Multilingual Offensive Language Detection

Mudit Chaudhary **Siddhant Garg** **Sridhama Prakhya** **Priyanka Gohil**
{mchaudhary, siddhantgarg, sprakhya, pgohil}@umass.edu

1 Introduction and Problem statement

The internet and social media have become a breeding ground for hate speech and offensive language. With the unprecedented rate at which the social media platforms generate offensive content along with the opportunity to remain anonymous, robust automated moderation systems are sorely needed. A significant challenge for hate speech detection is the context-dependent nature of the task, and lack of consensus on what constitutes the hate speech. This is further deteriorated by the fact that the social media generated content is filled with poorly written texts with lots of emoticons and hashtags (Kovács et al., 2021). In the multilingual context, this task becomes even more challenging because each language exhibits different complexities about dealing with different cultural ideas (Nozza, 2021). There many expressions that are not inherently offensive, but can be construed so in a specific context - different use of the same word, different audience, different speakers.

There are already many tasks like TRAC-2020 (Kumar et al., 2020), HASOC (Mandl et al., 2019), and GermEval (Struß et al., 2019) that address the issue of hate speech detection in multiple languages. In this work we also worked on developing a multilingual hate speech detection system that follows the OffenseEval task (Zampieri et al., 2020). The task provided a limited labeled dataset, called OLID for hate speech detection for five languages: Arabic, Danish, English, Greek, and Turkish and a relatively large English dataset, called SOLID, that is labeled in a semi-supervised manner for offensive language detection.

In this work, we performed the following three tasks:

- **Task A: Offensive language identification:** Identifying whether the text is offensive or inoffensive (done for all 5 languages).
- **Task B: Categorization of offensive language:** Identifying whether the offensive text is targeted or untargeted. (English only)
- **Task C: Offensive language target identification:** Identifying whether the targeted offensive language is targeting a group, individual, or others. (English only)

In this work, our major focus is on Task A because of the availability of labeled dataset in multiple languages for this task. In order to design a robust Multilingual Hate Speech Detection System, we conducted several experiments. Specifically we explored two directions, with different trade-offs, to tackle this problem.

Our first approach is to design a pipeline where a language classifier (transformer-based) is trained to detect the language of the tweet and then, that tweet will be fed to a BERT model (Devlin et al., 2018) that is pre-trained on that language only. Therefore, we will have five different BERT models for each of the languages, we are working on. One advantage of this method is that since, the BERT is pre-trained on a specific language, it is better able to process nuances associated with that language. However, a single BERT can have around 100M parameters and keeping separate BERT models for each language is not scalable.

Therefore, to address the scalability issue, we proposed a second approach where we leveraged a cross-lingual model, called XLM-RoBERTa (Conneau et al., 2019) to design a single system, that is capable of taking multilingual inputs. We developed multiple baseline models where we trained XLM-R separately on each of the language, and also trained it on all languages together, and compared the results with the separate pretrained BERT models. We also conducted experiments for zero-shot cross-lingual hate speech detection using soft-prompt tuning for cross-lingual transfer

from high resource languages to low-resource languages.

We also leveraged the SOLID dataset, that is available only for Task A in English, to augment the OLID dataset. This was done by translating SOLID data using Google Translate API[1]. We also conducted experiments to address the high class imbalance in the OLID dataset.

2 What we proposed and accomplished

- Develop a pipeline based approach which performs language classification, followed by a separate hate speech detector for each of the five languages.✓
- Build and train a single cross-lingual model that can perform offensive language detection (Task A) over all five languages.✓
- Perform Task B and Task C. ✓
- Zero-shot evaluation along with soft-prompt tuning. ✓
- ~~Use SOLID dataset with 9M tweets~~ Crawled 30K SOLID tweets due to Twitter API rate limit. We translated it to different languages for augmentation. ✓

3 Related Work

Hate speech detection is a long standing problem and there have been many works across different languages to detect offensive language like detecting insults and aggression (Kumar et al., 2018), racism (He et al., 2021), (Greevy and Smeaton, 2004) and not only in English (Mandl et al., 2019), (Zampieri et al., 2020) but in other languages also like Spanish (i Orts, 2019a), German (Wiedemann et al., 2018), Arabic (Safaya et al., 2020) and many other. Traditional approaches used template (Mondal et al., 2017) and keyword (MacAvaney et al., 2019) based detectors. Some of the methods used classical machine learning approaches with Bag-of-Words models with linear classifiers like Naive Bayes and SVM for this task (Greevy and Smeaton, 2004), (Kwok and Wang, 2013) and as with many machine learning based approaches the dynamic shifted to deep learning. There are several deep learning based solutions for offensive language detection task that use RNNs (Del Vigna12 et al., 2017), (Wang et al., 2019), CNNs (Badjatiya et al., 2017), (Gambäck and Sikdar, 2017), and transformer based ensemble models (Alonso et al., 2020). (Kovács et al., 2021) provide several insights and challenges related to nuances about what constitutes hate speech and different context in which a seemingly harmless phrase can be interpreted as offensive.

