An Ensemble approach to Cross-lingual Multilabel Toxicity Detection on Dutch Conversations

Anonymous ACL Rolling Review submission

Abstract

The task of multi-label toxicity detection is currently 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 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 (Vaswani et al., 2017; Devlin et al., 2018). The combination of multilingual embeddings and Transformer model improves performance in a cross-lingual setting. We analyse the specific challenges of cross-lingual tasks within lesser resourced language than English, namely; annotation, comparison, training distribution, and multi-lingual embeddings.

1 Introduction

Toxic comment detection is becoming an integral part of online discussion, and has now been practically implemented within most major social media platforms. However, that success is not shared equally across languages. Lesser resourced languages than English still lack the accurate pre-trained models that are readily available in more resourced languages. This is brought about in most parts by a lack of large annotated corpora, but is also exacerbated by inconsistent definitions of task across the sub-fields of offensive language, abusive language, cyberbullying, hate speech, and toxicity. Where quality data does exist, it is not always produced under comparable circumstances. This paper aims to overcome that challenge by leveraging recent advancements in multi-lingual language representations. We perform multi-label text classification on the Amica data set, using an ensemble approach of Transformer and LSTM models with multilingual embeddings (Vaswani et al., 2017; Devlin et al., 2018; Van Hee et al., 2018). The system is trained on English text and evaluated on Dutch text.

2 Related Research

Multilabel toxicity detection is an active area of research within Natural Language Processing, where numerous research groups and venues are currently engaging with it. Toxicity is seen as a subsection or interpretation of offensive language, and has been most notably tackled through a number of shared tasks. A current challenge within the sub-field of toxicity detection is the definition and operationalization of it as a concrete task. Though ample research is being done within the area, many projects take up alternative interpretations and definitions. This has lead to grey areas between terms like offensive language and profanity, or 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). Here we can give a brief overview of surrounding terms and highlight key studies addressing them.

2.1 Overlapping Terms

Toxicity The most general of these definition draws it’s origins from chemistry, where toxicity refers the extent to which 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. The word is less prevalent within day-to-day speech than ‘offensive’, but it raises less subjective questions (Georgakopoulos et al., 2018; Wulczyn et al., 2017a) .
Offensive Language In some definitions ‘Offensive Language’ is synonymous with profanity, though this is not always logical; for example, using profanity as a modifier to express a complement ‘You are a F’ing genius’. Also when subjective experience is considered, offensive language does not necessarily include all language that offends, which adds to possible confusion for annotators (Chen et al., 2012; Xiang et al., 2012).

Hate Speech Hate speech detection is the most prevalent within the field of NLP. Its distinguishing feature would be a focus on data that targets particular groups based on race, religion, or gender, though other definitions by no means exclude this. (Badjatiya et al., 2017; Burnap and Williams, 2016; Davidson et al., 2017; Del Vigna et al., 2017; Gambäck and Sikdar, 2017; Gitari et al., 2015; Warner and Hirschberg, 2012)

Abusive Language Abusive language would focus on insulting and demeaning language that is targeted at a specific individual or group (Mehdad and Tetreault, 2016; Park and Fung, 2017)

Cyberbullying Researcher’s targeting cyberbullying work typically with children’s language. Emmery et al. (Emmery et al., 2019) explore the relationship between bullying and toxicity (Dadvar et al., 2013; Dinakar et al., 2012; Van Hee et al., 2015; Zhong et al., 2016).

Online Harassment Online Harassment mimics cyberbullying in definition, though researchers will often work with adult’s data (Golbeck et al., 2017; Yin et al., 2009). Online Harassment also encompasses some behavioural research.

2.2 Multi-label Toxicity
Multilabel toxicity was defined by the Conversation AI group and Wulczyn et al.(Wulczyn et al., 2017c). The term goes beyond its counterparts by adding fine grained sub labels, into which other sub tasks and labels could be assigned. The original motivation of the Wulczyn et al. was for multi-label toxicity to serve as a compatible annotation model for tasks beyond the original Wikipedia dataset. For a detailed overview of this discussion look to van Aken et al. or Gunasekera et al. (van Aken et al., 2018; Gunasekara and Nejadgholi, 2018).

