UR_NLP @ HaSpeeDe 2 at EVALITA 2020: Towards Robust Hate Speech Detection with Contextual Embeddings

Julia Hoffmann
University of Regensburg
Julia1.Hoffmann@ur.de

Udo Kruschwitz
University of Regensburg
Udo.Kruschwitz@ur.de

Abstract

We describe our approach to address Task A of the EVALITA 2020 Hate Speech Detection (HaSpeeDe2) challenge. We submitted two runs that are both based on contextual embeddings – which we had chosen due to their effectiveness in solving a wide range of NLP problems. For our baseline run we use stacked embeddings that serve as features in a linear SVM. Our second run is a simple ensemble approach of three SVMs with majority voting. Both approaches outperform the official baselines by a large margin, and the ensemble classifier in particular demonstrates robust performance on different types of test data coming 6th (out of 27 runs) for news headlines and 10th (out of 27) for Twitter feeds.

1 Introduction

Hate speech in social media (and its automatic detection) has become a major problem in recent years. It can be generically defined as *"language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group"* (Davidson et al., 2017) and is often based on aspects like race, religion, ethnicity, and gender. The problem is that what is considered acceptable for some might not be for others. In addition to that, there is a fine line between freedom of expression on the one hand and censorship and illegal discrimination on the other (Zimmerman et al., 2018). In fact, this fine balance is reflected by the fundamental human rights (as outlined in articles 19 and 20 of (The United Nations, 1948) and (The United Nations General Assembly, 1966) which simultaneously provide rights to freedom of expression and prevent censorship and illegal discrimination. All this contributes to making automatically detecting hate speech a challenging task.

Nevertheless, social media platforms such as Twitter have defined clear guidelines prohibiting the use of hateful behaviour.[1] Accounts with such contents can be reported and are subsequently deleted. The challenge is to be able to detect such content automatically with both high precision and high recall.

The EVALITA evaluation campaign introduced a hate speech detection challenge applied to Italian social media in 2018 (Bosco et al., 2018). Its success led to the continuation of the challenge in 2020, now called HaSpeeDe 2, which is split up into three subtasks (Sanguinetti et al., 2020). This report discusses our two runs that we submitted to HaSpeeDe 2 Task A of EVALITA 2020 (Basile et al., 2020). We will first give some background on the problem aimed at motivating our choice of approach. We will then introduce our systems, report results and discuss some findings. We will also outline some scope for future developments.

2 Background

We will provide some background that should motivate the system architectures we developed. There are several aspects to be mentioned here.

First of all, given the impressive advances in a broad range of natural language processing tasks using a transformer-based architecture (Vaswani et al., 2017) capturing contextual embeddings – most prominently utilizing the various flavours of BERT (Devlin et al., 2019) – we decided to adopt a transformer architecture as well. There are two ways language models such as BERT could be used – using pre-training and fine-tuning or just feature-based without fine-tuning.

[1] https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

This leads us to the next design decision. The winning team in the 2018 HaSpeeDe competition, ItaliaNLP, submitted as one of their runs a SVM with three different feature categories, namely raw and lexical text, morpho-syntactic and lexicon features, which performed extremely well in particular when trained and tested on Twitter data (Cimino et al., 2018). Rather than designing an end-to-end neural architecture that would be fine-tuned on the available training data we therefore opted for a simpler and slightly more transparent architecture with an SVM backbone as our classifier, i.e. the feature-based approach mentioned above.

Ensemble methods have repeatedly been shown to outperform individual classifiers for a variety of tasks including hate speech detection. For example, an ensemble of ten simple neural classifiers proposed by (Zimmerman et al., 2018) outperformed a BERT-based approach on the standard HatebaseTwitter benchmark dataset (MacAvaney et al., 2019). Other recent examples that demonstrate the effectiveness of ensemble methods for hate speech detection include (Alonso et al., 2020; Nourbakhsh et al., 2019; Seganti et al., 2019; Zampieri et al., 2020; Badjatiya et al., 2017; Park and Fung, 2017). We should add that these findings are not limited to the area of hate speech detection as ensemble methods have a long history in being successfully utilized in a broad range of machine learning approaches, e.g. (Molteni et al., 1996). Simple but effective ensemble approaches have also been used for example in sentiment classification of tweets, e.g. (Hagen et al., 2015), and other social media classification tasks.

Finally, given the task definition in which the classifier was to be trained on social media data but then tested on both social media and news headlines we were aiming at an approach that would have a robust performance across domains rather than being tailored specifically to one type of data.