There have been tremendous efforts to develop systems that can detect offensive language and this is not limited to only English language. Many recent works have proposed annotated datasets and transfer learning strategies for hate speech detection. (Wiedemann et al., 2018) proposed a transfer learning method using BiLSTM-CNN model for hate speech detection on tweets in German Language. KanCMD dataset (Hande et al., 2020) was proposed for multi-task learning for jointly training the model for sentiment analysis and offensive language detection for Kannada language. Similarly, OGRT dataset (Pitenis et al., 2020a) is a Greek annotated dataset for offensive language identification. Many monolingual language specific models like GREEK-BERT (Koutsikakis et al., 2020), and Arabic-BERT (Safaya et al., 2020) have also been proposed recently that are pre-trained on respective languages.

With the large computation costs associated with transformer based models, the focus is being shifted to Multilingual Hate Speech Detection. Many tasks have also been introduced like SemEval-Offense Eval Task 2020 (Zampieri et al., 2020), HASOC (Mandl et al., 2019) for English, German and Hindi, HatEval 2019 (i Orts, 2019b) for English and Spanish, and TRAC-2020 (Kumar et al., 2020) for English, Bengali and Hindi. Due to relatively more popularity of English datasets as compared to other low-resource language datasets, zero-shot multilingual hate speech detection methods (Pelicon et al., 2021; Nozza, 2021), have also been proposed. In this work we have also conducted cross-lingual zero-shot evaluation, where we trained a multilingual model on one language and evaluated it on other languages.

4 Dataset

4.1 OLID

We used Offensive Language Identification Dataset (OLID) (Zampieri et al., 2020) that includes data in the following five languages: English, Danish, Turkish, Arabic and Greek. The datasets for non-English datasets are derived from

[1] https://cloud.google.com/translate

Language	Tweet	A	B	C
English	This account owner asks for peopl eto think rationally.	NOT	—	—
Danish	Du glemmer østeuropæerne som er de værste *Translation: You forget Eastern Europeans, who are the worst*	OFF	—	—
Turkish	Böyle devam et seni gerizekah *Translation: Go on like this, you idiot*	OFF	—	—
Arabic	لعنك الله أيها الحبان يا ابن الكلب. *Translation: May God curse you, O coward, O son of a dog.*	OFF	—	—
Greek	Παραδέξου το, είσαι ξεγυμνωμένος εδώ και καιρό *Translation: Admit it, you have been unfucked for a while now*	OFF	—	—
English	this job got me all the way fucked up real shit	OFF	UNT	—
	etf ari her ass tooo big	OFF	TIN	IND
	@USER We are a cocuntry of morons	OFF	TIN	GRP

Table 1: Annotated examples for all languages and subtasks

the Arabic Dataset (Mubarak et al., 2020), the Danish Dataset (Sigurbergsson and Derczynski, 2019) for Danish, Greek Twitter Dataset (Pitenis et al., 2020b), and the Turkish Dataset (Çöltekin, 2020). The distribution of the data across categories for all languages for task A is shown in Table 2, while Tables 3 and 4 present statistics about the data for the English tasks B and C, respectively. Labeled examples from the different datasets are shown in Table 1.

Lang--uage	Train			Test		
	OFF	NOT	Total	OFF	NOT	Total
English	4400	8840	13240	240	620	860
Danish	384	2576	2960	41	288	329
Turkish	6046	25231	31277	711	2804	3515
Arabic	1550	6289	7839	369	1458	1827
Greek	2486	6257	8743	242	1302	1544

Table 2: Task A (all languages): Statistics about the data

Lang--uage	Train			Test		
	TIN	UNT	Total	TIN	UNT	Total
English	3876	524	4400	213	27	240

Table 3: Subtask B (English): Statistics about the data

Lang--uage	Train			Test		
	IND	GRP	OTH	IND	GRP	OTH
English	2407	1074	395	100	78	35

Table 4: Subtask C (English): Statistics about the data

The dataset contains hierarchical three-level annotation schema also takes both target and types of offense into account. The details of each the tasks are briefly summarized in the following sections:

4.1.1 Task A

Task A deals with the identification of hate speech in tweets. As we can see from column A of Table 1, this task is binary classification with the following two labels.

- *NOT* : text that is neither offensive, nor profane.
- *OFF*: text containing inappropriate language, insults, or threats.

4.1.2 Task B

This task identifies whether a tweet is targeted or untargeted. As described by column B of Table 1, this task is also binary classification task with the following labels:

- TIN: targeted insults or threats towards a group or an individual.
- UNT: untargeted profanity or swearing.