2.3 Cross-lingual Language Classification
Beyond the primary task, this paper also performs cross-lingual classification of Dutch language test data from English language training data. This resourceful combination relies on recent advancements in multilingual models, and benefits under represented languages greatly. Data sets like that of Conversation AI are less available for Dutch making classification far 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).

2.4 Amica Project
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.

### Table 1: Annotation Labels within Amica Dataset

<table>
<thead>
<tr>
<th>Annotation Label</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>94.04</td>
</tr>
<tr>
<td>General_insult</td>
<td>1.96</td>
</tr>
<tr>
<td>Harmless_sexual_talk</td>
<td>0.97</td>
</tr>
<tr>
<td>Curse_Exclusion</td>
<td>0.65</td>
</tr>
<tr>
<td>Assertive_selfdef</td>
<td>0.54</td>
</tr>
<tr>
<td>Other_language</td>
<td>0.4</td>
</tr>
<tr>
<td>Sexual_harassment</td>
<td>0.33</td>
</tr>
<tr>
<td>General_defense</td>
<td>0.33</td>
</tr>
<tr>
<td>Defamation</td>
<td>0.18</td>
</tr>
</tbody>
</table>

Annotation labels and percentage label frequency

3 Data

3.1 Define Multi-label Toxicity
Multi-label toxicity is a six-part label used by the Wikipedia Talk Labels: Toxicity Data set (Wulczyn et al., 2017c). Table 2 describes those sub-labels:

<table>
<thead>
<tr>
<th>Toxicity</th>
<th>Severe Toxicity</th>
<th>Identity Attack</th>
<th>Insult</th>
<th>Profanity</th>
<th>Threat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toxicity</td>
<td>0.13</td>
<td>0.01</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Severe Toxicity</td>
<td>0.05</td>
<td>0.01</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Identity Attack</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Insult</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Profanity</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Threat</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Multi-label Toxicity Dataset [wiki-tox] was developed by researchers at Conversation AI (Wulczyn et al., 2017c), to be used in public shared tasks and to address the challenge of offensive conversations on online platforms. 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., 2017b,c). Their objective was to encourage active debate free from toxicity or prejudice.
<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOXICITY</td>
<td>Rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion.</td>
<td>“Bye! Don’t look, come or think of coming back!”</td>
</tr>
<tr>
<td>SEVERE_TOXICITY</td>
<td>A very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective.</td>
<td>“You’re a piece of scum.”</td>
</tr>
<tr>
<td>IDENTITY_ATTACK</td>
<td>Negative or hateful comments targeting someone because of their identity.</td>
<td>“Mate, sound like you are Jewish. Gayness is in the air”</td>
</tr>
<tr>
<td>INSULT</td>
<td>Insulting, inflammatory, or negative comment towards a person or a group of people.</td>
<td>“Do you know you come across as a giant prick?”</td>
</tr>
<tr>
<td>PROFANITY</td>
<td>Swear words, curse words, or other obscene or profane language.</td>
<td>“That guideline is bullshit and should be ignored.”</td>
</tr>
<tr>
<td>THREAT</td>
<td>Describes an intention to inflict pain, injury, or violence against an individual or group.</td>
<td>“I will arrange to have your life terminated.”</td>
</tr>
</tbody>
</table>

Table 2: Description and Example of labels from the Wikipedia Talk Labels: Toxicity Dataset

Data Collection of the wiki-tox dataset was intended to build an overview of attack conversations on wikipedia, as part of a collection of datasets focused on personal attacks and demographic abuse. The developers organised a Kaggle shared task, for which these datasets were amalgamated and focused on the task of toxicity (Wulczyn et al., 2017c). Messages were limited to comments only and fields relating to user source, message type, and context were removed though still aligned via a foreign key. The final dataset contained comment text and a binary label for six toxicity classes shown in Table 2.

Annotators were recruited by Conversation Ai from the Figshare platform. The same group of annotators were employed for all subtasks. Their data achieved a Cohen Kappa score of 0.45 which shows that their data had a ‘fair’ alignment, using Hayes and Krippendorff’s interpretation (Hayes and Krippendorff, 2007; Wulczyn et al., 2017b).

Comment Quality: Comments in the wiki-tox dataset are based around the task of editing Wikipedia pages. The dataset contains mainly active users, with 89% of users contributing more than 5 comments (Wulczyn et al., 2017b). Attacking or toxic comments are distributed across a broad range of users; almost 80% of toxic comments come from 9000 users, implying that toxicity is distributed generally throughout the corpus. Furthermore, Wulczyn studied Wikipedia article versus user tags as a predictor of toxicity, and found that article is a stronger predictor of toxicity (Wulczyn et al., 2017c).

