A Dutch Dataset for Cross-lingual Multi-label Toxicity Detection

Ben Burtenshaw and **Mike Kestemont**
Antwerp Centre for Digital Humanities and Literary Criticism
University of Antwerp
Prinsstraat 13, 2000
Antwerp (Belgium)
firstname.lastname@uantwerpen.be

Abstract

Multi-label toxicity detection is highly prominent, with many research groups, companies, and individuals engaging with it through shared tasks and dedicated venues. This paper describes a cross-lingual approach to annotating multi-label text classification on a newly developed Dutch language dataset, using a model trained on English data. We present an ensemble model of one Transformer model and an LSTM using Multilingual embeddings. The combination of multilingual embeddings and the Transformer model improves performance in a cross-lingual setting.

1 Introduction

Toxic comment detection is becoming an integral part of online discussion, and most major social media platforms use it. However, that success is not shared equally across languages. Low resource languages still lack the accurate pre-trained models that are readily available in more resourced languages, such as English. This is mostly due to a lack of annotated corpora. Inconsistent task definitions of task compound the problem. Where quality data does exist, it often uses alternative task definitions. This paper aims to overcome that challenge by annotating a new dataset and evaluating it within a cross-lingual experiment. We perform multi-label text classification, using an ensemble approach of Transformer and LSTM models with multilingual embeddings (Vaswani et al., 2017; Devlin et al., 2019; Van Hee et al., 2015a). The system is trained on English data by Wulczyn et al. and evaluated on newly annotated Dutch text from the Amica corpus (Wulczyn et al., 2017a; Van Hee et al., 2015a).

We selected multi-label toxicity over other label definitions based on its adaptability and feedback from annotators. Toxicity draws its origins from chemistry, referring to how a substance can damage an organism. From experience in annotator training and feedback, this is a straightforward term to communicate to annotators who relate quickly to the concept of harmful language that degrades a conversation or debate, much like a poison.

2 Related Research

The Conversation AI group defined multi-label toxicity, and Wulczyn et al.(Wulczyn et al., 2017c). The term goes beyond its counterparts by adding fine-grained sub-labels. The original motivation of Wulczyn et al. was for multi-label toxicity to serve as a compatible annotation model for tasks beyond the original Wikipedia dataset. Unlike other similar initiatives, their work focused on the risk that communities break down or turn silent, "leading many communities to limit or completely shut down user comments" (Wulczyn et al., 2017a,c). For a detailed overview of multi-label toxicity, look to van Aken et al., or Gunasekera et al. (Georgakopoulos et al., 2018; Wulczyn et al., 2017b).

A current challenge within the sub-field of toxicity detection is the definition and operationalisation as a concrete task. Though there is research within the area, many projects take up alternative interpretations and definitions. This has led to grey areas between terms like offensive language and profanity, cyberbullying, and online harassment. In practice, many projects are classifying the same data and phenomena under alternative definitions. This problem is explored in greater detail by Emmery and colleagues (Emmery et al., 2019).

Cross-lingual classification uses training material in one language and test material in another. In this paper, we use English language training data to improve performance on Dutch language test data. This resourceful combination relies on recent advancements in multilingual models and benefits underrepresented languages greatly. Data

Proceedings of 14th Workshop on Building and Using Comparable Corpora in conjunction with RANLP-2021, pages 75–79, held online, 6 September 2021.
https://doi.org/10.26615/978-954-452-076-2_010

Negative	94.04	Blackmail	0.11
insult	1.96	Racism	0.1
Harmless_sexual	0.97	Att_relatives	0.09
Curse_Exclusion	0.65	Powerless	0.06
Assertive_selfdef	0.54	Other	0.04
Other_language	0.4	Sarcasm	0.04
Sexual_harassment	0.33	Good	0.01
General_defense	0.33	pro_harasser	0.01
Defamation	0.18	Sexism	0.13

Table 1: Cyberbullying Labels within Amica Dataset and Frequency

sets like that of Conversation AI are less available for Dutch, making classification harder. There are a series of recent projects utilising multilingual pretrained models for cross-lingual classification of toxic comments (Pamungkas and Patti, 2019; Pant and Dadu, 2020; Stappen et al., 2020).

Amica was a collaborative project between Dutch-speaking NLP research groups into cyberbullying. Van Hee et al. facilitated the detailed annotation of many data sets for a range of bullying labels, using real and simulated conversations between children. Table 1 gives the label distribution.

3 Data

We use a newly annotated version of the AMiCA dataset, initially developed by Van Hee et al., for cyberbullying tasks. In addition, we performed further annotation for multi-label toxicity, following the label guidelines of Wulczyn et al..

3.1 AMiCA Instant Messages

Van Hee et al. developed the AMiCA dataset through anonymous donation and simulation outlined by Emmery et al.. Table 2 reveals the macro details of the data used with original cyberbullying token labels.