4.1.3 Task C

This task identifies target of offense if the tweet is offensive. This is important because the target of the offense in an important variable that allows us to discriminate between hate speech, which often is towards a group, or cyberbullying, which is typically towards individuals (Zampieri et al., 2020). This is 3-way classification task with the following labels:

- IND: the target is an individual, which can be explicitly mentioned or it can be implicit;

- GRP: the target is a group of people based on ethnicity, gender, sexual orientation, religious belief, or other common characteristic;

- OTH: the target does not fall into any of the previous categories, e.g., organizations, events, and issues.

The usage of five languages with a standardized schema for the purpose of detecting offensive speech is believed to improve dataset consistency. This strategy is in line with current best practices in abusive language data collection (Vidgen and Derczynski, 2020).

4.2 SOLID

Semi-Supervised Offensive Language Identification Dataset (SOLID) (Rosenthal et al., 2020) contains over nine million English tweets labeled in a semi supervised fashion. It follows the same structure for task A, B and C for English language as that of OLID dataset.

OLID was collected using a predefined list of keywords that were more likely to retrieve offensive tweets, which causes offensive tweets in OLID to be explicit and easier to classify. In contrast, the tweets collected for SOLID contain implicit and explicit offensive text. This gives the opportunity to study the performance of various models in hard classification cases.

We were provided with the links to the tweets along with the scores from 0-1 that measures the confidence values whether the tweet is offensive or not. Due to the Twitter API rate limit, we were able to crawl 30K high confidence tweets out of 9M tweets. Specifically, we crawled tweets with scores less than 0.2 that denotes inoffensive tweets and tweets with score greater than 0.7, which denotes highly offensive or abusive sentences. SOLID is labeled only for Task A and not for Task B and C.

4.3 Data preprocessing

For both OLID and SOLID dataset, all user mentions are substituted by @USER for anonymization. For further processing, we have tokenized the tweets as per model requirements and added special tokens, like [CLS] token at the start of input sequences for performing classification.

For language detection task, the first 2500 rows from all 5 language datasets are merged together to form a language detection dataset. An additional column 'language' has been added which specifies five labels corresponding to five languages. The combined dataset is shuffled, tokenized and processed as mentioned above.

5 Baselines

We used two XLM-R based classifiers as baselines for hate speech detection and a zero-shot cross-lingual baseline. Since XLM-R model is a cross-lingual model, and pretrained on over 100 different languages, it provides a strong baseline to compare them against the transformers that are pretrained only on one language and address some of the limitations of the cross-lingual models for hate speech detection. The classifiers were trained on the train split of the OLID dataset and evaluated on the validation split provided by the OffenseEval-2020 task. The proposed experiments in Section 6 also provide a baselines to address class imbalance problem associated with the OLID datasets.

5.1 Language-Specific Baseline

Five different XLM-R models were trained independently on each of the languages of the OLID dataset. Since there are very less labeled tweets for each of the languages, the model overfitted severely and L2 regularization was applied in order to reduce it. Furthermore, we compared the results of this experiment with the BERT-classifier models that were pre-trained on a specific language.

5.2 Cross-lingual Baseline

In order to scale the solution for detective hate speech for multiple languages, we trained a single XLM-R model on all the languages of the OLID dataset. Since the overall size of the dataset is increased as compared to the language-specific baseline in section 5.1, we expected better results from that baseline. However, we saw that the results were very similar even slightly worse. To improve upon this baseline, we augmented our dataset using the tweets from SOLID dataset.

5.3 Zero-Shot Baseline

This baseline is designed to address zero-shot transfer capabilities of the cross-lingual XLM-R model. In this experiment, the XLM-R model, trained on English, from language-specific baseline in section 5.1 was evaluated on other languages of OLID. There is an expected drop in the

F1 scores as compared to the experiments in section 5.1 and section 5.2. To improve the zero-shot learning for this task we also implemented soft-prompt tuning.

6 Our Approach

6.1 Language-specific BERT with language detection head

For offensive language identification (Task A), we have developed a pipeline-based approach as described in Figure 2. We have preprocessed the data as described in the data preprocessing section for the language detection task. This dataset is used to fine-tune the BERT base model (uncased) for identifying the language labels. The language detection results are used to create five separate datasets for each of the five languages. To perform offensive speech detection, language-specific BERTs were used over appropriate language datasets created in the last step.

Leveraging the transfer learning paradigm, we chose to use pre-trained language-wise BERT models for offensive language detection. Our working hypothesis for these experiments was that a per-language model, due to the crisp scope of its training objective, should be able to outperform a cross-lingual or multilingual model, which might suffer from noise compounded from multiple languages. Namely, we first set out by fine-tuning five BERT models which were pre-trained on their respective languages (Arabic, Danish, English, Greek, and Turkish) for the language classification task: Task A. As tokenization varies from language to language, we made sure to use the appropriate tokenization technique for the given language. From our experiments, through a random hyperparameter search, we found that a small learning rate—with a learning rate warm-up followed by decay—works the best for fine-tuning, since model parameters are not varied drastically from their pre-trained initialization.