![Figure 1: Correlation Matrix of Wiki Toxic Comment Data](image)
A correlation matrix between labels, higher correlations shown in darker colours based on the right-hand key.

3.3 Amica Cyberbullying Data
The Amica project is described in greater detail in the Related Research section 2.4. Van Hee et al. focused on the detection of cyberbullying and its place within other offensive language classification tasks. They developed datasets through multiple methods; anonymous donation, simulation and parsing from the web. These channels come with various ethical restrictions, which Emmery et al. outline (Emmery et al., 2019). We have limited this study to the simulated portion of the dataset.

3.3.1 Annotation of Multilabel Toxicity
In line with the wiki-tox dataset we annotated the Amica corpus using the same instructions outlined in the authors own work (Wulczyn et al., 2017c). We translated the instructions into Dutch, the native language of the annotators, and gave de-
Figure 2: Correlation Matrix of Amica Data
A correlation matrix between labels, higher correlations shown in darker colours based on the right-hand key.

Detailed instructions during an induction process and offered time clarification of the task.

**Interannotator Agreement** was calculated through two formulations, two discrete annotators and any two of the 6 annotators. Two Discrete Annotators takes the largest set of overlapping instances by the same two annotators, and achieves the Cohen Kappa score of 0.4503. Any Two Annotators increases the sample to include any two messages that were annotated by any two annotators, and achieves the Cohen Kappa score was 0.4089. Even though two discrete annotators achieved a higher Cohen Kappa score, we followed Wulczyn and colleagues in using the any two users method. This decision was motivated scale and comparison with the original data. The wiki-tox dataset achieved an inter-annotator agreement score of the 0.4469.

**Compare Toxicity and Cyberbullying** As a precursor to the main experiments, and to align this annotation process with Van Hee et al. and Emmery 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., 2015; Emmery et al., 2019). This process received an F1 score of 0.51, showing that multi-label toxicity is not aligned with cyberbullying.

### 3.4 Data Overview

The annotated data was stored in a SQL table and processed as a Pandas Dataframe. The resulting Dataframe was aligned with the wiki-tox dataset on multilabel toxicity for comparison, using the row index of the original Amica dataset expanded. Annotations are stored repetitively in rows, with all data on one line per annotation. This format is pivoted to per message, and multiple annotations of the same message are combined with multi-index references to be used as an inter-annotator agreement set. We maintain annotation lines in a distinct Dataframe, that matches the column labelling of the wiki-tox data set, using the Amica primary key as id. Table 3.4 shows the distribution of labels across the wiki-tox dataset and amica test portion.

<table>
<thead>
<tr>
<th>Label</th>
<th>Wiki Count</th>
<th>Mean</th>
<th>Std</th>
<th>Amica Count</th>
<th>Mean</th>
<th>Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>toxic</td>
<td>63978</td>
<td>0.10</td>
<td>0.29</td>
<td>4102</td>
<td>0.10</td>
<td>0.30</td>
</tr>
<tr>
<td>severe_toxic</td>
<td>63978</td>
<td>0.01</td>
<td>0.08</td>
<td>4102</td>
<td>0.01</td>
<td>0.08</td>
</tr>
<tr>
<td>obscene</td>
<td>63978</td>
<td>0.06</td>
<td>0.23</td>
<td>4102</td>
<td>0.06</td>
<td>0.24</td>
</tr>
<tr>
<td>threat</td>
<td>63978</td>
<td>0.00</td>
<td>0.06</td>
<td>4102</td>
<td>0.00</td>
<td>0.05</td>
</tr>
<tr>
<td>insult</td>
<td>63978</td>
<td>0.05</td>
<td>0.23</td>
<td>4102</td>
<td>0.06</td>
<td>0.23</td>
</tr>
<tr>
<td>identity_hate</td>
<td>63978</td>
<td>0.01</td>
<td>0.10</td>
<td>4102</td>
<td>0.01</td>
<td>0.11</td>
</tr>
</tbody>
</table>

### 4 Method

We used an ensemble of two component models; a finetuned multilingual BERT-base Bi-LSTM and an LSTM model using Multilingual Unsupervised and Supervised Embeddings (MUSE) (Conneau et al., 2017; Devlin et al., 2018).