One additional motivation for our work is the intention to develop approaches that can be applied to different languages (we will get back to that point when we outline future directions).

We will now demonstrate how those motivating considerations lead to the system architecture we propose.

3 System Architecture

We submitted two runs of which the first one can be considered our own baseline approach. We first present both architectures at a conceptual level and will go into the technical details when we discuss the experimental setup in the next section. Our runs are:

- Model 1: *Stacked embeddings as features of a linear SVM*
- Model 2: *Ensemble of several SVMs with different text representations* – both contextual embeddings and TF-IDF-based.

Both models can be realised in many different ways. The core idea, as motivated before, is to experiment with transformer-based contextual embeddings but to avoid fine-tuning and instead deploy a traditional, more transparent approach of SVM. The ensemble can consist of a variety of different systems that can be aggregated in many ways. In this paper (and as submitted) we treat each system as equally important and use a simple majority vote.

Stacked embeddings have been shown to be effective in NLP applications, e.g. (Akbik et al., 2018; Akbik et al., 2019). Conceptually there is some similarity to ensemble approaches in that a combination of differently derived embedding models turns out to be more effective than each approach individually.

3.1 Model 1: Stacked embeddings + SVM

Our own baseline model combines two different document embeddings: transformer document and document pool embeddings which are then fed into a linear SVM to train a classifier. We keep the architecture deliberately simple.

There is a wide range of transformer-based language models. One of our motivations was to train a classifier that will generalise beyond a specific domain but also has the potential to generalise beyond a specific language. We therefore opted for XLM-RoBERTa (XLM-R) that has been shown to outperform alternative multilingual models such as mBERT in various NLP tasks (Conneau et al., 2020). XLM-R is based on XLM and RoBERTa. It is trained on data covering 100 languages in a very large (2TB) CommonCrawl. Transformer document embeddings are obtained from (the large version of) XLM-R. In addition to that we use document pool embeddings which consist of word embeddings using Flair (Akbik et al., 2019). The exact experimental choices are described further down.

3.2 Model 2: Ensemble of SVMs

Our second system is an ensemble classifier consisting of three SVMs each trained on a different text representation, namely:

- Transformer document embeddings using XLM-R
- Document pool embeddings
- Straightforward TF-IDF.

The first two of these are exactly the same as we have seen in Model 1 except that they are not stacked but fed into different classifiers. Again we observe that the general setup is kept simple to avoid overfitting for the specific problem at hand thereby allowing more scope for future experiments.

4 Experimental Setup

We applied our systems to *Task A - Hate Speech Detection (Main Task)*.

4.1 Data Sets

Training and test data is briefly described here.

- *Training Data Set*: the training data set consists of 6,839 tweets in total, 2,766 of them classified as hate speech. The corpus has three columns: tweet ID, text and the label (0 = no hate speech, 1 = hate speech). Table 1 summarises these numbers.

Label	Training Data Set
0	4,073
1	2,766
Total	6,839

Table 1: Training Data

- *Test Data Set*: unlike training data which was all Twitter feeds, there were two sets of test data, the first one sampled from Twitter and the second one from news headlines. The Twitter test set has 1,263 entries in total, the news test set 500. The two columns in both sets are the ID and the text of the tweet and news headlines, respectively. The classes 0 and 1 in the Twitter test set include 641 and 622 tweets respectively. In the news headline test set 319 entries have the label 0, 181 the label 1 (see Table 2).

Label	Twitter Test Set	News Test Set
0	641	319
1	622	181
Total	1,263	500

Table 2: Test Data

4.2 Data Preprocessing

In line with our overall aim of simplicity and generalisibility (rather than tuning) we applied a simple pre-processing pipeline that would apply to both Twitter data as well as news headlines. There are only small variations in the different normalization steps as follows.

For any embedding-based processing the text was lower-cased and punctuation was removed so that any input, be it tweet or news headline, would be represented as a string of unpunctuated tokens. For the calculation of our (sparse) TF-IDF representation the text was tokenized and in addition to that stopwords were removed. After that each token was vectorized using TF-IDF. Figure 1 shows an overview of the preprocessing.