Overall, as we'll present in later sections, pre-trained BERT models fared quite well for most of the language classification tasks. Coming to the reason for such good performance, it's widely known that encoder-based models, like BERT, perform well on classification tasks; however, we also believe that for hate speech detection tasks, there are certain language-specific cues (such as the use of directed profanity) that make classification rather more explicit.

6.2 Cross-lingual Hate Speech Detection

As discussed in 6.1, BERT models that are pre-trained on a specific language and then finetuned for hate speech detection task, achieve high F1 scores on respective languages as seen in Table10. However, with a large number of parameters, over 100M, for each of the BERT model, this approach is not scalable. Moreover, pretraining datasets is not freely available for all the languages so language-specific BERT may not be able to scale to low-resource languages. To address these issues, we attempt to train a single, unified hate speech detector that is able to take input in any language and give accurate predictions. To this extent, we used XLM-R model as a cross-lingual hate speech detector, that is pretrained on over 100 languages and holds some capability to transfer knowledge from one language to other language.

As described in section 5.1 and section 5.2, we trained two XLM-R classifiers as baseline models on OLID dataset. However, OLID is a small labeled dataset with a high class imbalance. From Table 2, we can see that there disproportionately high number of inoffensive tweets as compared to the offensive tweets, for both train and test splits. To address the class imbalance issue we proposed two approaches: uniform sampling of dataset and data augmentation using SOLID.

6.2.1 Uniform Sampling

Following the given training and validation splits, the model sees more number of inoffensive tweets than offensive tweets during training, hence it can become biased towards predicting a tweet as inoffensive. Moreover, there are different number of tweets for each of the languages too. To balance the amount of information that our cross-lingual model takes as input, we uniformly sampled the inputs in hierarchical manner. Specifically, for each input sentence, first a language was selected uniformly, then the label, *offensive* or *inoffensive* was sampled uniformly and then the tweet, corresponding to the sampled language and label, was randomly selected without replacement. In this way, the model learns from a well-balanced input data when trained for large number of iterations.

6.2.2 Data Augmentation with SOLID

SOLID dataset contains large number of tweets in English with scores about whether the sentence is offensive or not. We used this dataset to augment OLID dataset in two ways. First, we sampled 2000

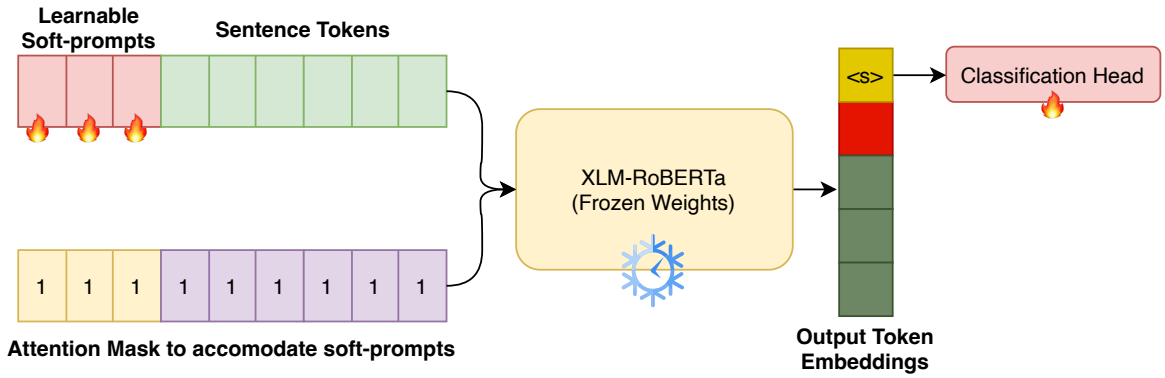

Figure 1: Overview of Prompt-XLM-RoBERTa Classification model. Only the modules marked with 🔥 are updated after backpropagation.

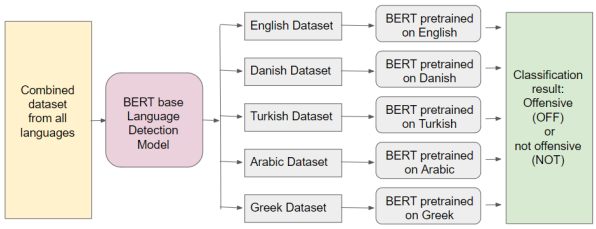

Figure 2: Offensive language identification using language-specific BERT with language detection head

tweets from SOLID with 1000 tweets labeled as *offensive* and other 1000 as *inoffensive*. Then these tweets were translated to other four languages using Google Translate API resulting in a balanced dataset of 10,000 tweets. These 10,000 tweets together with the OLID dataset were used to train XLM-R model for hate speech detection.

In an another experiment, we only took tweets from the SOLID dataset that were labeled as *offensive*. These constituted around 3000 tweets and each of these tweets were then, translated to other four languages resulting over 15,000 tweets marked as *offensive*. These tweets were then combined with the OLID dataset in order to reduce the imbalance between the two categories and XLM-R model was trained on this augmented dataset.