#### 4.1 BERT-base Bi-LSTM

A Bidirectional Encoder Representation from Transformers or BERT model is a pretrained model that uses bidirectional training to learn contextual attention at a word and sub-word level (Devlin et al., 2018). Where earlier attention based neural approaches would align words semantically with their previous counterparts, to achieve contextually relevant interpretations of words, BERT models take this to a wider vocabulary; relying on massive corpora and training resources. This can be leveraged as a base resource, upon which task specific classifications can be finetuned. In turn, the task specific training examples can be fewer and even lack linguistic phenomena, which the resulting model can rely on its prior training to interpret (Sanh et al., 2020).

**Vocabulary and Tokenisation** Transformer models require distinct vocabulary representations and tokenisation, because they have already been trained with a certain vocabulary representation...
that subsequent training should align with (Zhang et al., 2020). Data needs to be tokenised in the same way to produce a corresponding network, rather than produce duplicate nodes around variant examples of words. Fortunately researchers generally publish tokenisers alongside models (Devlin et al., 2018).

**Training** The BERT-model was trained for 4 epochs over a 10-fold cross validated dataset. The mean validation and training loss for all splits of the data was 0.05.

### 4.2 LSTM

A Long Short-term Memory (LSTM) network uses a Recurrent Neural Network architecture with both forward and backward connections between nodes. They have been applied successfully as stand-alone models on text classification generally, as well as specifically for multilabel toxicity (Nowak et al., 2017). Their backward connection of the network nodes (representing words), allows them to recognises relations over spans of text and across sentences, where the forward passing RNN’s fall short. The LSTM is a simplified version of the above BERT model that uses a bi-LSTM architecture. However, it does allow us to use a specific set of embeddings, and further study has shown that the two approaches’ predictions are not aligned (Sundermeyer et al., 2012). Therefore, it lends itself to worthwhile comparison within an ensemble.

**Embeddings** The LSTM model used MUSE embeddings of word and character representations (Peters et al., 2018; Tulkens et al., 2015; Ginter et al., 2017).

**Training** The LSTM model was trained 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 an Random Forest ensemble of the LSTM and BERT models, on a cross validated training set with grid-searched parameters (Breiman, 2001a; Nowak et al., 2017).

**Random Forest** A random forest is a meta estimator that builds decision tree classifiers on top of sub-models’ predictions for a dataset. In short, deciding which model to rely on in relation to each sample (Breiman, 2001b; Pedregosa et al., 2011). When models reflect meaningfully different distributions of correct predictions, this can improve accuracy and generalise performance.

**Overfitting in Ensembles** A key risk in ensemble training is overfitting. In short, overfitting is when a predictive model becomes reliant on features within its training data that are not present in a general setting (Pourtaheri and Zahiri, 2016). In ensemble models specifically, overfitting is often caused by improper data segmentation, leading to models being tested on data they have previously been trained on. At the ensemble stage, the component models then supply very clear signals for a sample. The ensemble selection relies on these clear predictions, that do not exist in the test phase. To mitigate this all ensemble models have used a stratified $k$-fold structure (Yadav and Shukla, 2016).

### 4.4 Training

**Stratified Kfold** It is methodologically flawed to rely on a single set of training and testing data. A model is likely to repeat the labels of the training samples, and results would reflect qualities specific to the test data; this is also overfitting. To avoid it, it is common practice to use multiple ($k$) ‘folds’ of the data, where an even distribution of samples are tested as well as trained upon (Fushiki, 2011; Pedregosa et al., 2011). For each fold, the predicted validation labels were recorded. This allowed us to collect predictions for all comments in the training data without the component models training on the same comments that they were tested on. These predictions were then used to train the ensemble model.

**Decision Threshold** We use Receiver Under the Curve analysis (ROC) to present decisions from the component models (Fawcett, 2006). We found that using an optimised decision threshold displayed both a fairer comparison with baseline models, and aided the ensemble model when combined with decision probability. Figure 3 shows the ROC curve for each model over each label.