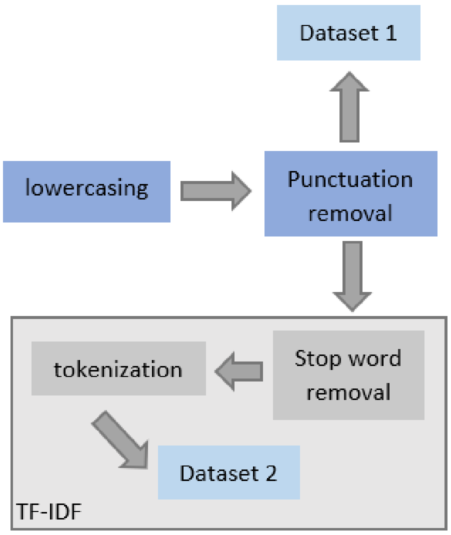


Figure 1: Data Preprocessing

4.3 Implementation

All implementation was done in Python. For all text and document embeddings we used *flairNLP*[2]. Our SVMs were developed using *scikit-learn* (Pedregosa et al., 2011), and for the preprocessing of the TF-IDF version and TF-IDF calculation we used *NLTK*[3] and *scikit-learn*.

Stacked embeddings + SVM: as outlined, we use stacked embeddings composed of *Transformer Document* and *Document Pool Embeddings*. The Transformer Document Embeddings are obtained using XLM-R. Document Pool Embeddings are calculated using a mean-pooling over all word embeddings. It consists of forward and backward embeddings for the Italian language as provided by flair (Akbik et al., 2018) and as recommended. An overview is given in Figure 2.

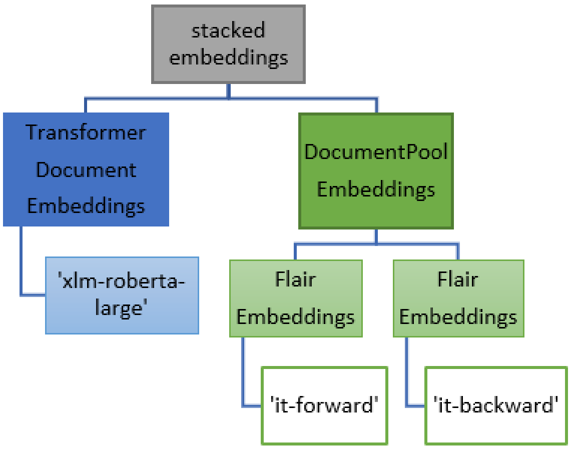


Figure 2: Embeddings in our Baseline (Model 1)

Flair allows for the easy combination of embeddings to create stacked embeddings – one for each input text. These vectors (together with the labels) are then used to train the SVM. Using grid-search on the training data the most suitable parameter settings were determined, and Table 3 specifies the settings which were then used in the submitted run.

Parameter	Value
C	1.0
kernel	'linear'
degree	3
gamma	1

Table 3: Parameters of the SVM (Baseline)

Ensemble of SVMs: three different feature representations are used to train one SVM each as illustrated in Table 4. The first two incorporate the same representations as already seen in Figure 2.

Classifier	Features
SVM2.1	Transformer Document Embeddings
SVM2.2	Document Pool Embeddings
SVM2.3	TF-IDF

Table 4: Overview of SVM Ensemble

Again we used grid-search for parameter tuning (see Table 5).

Parameter	SVM2.1	SVM2.2	SVM2.3
C	1.0	1.0	1.0
kernel	'linear'	'linear'	'rbf'
degree	3	3	3
gamma	1	1	1

Table 5: Parameters of the SVMs for Model 2 (Ensemble of SVMs)

Input is run against each classifier, and through majority voting over these three predictions the final classification category is determined.

5 Results

We first present detailed results and then discuss our findings and insights. We start with our baseline approach and then move on to the classifier ensemble. Macro-F1 is the official metric for this competition. In addition to that we look at Precision, Recall and F1 at category-level and also include confusion matrices for each approach (Model 1 and Model 2) and test set (Twitter data and news headlines). There were 27 runs submitted for each dataset and the official baseline was a linear SVM with TF-IDF of word and char-grams.

5.1 Model 1: Our Baseline

Twitter Data: Training and testing on Twitter data results in a Macro-F1 score of 0.7399 which makes it into position 16 (out of 27). The official task baseline is 0.7212. Details are displayed in Table 6 and Figure 3.

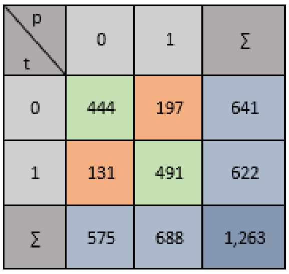
News Headlines: On the news headlines test data we get a Macro-F1 of 0.6684 with official baseline result of 0.6210 (rank 12). More details are in Table 7 and Figure 4.