6.3 Soft Prompt Tuning for Zero-Shot Cross-Lingual Transfer

In order to evaluate whether soft-prompt tuning is a practical method for cross-lingual transfer, we augment XLM-RoBERTa with learnable soft-prompts as described by Lester et al. (2021). We hypothesized that by keeping our cross-lingual language model's untouched, the model will not be biased towards the language it is being trained on and will suffer a smaller performance degradation during zero-shot transfer to other languages.

To build such a model, we freeze all the weights of our backbone model i.e., XLM-RoBERTa. Then, we add $n\_prompts$ learnable soft prompt tokens with dimensionality same as model's hidden size (768) to the model's input. To perform unmasked self-attention over the soft-prompts, we extend the existing attention mask. Classification head with learnable weights is attached to the start of the sequence token `<s>`. We present our Prompt-XLM-RoBERTa model in Figure 1.

We also provide an option to choose the number of soft-prompts $n\_prompts$. We also find that prompt-tuning with a larger learning rate converged faster and performed better during evaluation.

The model was only trained on English OLID data and evaluated on other 4 languages in a zero-shot fashion.

6.4 Tasks B & C

For Tasks B & C, we used the same pre-trained, uncased BERT model for English language classification as Task A. Our data pre-processing steps were also identical to the one used for the English model in Task A. We used a similar approach for tokenization and hyperparameter tuning as described in section 6.1. However, unlike Tasks A & B which were binary in nature, we updated the classification head to produce a ternary output for Task C.

Language	Model Variant	Precision	Recall	F1-score
Arabic	XLM-OLID-seperate	**0.8441**	0.7777	**0.8095**
	XLM-OLID-cross	0.8267	0.7886	0.8072
	XLM-OLID-Uniform	0.8104	**0.7994**	0.8049
	XLM-OLID-SOLID-2000-N/OFF	0.84713	0.7208	0.778916
	XLM-OLID-SOLID-3000-OFF	0.7268	0.764	0.745
Danish	XLM-OLID-seperate	0.7272	**0.5853**	**0.6486**
	XLM-OLID-cross	0.6451	0.4878	0.5555
	XLM-OLID-Uniform	**0.7826**	0.439	0.5625
	XLM-OLID-SOLID-2000-N/OFF	0.5641	0.5365	0.55
	XLM-OLID-SOLID-3000-OFF	0.5	**0.5853**	0.5393
English	XLM-OLID-seperate	0.6753	**0.7541**	**0.7125**
	XLM-OLID-cross	0.7692	0.625	0.6896
	XLM-OLID-Uniform	0.7149	0.6583	0.6854
	XLM-OLID-SOLID-2000-N/OFF	0.788	0.6041	0.683
	XLM-OLID-SOLID-3000-OFF	**0.8042**	0.6333	0.7086
Greek	XLM-OLID-seperate	**0.5966**	0.8801	**0.7111**
	XLM-OLID-cross	0.5411	0.8966	0.6749
	XLM-OLID-Uniform	0.5414	**0.9173**	0.6809
	XLM-OLID-SOLID-2000-N/OFF	0.5447	0.9049	0.6801
	XLM-OLID-SOLID-3000-OFF	0.5513	0.909	0.68642
Turkish	XLM-OLID-seperate	0.7475	0.6329	**0.6854**
	XLM-OLID-cross	0.7346	0.6385	0.6832
	XLM-OLID-Uniform	0.6671	**0.6962**	0.6813
	XLM-OLID-SOLID-2000-N/OFF	0.722	0.5682	0.6362
	XLM-OLID-SOLID-3000-OFF	**0.7485**	0.5611	0.64147

Table 5: The table shows the results of the baselines (first 2 rows) and the results of cross-lingual hate speech detector. XLM-OLID-separate: separate XLM-R models are trained for each languag on OLID dataset (section 5.1). XLM-OLID-cross: unified XLM-R model is trained on OLID (section 5.2), XLM-OLID-Uniform: Unified XLM-R model trained on OLID with uniform sampling (section 6.2.1), XLM-OLID-SOLID-2000-N/OFF: XLM-R model trained on OLID and combined with translated tweets of SOLID dataset for both the labels (section 6.2.2), XLM-OLID-SOLID-3000-OFF: XLM-R model trained on complete OLID and only offensive tweets from SOLID (section 6.2.2). From the results we can see that **XLM-OLID-separate achieves best F1-score for all languages**.

Language	Model	F1-score
Arabic	BERT-separate	**0.8187**
	XLM-OLID-separate	0.8095
Danish	BERT-separate	0.5714
	XLM-OLID-separate	**0.6486**
English	BERT-separate	0.7113
	XLM-OLID-separate	**0.7125**
Greek	BERT-separate	**0.7419**
	XLM-OLID-separate	0.7111
Turkish	BERT-separate	**0.7076**
	XLM-OLID-separate	0.6854

Table 6: Comparison of results of per-language pre-trained BERT with XLM-R models trained separately on OLID dataset.