Bullying Tokens	2,343
Negative Tokens	2,546
All Tokens	62,340
Mean Tokens per msg	12

Table 2: AMiCA data lexical statistics

3.2 Multi-label Toxicity Annotation

To annotate the AMiCA dataset for Multi-label toxicity labels, we used the annotation instructions outlined in (Wulczyn et al., 2017c). We translated the instructions into Dutch, the native language of the annotators, and gave detailed guidance with an introductory tutorial and handout. Table 3 describes the sub-labels: Toxicity, Severe Toxicity, Identity Attack, Insult, Profanity, and Threat.

TOXICITY Rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion.
SEVERE_TOXICITY A very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion.
IDENTITY_ATTACK Negative or hateful comments targeting someone because of their identity.
INSULT Insulting, inflammatory, or negative comment towards a person or a group of people.
PROFANITY Swear words, curse words, or other obscene or profane languages.
THREAT Describes an intention to inflict pain, injury, or violence against an individual or group.

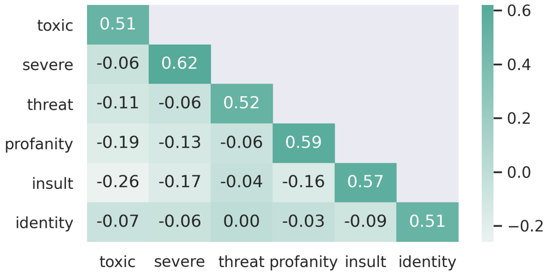
Table 3: Description and Example of labels from the Wikipedia Talk Labels: Toxicity Dataset

We stored the annotated data in a SQL table using the row index of the original AMiCA annotations for cyberbullying. Table 4 shows the distribution of labels across the English data by Wulczyn et al. and the newly annotated data.

Interannotator Agreement We calculated interannotator agreement using the largest set of overlapping instances by the same two annotators achieving a **Krippendorf score of 0.4483**, revealing that there was substantial agreement between annotators. We can compare this to that of Wulczyn et al., which scored 0.45 (Wulczyn et al., 2017a). We can delve further into inter-annotator relations through multi-label use. Figure 1 reveals the Cohen Kappa between labels. We see that all six true label pairs (i.e. TOXIC & TOXIC) achieve a fair to substantial correlation and that all false label pairs (i.e. INSULT & THREAT) do not correlate.

	New		Wulczyn 2017	
	n	%	n	%
toxic	3157	31%	15294	44%
severe	833	8%	1596	5%
threat	851	8%	8449	24%
profanity	1165	11%	478	1%
insult	1276	13%	7877	22%
identity	1339	13%	1405	04%
Total	10189		35099	

Table 4: Annotated Labels in Dutch (New) and English (Wulczyn 2017) data. n shows the number of comments for each label and % shows the percentage of the total comments for that label.



	toxic	severe	threat	profanity	insult	identity
toxic	0.51					
severe	-0.06	0.62				
threat	-0.11	-0.06	0.52			
profanity	-0.19	-0.13	-0.06	0.59		
insult	-0.26	-0.17	-0.04	-0.16	0.57	
identity	-0.07	-0.06	0.00	-0.03	-0.09	0.51

Figure 1: Correlation Matrix of Toxic labels on Annotated Amica Dataset

Compare Toxicity and Cyberbullying As a precursor to the main experiments, and to align the new annotation with Van Hee et al., we tested how cyberbullying acts as a naive predictor of toxicity using the combined labels for each class and F1 Score (Van Hee et al., 2015b; Emmery et al., 2019). We calculated an F1 score of 0.51, revealing that multi-label toxicity does not align with cyberbullying.

4 Method

We performed cross-lingual classification using an Ensemble approach of two component models, a fine-tuned multilingual BERT-base and an LSTM model using Multilingual Unsupervised and Supervised Embeddings (MUSE) (Conneau et al., 2017; Lample et al., 2017). We also used two baseline models for comparison, an LSTM without multilingual embeddings and a Support Vector Machine.

4.1 Fine-tuned BERT-base

We fine-tuned a Multilingual BERT-base model and 3 linear layers. A Bidirectional Encoder Representation from Transformers or BERT model is a pre-trained model that uses bidirectional training to learn contextual attention at a word and sub-word level (Devlin et al., 2019). We used sub-word token representation that aligns with the base vocabulary representation (Zhang et al., 2020). We fine-tuned the BERT model for 4 epochs over a 10-fold cross-validated dataset. The mean validation and training loss for all folds of the data was 0.05.