[2] https://github.com/flairNLP/flair

[3] https://www.nltk.org

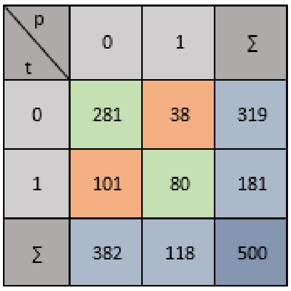
Metric	0	1
Precision	0.7722	0.7137
Recall	0.6927	0.7894
F1	0.7303	0.7496

Table 6: Results: Model 1 (Stacked embeddings + SVM) on Twitter Data



p t	0	1	Σ
0	444	197	641
1	131	491	622
Σ	575	688	1,263

Figure 3: Confusion Matrix: Model 1 (Stacked embeddings + SVM) on Twitter Data (p = predicted, t = true)

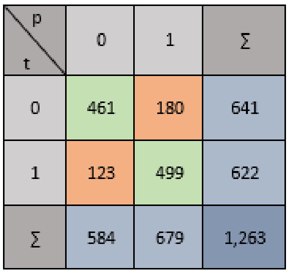


p t	0	1	Σ
0	281	38	319
1	101	80	181
Σ	382	118	500

Figure 4: Confusion Matrix: Model 1 (Stacked embeddings + SVM) on News Data (p = predicted, t = true)

5.2 Model 2: Ensemble

Twitter Data: Our ensemble approach gets a Macro-F1 of 0.7599 (rank 10). More details are included in Table 8 and Figure 5.



p t	0	1	Σ
0	461	180	641
1	123	499	622
Σ	584	679	1,263

Figure 5: Confusion Matrix: Model 2 (Ensemble of SVMs) on Twitter Data (p = predicted, t = true)

Metric	0	1
Precision	0.7356	0.6780
Recall	0.8809	0.4420
F1	0.8017	0.5351

Table 7: Results: Model 1 (Stacked embeddings + SVM) on News Data

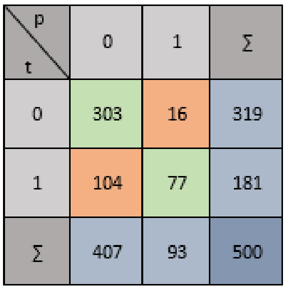
Metric	0	1
Precision	0.7894	0.7349
Recall	0.7192	0.8023
F1	0.7527	0.7671

Table 8: Results: Model 2 (Ensemble of SVMs) on Twitter Data

News Headlines: On the news headlines test data we get a Macro-F1 of 0.6984 with an official baseline result of 0.6210 (rank 6). More details can be found in Table 9 and Figure 6.

Metric	0	1
Precision	0.7445	0.8280
Recall	0.9498	0.4254
F1	0.8347	0.5620

Table 9: Results: Model 2 (Ensemble of SVMs) on News Data



p t	0	1	Σ
0	303	16	319
1	104	77	181
Σ	407	93	500

Figure 6: Confusion Matrix: Model for 2 (Ensemble of SVMs) on News Data (p = predicted, t = true)

6 Discussion

Our first observation we derive from the results is that the ensemble approach we proposed for this task does provide a robust and solid performance – solid in that it scores well in the ranked list of systems and robust in that it also ranks highly when applied to out-of-domain data (coming 6th out of 27 submitted runs on data it had not been trained on). Given the simplicity of our system architecture and the composition of the official baseline system we also note the superiority of transformer-based contextual embeddings over bag-of-words approaches (while this comes as no surprise it is still worth pointing out). Moving from a feature-based to a pre-training plus fine-tuning approach will most certainly further push up the scores.

Looking at the balance between precision and recall, we find that both our approaches have a tendency to return a fair number of *false positives for the Twitter data* set. This could indicate that words and phrases used to express hateful content is quite common in social media even if it does not actually represent hate speech. On the other hand, we record a large proportion of *false negatives* when classifying *news headlines*. This could be an indicator of a more subtle way in which hate speech is expressed in traditional news outlets.

Generally speaking, both models perform better on Twitter data than on news headlines – again an insight that was to be expected due to the training data. However, the fact that our approach managed to score higher in the ranked list of systems for data it was not trained on is a result that confirms our initial assumptions – that using a corpus with a very broad range of topics, styles and languages as our core language model would help in making the system transfer more easily to unseen input.