6.5 Technical Details

We used PyTorch[2]==1.10.2 with HuggingFace transformers==4.18.0[3] for using XLM-R, and BERT based models. The models were trained on Google Cloud Platform[4] with the GPUs on Google Colab Pro[5]. We were able to sync colab with our code repository with Google Drive and accessed our training/evaluation scripts from a Colab notebook session. More specifically, we did this by using the "`!`" symbol to run shell scripts from within a notebook cell; here's an example of the syntax we used: `python hello_world.py`.

[2]https://pytorch.org
[3]https://huggingface.co/docs/transformers/index
[4]https://cloud.google.com
[5]https://colab.research.google.com

Language	Model	F1-score
Arabic	XLM-R-separate-ZS	0.4387
	soft-prompt-ZS	**0.46**
Danish	XLM-R-separate-ZS	**0.4316**
	soft-prompt-ZS	0.37
English	XLM-R-separate-ZS	**0.8302**
	soft-prompt-ZS	0.58
Greek	XLM-R-separate-ZS	**0.4128**
	soft-prompt-ZS	0.24
Turkish	XLM-R-separate-ZS	**0.4148**
	soft-prompt-ZS	0.37

Table 7: Comparison of results of zero-shot evaluation of XLM-R model (trained on English) with soft-prompt tuning with 30 prompts. ZS: Zero-Shot

Furthermore, for some experiments, we leveraged the ability to simultaneously run a terminal session while running a notebook. This allowed us to run two experiments within one Colab notebook session. Nevertheless, with the above setup, it should be noted that GPU memory is shared amongst the terminal and notebook sessions.

7 Experiments and Results

We present the results of all the experiments in this section. We used Precision, Recall and F1 score as the metrics for evaluating our models. We also reported average accuracies but since the dataset is highly imbalanced, accuracy is not a good metric. For example, from Table 11, we can see that the accuracy scores are high but F1-scores are comparatively lower. Therefore, we report F1-scores for all the experiments.

All the transformer based models were trained using AdamW optimizer with initial learning rate of 2×10^{-5} and weight decay of 0.5 to reduce overfitting. All the models were trained on Colab GPUs for 50 epochs and batch size of 16. We also used learning rate warmup starting from 0 to initial learning rate for 10 epochs and then decayed to 0 uniformly.

7.1 Zero-Shot Cross-Lingual Hate Speech Detection

Soft-prompt tuning: Results for Prompt-XLM-RoBERTa are presented in Table 8 with $n\_prompts = 20, 30$. Since XLM-R is a cross-lingual model, we also conducted zero-shot evaluation of XLM-R model by first training it only on English tweets of OLID dataset and then evaluating the trained model on the validation splits of all the other languages present in OLID. Table 7 compares the resulting F1-scores of zero-shot evaluation of XLM-R model with soft-prompt tuning. For languages Danish, English, Greek, and Turkish, the XLM-R model outperformed soft-prompt tuning significantly but soft-prompt tuning performed better than English trained XLM-R on Arabic.

7.2 Language-specific BERT with language detection head

For language detection, we experimented with varying number of tweets per language and results were highly dependent on it. Hence, for the best results, we considered equal number of tweets per language, that is, 2500 tweets per language. Our language detection BERT model results are shown in table 9. All tweets in Arabic and Greek languages were identified correctly. Very few tweets from English, Danish and Turkish were classified incorrectly. This is mostly due to the similarity of characters in these three languages.

We present the results for our per-language BERT models for Task A, i.e., offensive language detection in Table 10. Interestingly, our classification results are mostly the same across different languages, barring Danish and Arabic. We believe the poor result in Danish is because of the low-quality corpus that the pre-trained model was trained on. Although we don't substantiate this here, we believe that the superior performance in Arabic could be due to explicit syntactic signals which make it easy to classify offensive language.

7.3 Cross-lingual Hate Speech Detector

In order to design a single unified system for hate speech detection, we trained several variants of XLM-R models on OLID and SOLID dataset. The results are compiled in Table 5 which can be summarized as follows:

- XLM-OLID-separate: Baseline experiment in which XLM-R models were trained independently on each of the five languages on OLID dataset.

- XLM-OLID-cross: Another baseline experiment to understand the cross-lingual capability of XLM-R for hate speech detection. In this experiment a single XLM-R model was trained on the complete OLID dataset.

Language	Precision		Recall		F1-Score		Accuracy	
	20	30	20	30	20	30	20	30
English	0.59	0.56	0.52	0.60	0.55	0.58	0.77	0.76
Arabic	0.49	0.47	0.35	0.44	0.41	0.46	0.80	0.79
Danish	0.27	0.27	0.63	0.60	0.38	0.37	0.76	0.75
Greek	0.31	0.26	0.45	0.22	0.37	0.24	0.76	0.78
Turkish	0.36	0.30	0.51	0.48	0.42	0.37	0.72	0.76
All	0.39	0.36	0.47	0.45	0.43	0.40	0.75	0.73

Table 8: Evaluation result of Prompt-XLM-RoBERTa with $n\_prompts = \{20, 30\}$ trained on English and evaluated on other languages in zero-shot fashion.