4.2 LSTM and MUSE Embeddings

We trained a Long Short-term Memory (LSTM) network with Multilingual Universal Sentence Embeddings (MUSE) (Hochreiter and Schmidhuber, 1997; Conneau et al., 2017; Lample et al., 2017). We train the LSTM model for 12 epochs over a 10-fold cross-validated dataset. The mean validation and training loss for all splits of the data was 0.03.

4.3 Ensemble

We used a Random Forest ensemble of the LSTM and BERT models on a cross-validated training set with grid-searched parameters (Breiman, 2001; Nowak et al., 2017). A key risk in ensemble training is overfitting (Pourtaheri and Zahiri, 2016), to mitigate this all models have used a stratified k-fold structure (Yadav and Shukla, 2016).

4.4 Training and Fine-tuning

We used a stratified k-fold configuration of the English and Dutch data to train and fine-tune models. First, we trained and fine-tuned models on a ‘train‘ portion and collected the predicted labels on ‘test‘ portions of the folds, split for English and Dutch data. This allowed us to reveal language performance separately. Next, we trained the ensemble model on component model predictions. Finally, we used an exhaustive grid search to select hyperparameters (Bergstra and Bengio, 2012) and a Receiver Under the Curve analysis (ROC) to select decision thresholds from the component models (Fawcett, 2006).

5 Results

Table 5 reveals results for *baselines*, component models, and ensemble model. We express results as Area Under the Curve, mean Precision, mean Recall, mean F1 for all labels. Baseline models are a Support Vector Machine of Continuous Bag-of-Words representations and an LSTM without Multilingual Universal Sentence Embeddings. Both component models achieved relevant F1 scores for the multi-label classification of toxicity, and the ensemble approach achieved the highest score. We also find that component models were able to overcome the low precision score seen in baseline methods.

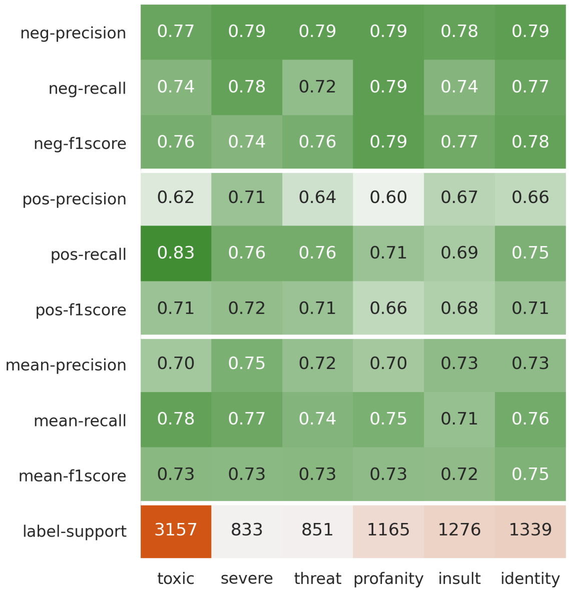
	AUC	Pre	Rec	F1
Ensemble	0.9401	0.7023	0.8789	0.7323
BERT	0.9113	0.6745	0.8412	0.7017
MUSE	0.8552	0.6301	0.7838	0.6512
LSTM w/o MUSE	0.7519	0.5692	0.7021	0.5845
SVM & CBOW	0.5702	0.4239	0.5217	0.4419

Table 5: Results Table of *baselines*, component, and ensemble models. Results are expressed as AUC, mean Precision, mean Recall, mean F1 for all labels.

6 Analysis

We performed error analysis to interpret model performance in relation to labels and the language of comments.

Sub-label Performance Figure 2 reveals the Precision, Recall, and F1 Score of the Ensemble model on all labels. Furthermore, we can see that the model performs better at negative label prediction, a common trait in transformer model classification.



	toxic	severe	threat	profanity	insult	identity
neg-precision	0.77	0.79	0.79	0.79	0.78	0.79
neg-recall	0.74	0.78	0.72	0.79	0.74	0.77
neg-f1score	0.76	0.74	0.76	0.79	0.77	0.78
pos-precision	0.62	0.71	0.64	0.60	0.67	0.66
pos-recall	0.83	0.76	0.76	0.71	0.69	0.75
pos-f1score	0.71	0.72	0.71	0.66	0.68	0.71
mean-precision	0.70	0.75	0.72	0.70	0.73	0.73
mean-recall	0.78	0.77	0.74	0.75	0.71	0.76
mean-f1score	0.73	0.73	0.73	0.73	0.72	0.75
label-support	3157	833	851	1165	1276	1339

Figure 2: Classification Report from Ensemble Approach on all toxicity labels

Cross-lingual Performance We explored the models' cross-lingual performance by comparing their scores on the English and Dutch data, shown in Table 6. Logically, the LSTM with MUSE embeddings performs poorly on English data, without relevant embedding weights. On the other hand, the BERT model performs well in both languages, and the Ensemble model relies on that when classifying English Data.