This leads us to an area of future research. While it would be possible to improve the performance of our system by making the preprocessing, the language model and any fine-tuning step match more closely the expected test data – e.g. by using AlBERTo, a BERT-based transformer trained on Italian Twitter data (Polignano et al., 2019) – we are actually aiming at something else. As part of the COURAGE research project[4] we are exploring ways to help teenagers manage social media exposure by providing a virtual companion that would, among other things, automatically identify examples of hate speech, bullying or other toxic content. Given this is a multi-national effort we are interested in architectures that work for languages including Italian, Spanish, German and English with as little fine-tuning as possible. The ensemble introduced here with its multilingual transformer backbone turns out to be a step in that direction.

7 Conclusion

We presented a simple but effective architecture to detect hate speech in Italian social media and news headlines. Our ensemble-based architecture relies on contextual embeddings trained on a large multilingual corpus which we see as the basis for the robustness of the approach. There is plenty of room for further improvement and the results we report here will serve as a benchmark in this development.

Acknowledgements

This work was supported by the project *COURAGE: A Social Media Companion Safeguarding and Educating Students* funded by the Volkswagen Foundation, grant number 95564.

References

A. Akbik, D. Blythe, and R. Vollgraf. 2018. Contextual string embeddings for sequence labeling. In E. M. Bender, L. Derczynski, and P. Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1638–1649. Association for Computational Linguistics.

A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Demonstrations*, pages 54–59, Minneapolis, Minnesota, June. Association for Computational Linguistics.

P. Alonso, R. Saini, and G. Kovács. 2020. Hate Speech Detection Using Transformer Ensembles on the HASOC Dataset. In A. Karpov and R. Potapova, editors, *Speech and Computer*, pages 13–21, Cham. Springer International Publishing.

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*, pages 759–760. International World Wide Web Conferences Steering Committee.

V. Basile, D. Croce, M. Di Maro, and L. C. Passaro. 2020. EVALITA 2020: Overview of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

[4] https://www.upf.edu/web/courage

C. Bosco, F. Dell'Orletta, F. Poletto, M. Sanguinetti, and M. Tesconi. 2018. Overview of the EVALITA 2018 hate speech detection task. In T. Caselli, N. Novielli, V. Patti, and P. Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org.

A. Cimino, L. De Mattei, and F. Dell'Orletta. 2018. Multi-task learning in deep neural networks at EVALITA 2018. In T. Caselli, N. Novielli, V. Patti, and P. Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org.

A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, editors, *Proceedings of ACL*, pages 8440–8451. Association for Computational Linguistics.

T. Davidson, D. Warmsley, M. W. Macy, and I. Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*, pages 512–515. AAAI Press.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

M. Hagen, M. Potthast, M. Büchner, and B. Stein. 2015. Webis: An ensemble for twitter sentiment detection. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 582–589.

S. MacAvaney, H. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder. 2019. Hate speech detection: Challenges and solutions. *PLoS ONE*, 14:1–16.

F. Molteni, R. Buizza, T. N Palmer, and T. Petroliagis. 1996. The ECMWF ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, 122(529):73–119.

A. Nourbakhsh, F. Vermeer, G. Wiltvank, and R. van der Goot. 2019. sthruggle at SemEval-2019 task 5: An ensemble approach to hate speech detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 484–488, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

J. H. Park and P. Fung. 2017. One-step and two-step classification for abusive language detection on twitter. In *Proceedings of The First Workshop on Abusive Language Online*, pages 41–45. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, and V. Basile. 2019. Alberto: Italian BERT language understanding model for NLP challenging tasks based on tweets. In R. Bernardi, R. Navigli, and G. Semeraro, editors, *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*, volume 2481 of *CEUR Workshop Proceedings*. CEUR-WS.org.

M. Sanguinetti, G. Comandini, E. Di Nuovo, S. Frenda, M. Stranisci, C. Bosco, T. Caselli, V. Patti, and I. Russo. 2020. HaSpeeDe 2@EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

A. Seganti, H. Sobol, I. Orlova, H. Kim, J. Staniszewski, T. Krumholc, and K. Koziel. 2019. NLPR@SRPOL at SemEval-2019 task 6 and task 5: Linguistically enhanced deep learning offensive sentence classifier. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 712–721, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

The United Nations General Assembly. 1966. International covenant on civil and political rights. *Treaty Series*, 999:171, December.

The United Nations. 1948. *Universal Declaration of Human Rights*. The United Nations, December.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, USA. Curran Associates Inc.

M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, and Ç. Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). *CoRR*, abs/2006.07235.

S. Zimmerman, U. Kruschwitz, and C. Fox. 2018. Improving Hate Speech Detection with Deep Learning Ensembles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).