Language	Total tweets	Correctly classified
English	860	841
Danish	329	327
Turkish	3515	3510
Arabic	1827	1827
Greek	1544	1544

Table 9: Language Detection result on combined OLID test data from all languages

- XLM-OLID-Uniform: In order to mitigate the class imbalance issue, that is present in OLID, we proposed a uniform sampling approach for OLID dataset in section 6.2.1 that samples different tweets from all the languages along with all the labels uniformly.
- XLM-OLID-SOLID-2000-N/OFF: This experiment covers the data augmentation strategy that was proposed in section 6.2.2. 2000 highest confidence tweets from the crawled SOLID dataset were taken and then translated to other languages using Google Translate API. XLM-R model was trained on these SOLID tweets along complete OLID dataset.
- XLM-OLID-SOLID-3000-OFF: To augment the data and reduce the class imbalance issue, all the offensive tweets from SOLID dataset, that consisted of 3000 tweets, were combined with OLID and XLM-R model was trained on the complete data.

Table 5 shows the comprehensive results of the above experiments. We saw that the F1-score of XLM-OLID-separate was highest among all the variants across all the languages that is counter-intuitive with the proposed experiments. We believe that augmenting OLID with SOLID dataset did not help because SOLID was labeled in a semi-supervised manner and thus, its distribution might be significantly different from OLID dataset. Also, we used Google Translate API to translate the SOLID tweets to other languages, therefore, the performance also depends on the quality of translations which may not be able to cover cultural specific keywords. Furthermore, SOLID was labeled according to what is offensive in English but the same label may not translate parallelly to other languages.

From Table 6, we see that language specific pre-training of BERT model also boosts up the F1-scores for Arabic, Danish, Greek and Turkish. For English, the F1-score of XML-R is slightly better but that is expected because XLM-R is trained on over 100 languages among which proportion of English data is highest.

7.4 Tasks B and Tasks C

Table 11 shows the results of subtasks B & C. These tasks are only English-specific due to the lack of training data for other languages. For task B, we were able to achieve a good F1 score of 0.95, while only 0.58 for task C. The poor performance on Task C could be explained due to the nature of the task—as, it's a ternary classification task, the number of potential predicted classes increases, making classification harder than Tasks A & B. Also, as Task C is ternary, we used macro-averaging to calculate precision/recall/F1-score.

8 Error analysis

8.1 Prompt-XLM-RoBERTa Error Analysis

We found a total of 96 False Negatives and 112 False Positive during English evaluation.

For most of the predicted false positives, seems like the model is focusing on certain keywords such as 'guns', 'gun control', 'crime', and some words which have a 'hateful' connotation if presented out of context. For example, "Killary how does stricter **gun control** work. Looking

Language	Prec.	Recall	F1	Acc.
English	0.7113	0.7113	0.7113	0.8388
Arabic	0.8576	0.7832	0.8187	0.9299
Danish	0.5946	0.5500	0.5714	0.8991
Greek	0.6551	0.8554	0.7419	0.9067
Turkish	0.7184	0.6972	0.7076	0.8835
Average	0.7074	0.7194	0.7102	0.8916

Table 10: Language-wise results for pre-trained BERT models fine-tuned on Task A. Note that accuracy isn't the best possible metric to gauge model performance due to large class imbalance.

Task	Precision	Recall	F1	Acc.
B	0.9207	0.9812	0.9500	0.9083
C	0.6178	0.5865	0.5820	0.6995

Table 11: Results for pre-trained BERT models fine-tuned on Tasks B & C. Note that accuracy isn't the best possible metric to gauge model performance due to large class imbalance.

at Chicago which has some of the strictest gun control laws in the country seems to have problems with **shootings** nearly everyday. Please explain that!", "Be sure to send out the left's antifa **thugs**.", " who cares about the farm. He had no reason to commit **murder**. End of story."

For false negatives, the tweets were more cultural context dependent and subtle as far as negative keywords are concerned. For example, "DavidHogg, you're nobody.", "#OITNB. She is the worst public defender. Trailing her pen uner the words."

The model is able to relate certain keywords that have 'hateful' connotations but is not able to learn, relate, and utilize world knowledge. This might be because we are freezing language model's weights and it has not been pre-trained on data with similar distribution.

8.2 Language Detection error analysis

The results are highly dependent on the distribution of tweets across each language. For example, when the number of examples in English language are more than those in other languages, the model gave 100% accuracy for detecting the English tweets and performed poorly for detecting other languages. The best results were obtained when the number of tweets were the same and maximum possible number for all languages i.e. 2500.

The misclassified examples showed a pattern: misclassified labels in English, Danish and Turkish languages were only amongst these three languages and were never misclassified as Greek or Arabic. This is mostly due to the similarity of characters in these three languages. Overall, the error is very low for language detection.

