech is less prevalent than in the biased training set. The evaluation set is also characterized by a new topic, COVID-19; this ensures that our model is robust to small contextual shifts that may be present in the test data. For the evaluation set, 10,000 tweets were selected to be annotated.

2.3 Annotation Procedure

Each tweet was annotated twice: In 90% of the cases by two different annotators (to estimate inter-annotator agreement) and in 10% of the cases by the same annotator (to assess the self-agreement). Special attention was devoted to an evening out the overlap between annotators to get agreement estimates on equally sized sets. Ten annotators were engaged for our annotation campaign. They were given annotation guidelines, a training session and a test on a small set to evaluate their understanding of the task and their commitment before starting the annotation procedure. The annotation process lasted four months, and it required about 1,200 person-hours for the ten annotators to complete the task.

In the training set, intentionally biased in favour of hate speech, about 1% of tweets were labelled as violent, 34% as offensive (to either individuals or groups), 4% as inappropriate (mostly containing swear words), and the remaining 61% as acceptable. In the evaluation set, which is a random selection of 10.000 Slovenian tweets, only 69 tweets were labelled as violent by at least one annotator, which is about 0.3%.

[2]Some annotators skipped some examples.

Training set		Evaluation set	
Annotated for Vrsta:	99809	**Annotated for Vrsta:**	20000
0 ni sporni govor	60981	0 ni sporni govor	13273
1 nespodobni govor	3817	1 nespodobni govor	285
2 žalitev	34244	2 žalitev	6373
3 nasilje	767	3 nasilje	69
Annotated for Tarča:	34204	**Annotated for Tarča:**	6430
1 ksenofobija in rasizem	1103	1 ksenofobija in rasizem	125
2 begunci/migranti	1011	2 begunci/migranti	68
3 islamofobija	527	3 islamofobija	21
4 antisemitizem	55	4 antisemitizem	10
5 druge religije	172	5 druge religije	15
6 homofobija	304	6 homofobija	16
7 seksizem	773	7 seksizem	68
8 ideologija	6231	8 ideologija	839
9 novinarji in mediji	2517	9 novinarji in mediji	682
10 politika/-i	10924	10 politika/-i	2623
11 posameznik	7016	11 posameznik	1318
12 drugo	3571	12 drugo	645

Figure 1: Number of annotated examples for hate speech type and target. The class distribution is severely unbalanced.

The training dataset for hate speech type includes 34,204 examples and the evaluation dataset includes 6,430 examples. Many of the examples are repeated (by two annotations for the same tweet), yet conflicting (due to annotator disagreement). The training and evaluation sets for hate speech type and target are summarized in Table 1.

The overall annotator agreement for hate speech target on the training set is 63.1%, and Nominal Krippendorf Alpha is 0.537. The annotator agreement for hate speech target on the evaluation set is 62.8%, and Nominal Krippendorf Alpha is 0.503. These scores indicate that the dataset is of high quality compared to other datasets annotated for hate speech, yet the relatively low agreement indicates that the annotation task is difficult and ambiguous even for humans.

3 EXPERIMENTS

We compare different machine learning algorithms on the hate speech target identification task. They belong to one of the following three categories: classical, representation optimization and deep learning. The results are presented in Table 1.

3.1 autoBOT - an autoML for texts

With the increasing amounts of available computing power, *automation* of machine learning has become an active research endeavor. Commonly, this branch of research focuses on automatic model selection and configuration. However, it has recently also been focused on the task of obtaining a suitable representation when less-structured inputs are considered (e.g. texts). This work represents, to our knowledge, one of the first attempts to solve a Slovene-based text classification task with an existing autoML approach. The in-house developed method, called autoBOT [10], has already shown promising results on multiple shared tasks (and in extensive empirical evaluation). Albeit it commonly scores on average worse than large, multi million-parameter neural networks, it remains interpretable and does not need any specialized hardware. Thus, this system serves as an easy-to-obtain baseline which commonly performs better than *ad hoc* approaches such as, e.g. word-based features coupled

with, e.g. a Support Vector Machine (SVM). The tool has multiple configurations which determine the feature space that is being *evolved* during the search for an optimal configuration of both the representation of a given document, but also the most suitable learner. We left all settings to default, varying only the representation type, which was either symbolic, neuro-symbolic-lite, neuro-symbolic-full or neural. Detailed descriptions of these feature spaces are available online[3]. The main difference between these variants is that the neuro-symbolic ones simultaneously consider both symbolic and sub-symbolic feature spaces (e.g. tokens and embeddings of the documents), whilst symbolic or neural-only consider only one type. The neural variant is based on the two non-contextual doc2vec variants and commonly does not perform particularly well on its own.