	All	EN	NL
Ensemble	0.6401	0.7587	0.7323
BERT	0.7112	0.7213	0.7017
MUSE	0.4812	0.4512	0.6512

Table 6: Cross-lingual Performance: F1 Scores of underlinecomponent and ensemble models. **EN** are scores on the Wulczyn data, **NL** are score on the new Dutch data.

7 Summary

We have demonstrated that by using multilingual pre-trained language models within an ensemble approach, we can classify multi-label toxicity in an alternate language. Furthermore, we have demonstrated that the BERT model's underlying training affects target language performance by analysing the performance of baseline, component and ensemble models in cross-lingual features. Furthermore, Table 5 reveals that component models were able to overcome an excess of false positives that hindered baseline methods.

References

James Bergstra and Yoshua Bengio. 2012. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research* 13 (2012), 25.

Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (Oct. 2001), 5–32. https://doi.org/10.1023/A:1010933404324

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word Translation without Parallel Data. *arXiv preprint arXiv:1710.04087* (2017). arXiv:1710.04087

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]* (May 2019). arXiv:1810.04805 [cs]

Chris Emmery, Ben Verhoeven, Guy De Pauw, Gilles Jacobs, Cynthia Van Hee, Els Lefever, Bart Desmet, Véronique Hoste, and Walter Daelemans. 2019. Current Limitations in Cyberbullying Detection: On Evaluation Criteria, Reproducibility, and Data Scarcity. *arXiv:1910.11922 [cs]* (Oct. 2019). arXiv:1910.11922 [cs]

Tom Fawcett. 2006. An Introduction to ROC Analysis. *Pattern recognition letters* 27, 8 (2006), 861–874.

Spiros V. Georgakopoulos, Sotiris K. Tasoulis, Aristidis G. Vrahatis, and Vassilis P. Plagianakos. 2018. Convolutional Neural Networks for Toxic Comment Classification. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*. 1–6.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised Machine Translation Using Monolingual Corpora Only. *arXiv preprint arXiv:1711.00043* (2017). arXiv:1711.00043

Jakub Nowak, Ahmet Taspinar, and Rafał Scherer. 2017. LSTM Recurrent Neural Networks for Short Text and Sentiment Classification. In *Artificial Intelligence and Soft Computing (Lecture Notes in Computer Science)*, Leszek Rutkowski, Marcin Korytkowski, Rafał Scherer, Ryszard Tadeusiewicz, Lotfi A. Zadeh, and Jacek M. Zurada (Eds.). Springer International Publishing, Cham, 553–562. https://doi.org/10.1007/978-3-319-59060-8_50

Endang Wahyu Pamungkas and Viviana Patti. 2019. Cross-Domain and Cross-Lingual Abusive Language Detection: A Hybrid Approach with Deep Learning and a Multilingual Lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics, Florence, Italy, 363–370. https://doi.org/10.18653/v1/P19-2051

Kartikey Pant and Tanvi Dadu. 2020. Cross-Lingual Inductive Transfer to Detect Offensive Language. *arXiv:2007.03771 [cs]* (July 2020). arXiv:2007.03771 [cs]

Zeinab Khatoun Pourtaheri and Seyed Hamid Zahiri. 2016. Ensemble Classifiers with Improved Overfitting. In *2016 1st Conference on Swarm Intelligence and Evolutionary Computation (CSIEC)*. 93–97. https://doi.org/10.1109/CSIEC.2016.7482130

Lukas Stappen, Fabian Brunn, and Björn Schuller. 2020. Cross-Lingual Zero- and Few-Shot Hate Speech Detection Utilising Frozen Transformer Language Models and AXEL. *arXiv:2004.13850 [cs, stat]* (April 2020). arXiv:2004.13850 [cs, stat]

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2015a. *Automatic Detection and Prevention of Cyberbullying*.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste. 2015b. Detection and Fine-Grained Classification of Cyberbullying Events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*. INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria, 672–680.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv:1706.03762 [cs]* (Dec. 2017). arXiv:1706.03762 [cs]

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017a. Ex Machina: Personal Attacks Seen at Scale. *arXiv:1610.08914 [cs]* (Feb. 2017). arXiv:1610.08914 [cs]

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017b. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web*. 1391–1399.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017c. Wikipedia Talk Labels: Personal Attacks. https://doi.org/10.6084/M9.FIGSHARE.4054689

S. Yadav and S. Shukla. 2016. Analysis of K-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification. In *2016 IEEE 6th International Conference on Advanced Computing (IACC)*. 78–83. https://doi.org/10.1109/IACC.2016.25

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-Aware BERT for Language Understanding. *arXiv:1909.02209 [cs]* (Feb. 2020). arXiv:1909.02209 [cs]