**Hyper-parameter Optimisation** A straightforward way of selecting model parameters is to use grid search; a complete search over a range of specified parameter values (a grid) for a predictor model (Bergstra and Bengio, 2012). Each combination of parameters is tested and the best performing set is used for evaluation.
5 Evaluation

We have evaluated the multilabel predictions using precision, recall, and accuracy via an f1-score. For multilabel classification errors are not discretely true or false. A prediction containing a subset of the actual classes should be considered more accurate than a prediction that contains none. We average each label of the multilabel process using two different methods; micro-averaging and macro-averaging. In micro-averaging all true-positives, true-negatives, false-positives and false-negatives for each class are summed and averaged. Macro-averaging takes the average of the precision and recall of the system on each label.

6 Results

Both component models achieved relevant F1-scores for the multilabel classification of toxicity, and the ensemble approach achieved the highest score. All results are presented in table 6.

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble</td>
<td>0.94</td>
<td>0.70</td>
<td>0.87</td>
<td>0.73</td>
</tr>
<tr>
<td>BERT</td>
<td>0.91</td>
<td>0.67</td>
<td>0.84</td>
<td>0.70</td>
</tr>
<tr>
<td>MUSE</td>
<td>0.85</td>
<td>0.63</td>
<td>0.78</td>
<td>0.65</td>
</tr>
<tr>
<td>Vanilla LSTM</td>
<td>0.75</td>
<td>0.56</td>
<td>0.70</td>
<td>0.58</td>
</tr>
<tr>
<td>SCM &amp; CBOW</td>
<td>0.57</td>
<td>0.42</td>
<td>0.52</td>
<td>0.44</td>
</tr>
</tbody>
</table>

Table 3: Results Table

Results Table of baselines, component, and ensemble models. Results are expressed as AUC, mean Precision, mean Recall, mean F1 for all label. A detailed classification report of each label is shown in Figure 4.

7 Analysis

When using an ensemble approach, it is useful to break down the results of component models to interpret their contribution to an augmented approach. In theory the most effective ensemble will utilise diverse specialisations in component models to contribute to overall accuracy. Component models may perform worse in stand alone settings yet effectively when combined.

7.1 Sub-word Features

The BERT architecture is trained on sub-word features, and we expect that models perform better on sub-word representations than their counterparts. We can analyse this quality by testing their performance against manipulated word samples. In figure 5 each model is tested on samples with deleted characters, and a vanilla Bi-LSTM with word embeddings is added for comparison. By deleting characters within a word, samples words will no longer align with a weight matrix based on words. Models that have trained on subword features should be able to overcome this. We can see that the vanilla model’s performance falls below chance with the loss of one character, showing that the sample words no longer align with the word-based model’s weight matrix. The BERT remains above 0.7 f1 score with 9 characters removed per sentence, revealing that its predictions are less dependent representation.
7.2 Alphanumeric Density

Literature shows that non-alphanumeric characters are a relevant feature in text sentiment classification, be that in the forms of emoji or excessive punctuation (Mohammad, 2018; Zhang et al., 2015). By separating the training samples into bins of alphanumeric density, we can analyse how the models perform based on the presence of non-alphanumeric characters.

7.3 Sentence Length

Sentence length is often a key factor in model performance, mostly defined by the model architecture and implementation. We can see that the MUSE LSTM is substantially more effective at classifying longer sentences, and therefore suitable to enter an ensemble with the BERT model.

7.4 Cross-lingual Performance

The use of English training data and Dutch test data should affect model parameters. We would expect that the model retains relevant weights relating to English language phenomena. To interpret this we created two cross-sections of 500 samples; one containing English loan words and another with no English loan words. We can then compare the performance of the ensemble model between these two cross-sections. The ensemble model achieved an F1-score of 0.9587 on the English section and 0.9189 on the no English. Dutch and English are both Germanic languages and share many grammatical similarities. Moreover, English loan words are frequently used within slang and popular culture.

8 Summary

We have demonstrated that by using multilingual pretrained language models within an ensemble approach, text examples in a target language can be classified for multilabel toxicity. By analysing the performance of baseline, component and ensemble models in relation to textual features we have demonstrated that the BERT model’s 7.1 subword features are integral to classifying text, and that the models’ underlying training affect target language performance. However, detailed analysis of the ensemble model shows that there are an excess in false positives caused by a lack of training examples for some labels. We found that this phenomenon existing in the English language approaches and was exacerbated in a cross-lingual setting.
References