3.2 Deep Learning

We trained a modelbased on the SloBERTa pre-trained language model [11]. SloBERTa is a transformer-based language model that shares the same architecture and training regime as the Camembert model [7] and is pre-trained on Slovenian corpora. For fine-tuning of the SloBERTa language model, we first split the original training set into training and validation folds in the 90%:10% ratio. We used the suggested hyperparameters for this model. We used the Adam optimizer with the learning rate of $2e-5$ and learning rate warmup over the first 10% of the training instances. We used a weight decay set to 0.01 for regularization. The model was trained for maximum 3 epochs with a batch size of 32. The best model was selected based on the validation set score. We performed the training of the models using the HuggingFace Transformers library [12].

We tokenized the textual input for the neural models with the language model's tokenizer. For performing matrix operations efficiently, all inputs were adjusted to the same length. After tokenizing all inputs, their maximum length was set to 256 tokens. Longer sequences were truncated, while shorter sequences were zero-padded. The fine-tuned model is available at the HuggingFace repository[4].

3.3 Other Baseline Approaches

The two mentioned approaches have demonstrated state-of-the-art performance; however, to establish their performance on this new task, we also implemented the following baselines. First, a simple majority classifier to establish the worst-case performance. Next, a doc2vec-based representation learner was coupled with a linear SVM (doc2vec). The *svm-word* is a sparse TF-IDF representation of the documents coupled with a linear SVM. Similarly, the *svm-char*, however, the representations are based on characters in this variant. The two alternatives use logistic regression (*lr-word*, *lr-char*). As another strong baseline, we used a multilingual language model called MPNet to obtain contextual representations, coupled with an SVM classifier. The baseline doc2vec model was trained for 32 epochs with eight threads. The min_count parameter was set to 2, window size to 5 and vector size to 512. For SVM and logistic regression (*LR*)-based learners, a grid search including the following regularization values was traversed: {0.1, 0.5, 1, 5, 10, 20, 50, 100, 500}.

4 RESULTS

The classification results for the discussed learning algorithms are given in Table 1. The results are sorted by learner complexity. The SloBERTa-based predictor performed the best, however, is also the one which includes the highest number of tunable parameters (more than 100m). The next series of learners are based on autoBOT's evolution and perform reasonably well. Interestingly, autoBOT variants which exploit only symbolic features perform better than the second neural network-based baseline which was not pre-trained specifically for Slovene – the *mpnet*. The remaining baselines perform worse, albeit having a similar number of final parameters to the final autoBOT-based models (tens of thousands at most). The autoBOT-neural, which implements the two main doc2vec variants, performs better than the naïve doc2vec implementation, however not notably better.

To better understand the key properties of the data set which carry information relevant for the addressed predictive task, we additionally explored *autoBOT-symbolic*'s 'report' functionality, which offers insight into the importance of individual feature subspaces. Each subspace and each feature in the subspace has a weight associated with it: the larger the weights, the more relevant a given feature type was for the learner. Visualization of these importances is shown in Table 2. It can be observed that character-based features were the most relevant for this task. This result is in alignment with many previous results on tweet classification, where e.g. punctuation-level features can be surprisingly effective. Furthermore, relational token features were also relevant. This feature type can be understood as skip-grams with dynamic distances between the two tokens. This feature type indicates that short phrases might have been of relevance. Interestingly, keyword-based features were not relevant for the learner. Further, autoBOT, being effectively a fine-tuned linear learner, also offers direct insight into fine-grained performances. Examples for the top five features per type are shown in Table 2.

5 CONCLUSION

In this work we present a new dataset of Slovenian tweets annotated for hate speech targets. To develop effective computational models to solve this task we use two approaches: the autoML approach combining symbolic and neural representations and a contextually-aware language model SloBERTa.

The results show that the context-aware SloBERTa model significantly outperforms all the other trained models. This result, together with the lower inter-annotator scores, confirm our initial assumption that hate speech target identification is a complex semantic task that requires a more complex understanding of the text that goes beyond simple pattern matching. However, the seemingly simpler models may still offer distinct advantages over the more complex neural models. First, the auto-ML models tested in this work are easily interpretable, offering insights into textual features which contribute to the classification. On the other hand, the neural language models generally work as black-boxes, and the extent of their interpretability is still an open research question. Second, the auto-ML models are significantly more straightforward to deploy as they tend to be much less computationally demanding both in terms of RAM and CPU usage. Neural language models are able to solve harder tasks but their increased number of parameters usually makes them a considerable challenge to deploy in a scalable fashion.

ACKNOWLEDGEMENTS

We would like to thank the Slovenian Research Agency for the financing of the second researcher (young researcher grant) and the financial support from research core funding no. P2-103. The

[3] autoBOT feature spaces: https://skblaz.github.io/autobot/features.html
[4] Hate speech target classification model: https://huggingface.co/IMSyPP/hate_speech_targets_slo

Table 1: Overview of the classification results. The SloBERTa model significantly outperforms all the other models and reaches inter-annotator agreement.

Classification model	Accuracy	Macro Rec	Macro Prec	Macro F1
majority	40.79%	8.33%	3.40%	4.83%
doc2vec	43.25%	20.65%	20.67%	19.76%
AutoBOT-neural (9h)	45.79%	15.37%	20.00%	16.10%
svm-word	50.39%	21.40%	25.75%	22.02%
lr-word	50.39%	21.40%	25.75%	22.02%
lr-char	51.21%	25.14%	28.17%	26.10%
svm-char	51.90%	23.47%	27.59%	24.20%
AutoBOT-neurosymbolic-lite (4h)	54.26%	27.34%	35.06%	28.90%
Paraphrase-multilingual-mpnet-base-v2 + Linear SVM	55.40%	40.24%	44.29%	41.20%
AutoBOT-symbolic (9h)	55.99%	29.68%	37.86%	31.32%
AutoBOT-neurosymbolic-full (4h)	56.28%	32.29%	37.83%	33.07%
SloBERTa	63.81%	53.03%	45.63%	48.28%

Table 2: Most relevant features per feature subspace. Feature subspaces are ordered relative to their importance. Individual numeric values next to each feature represent that feature's importance for the final learner. The features are sorted per-type. Note the word_features and their alignment with what a human would associate with hate speech.

char_features	ta s : 3.56	ni d : 2.73	lič : 2.69	ola : 2.58	ne m : 2.5
relational_features_token	pa−3−je : 2.23	pa−2−se : 2.12	v−2−pa : 1.78	ne−1−pa : 1.75	v−2−se : 1.71
pos_features	nnp nn nnp : 1.77	nnp jj nn : 1.75	nnp jj : 1.57	cc : 1.46	nn nn rb : 1.45
word_features	idioti : 1.09	riti : 0.95	tole : 0.95	sem : 0.94	fdv : 0.93
relational_features_char	e−3−d : 1.74	i−3−s : 1.56	n−3−z : 1.48	h−5−v : 1.43	z−4−t : 1.4
topic_features	topic_12 : 0.14	topic_2 : 0.02	topic_0 : 0.0	topic_1 : 0.0	topic_3 : 0.0
keyword_features	007amnesia : 0.0	15sto : 0.0	24kitchen : 0.0	2pira : 0.0	2sto7 : 0.0

work was also supported by European Union's Horizon 2020 research and innovation programme project EMBEDDIA (grant no. 825153) and the European Union's Rights, Equality and Citizenship Programme (2014-2020) project IMSyPP (grant no. 875263).[5]

REFERENCES

[1] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, 759–760.

[2] B. Evkoski, I. Mozetic, N. Ljubesic, and P. Kralj Novak. 2021. Community evolution in retweet networks. *arXiv preprint arXiv:2105.06214*.

[3] B. Evkoski, A. Pelicon, I. Mozetic, N. Ljubesic, and P. Kralj Novak. 2021. Retweet communities reveal the main sources of hate speech. (2021). arXiv: 2105.14898 [cs.SI].

[4] N. Ljubešić, D. Fišer, and T. Erjavec. 2019. The frenk datasets of socially unacceptable discourse in slovene and english. (2019). arXiv: 1906.02045 [cs.CL].

[5] N. Ljubešić, D. Fišer, and T. Erjavec. 2014. TweetCaT: a tool for building Twitter corpora of smaller languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA), Reykjavik, Iceland, (May 2014).

[6] S. Malmasi and M. Zampieri. 2018. Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30, 2, 187–202.

[5] The content of this publication represents the views of the authors only and is their sole responsibility. The European Commission does not accept any responsibility for use that may be made of the information it contains.

[7] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, and B. Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, (July 2020), 7203–7219.

[8] A. Pelicon, R. Shekhar, B. Škrlj, M. Purver, and S. Pollak. 2021. Investigating cross-lingual training for offensive language detection. *PeerJ Computer Science*, 7, e559.

[9] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki. 2017. Measuring the reliability of hate speech annotations: the case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

[10] B. Škrlj, M. Martinc, N. Lavrač, and S. Pollak. 2021. Autobot: evolving neuro-symbolic representations for explainable low resource text classification. *Machine Learning*, 110, 5, 989–1028. ISSN: 1573-0565. DOI: 10.1007/s10994-021-05968-x.

[11] M. Ulčar and M. Robnik-Šikonja. 2021. Slovenian RoBERTa contextual embeddings model: SloBERTa 2.0. Slovenian language resource repository CLARIN.SI. (2021). http://hdl.handle.net/11356/1397.

[12] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771.

[13] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL*.