8.3 Error analysis of per-language pre-trained BERT models

Task A: Looking at misclassified examples in the dataset, it's very apparent that our BERT model is not the best at classifying sentences that are sarcastic in nature. For instance, there are sentences that are not offensive, but contain words that express negative sentiment, and because of this, the sentence gets misclassified as being offensive. Here's an example of such a sentence that was misclassified as being offensive, most likely due to the words "screw up" being present: "*#SEO #Tips: You are the master of your own fate online, so be wise and don't expect pity. If you screw up, nobody else is to blame.*" Quantitatively, for this task, there were both 69 false positives and false negatives.

8.4 Analyzing Cross-lingual XLM-R Hate Speech Detector

We present a detailed error analysis by comparing our predictions on baseline models, and cross-lingual XLM-R model and a model variant that was trained on OLID plus only the offensive tweets of the SOLID dataset. Some examples of the prediction are shown in Table 12 for English language and Table 13 that shows some examples of the tweets in other four languages along with their English translations done by Google Translate API.

For English language, we noticed that there are lots of false positives for all the models. This can be seen in the results of Table 5, where precision value for is very low, specially for XLM-OLID-separate model with 0.67 precision. The reason for this the severe overfitting that we already expected that is due to class imbalance and the proposed variants improved the precision score for English langauge. In particular XLM-OLID-SOLID-3000-OFF shows the highest value of precision score of 0.8042 for English language. But some of the difficult examples were leading to false positive predictions across all the variants of XLM-R cross-lingual models. In Table 12, the

Language	Tweet	Ground Truth	XLM-Separate	XLM-cross	XLM-OLID-SOLID-OFF
English	I guess that's where swamp ass originated	*NOT*	*OFF*	*OFF*	*OFF*
	ConstitutionDay It's very odd for the alt right conservatives to say that we are ruining the constitution just because we want GunControlNow but they are the ones ruining the constitution getting upset because foreigners are coming to this land who are not White wanting to live	*OFF*	*OFF*	*NOT*	*NOT*
	Conservatives Govt have run up debt in spite of austerity cuts while the rich have doubled their wealth. inequality URL via	*OFF*	*NOT*	*NOT*	*NOT*
	ChristineBlaseyFord is only being harassed by members of the Antifa the radical arm of the Democratic Party funded by George Soros. VoteRedToSaveAmerica	*OFF*	*NOT*	*NOT*	*NOT*
	5 years ago Antifa rapper Pavlos Fyssas surrounded by fascists of the Golden Dawn Party in Keratsini, Athens, and stabbed to death. Five years later WE DO NOT FORGET WE DO NOT FORGIVE	*NOT*	*OFF*	*OFF*	*OFF*
	This blow hard was going to push this for his votes. Bad enough his names on it, it will always be the Tappanzee Bridge to us NYrs who had no say in it. I hope they investigate why this was going to be opened and risk lives and no one knew it wasn't ready	*NOT*	*OFF*	*OFF*	*OFF*

Table 12: Error Analysis of Cross-lingual XLM-R Hate Speech Detector for English Language with baselines and proposed approaches. Here *OFF*: offensive, *NOT*: Not Offensive. Some of the keywords are also marked in red. Note that in tweet 1, the ground truth label is *NOT* but all the models predicted as *OFF*. Moreover, bottom two rows shows the examples for false positive predictions and finally tweets 2 and 3 shows that all the models is not able to understand the context of the tweet because of limited data.

last two rows shows the examples of the tweet along with the ground truth label and three different model predictions. Note that all the models predicted *OFF* but the ground truth label was *NOT*

Other Languages: In order to understand the cross-lingual detections, we also sampled some tweets from Arabic, Danish, Greek, and Turkish and presented their predictions in Table 13. Since the languages are from different cultural background, and since we are not experts, the analysis is from our personal experience and not verified. According to our team, the examples shown in Table 13 shows that not all the annotations might be correct. For example, the second Arabic tweet translated to *Oh your lie, oh Atwi, fear God*, or the Danish tweet translated to *do you smoke cannabis??* do not seem offensive or hateful to us but the groundtruth label for these tweets are provided as *OFF*. Moreover, the second Greek tweet translated to *Mr. Apostole we have lost the game. White women like to jump with blacks. Remember Obama's mother who married a Kenyan, divorced her in two years and then remarried an Asian who was beating her.* is annotated as *NOT* but with different context or audiences it may sound as offensive. There were different opinions about this in our team too that further highlight the subjective nature of the problem.

One of the most **surprising results is that augmenting with SOLID data did not increase the performance**, instead the performance was either similar or slightly worse. We reason that this is because, every culture and language have their respective complexities and nuances in dealing with hate-speech. A phrase, inoffensive in one language may be considered offensive in another language. As studied by (Nozza, 2021), the English language may not be able to capture common (taboo) language-specific expressions and do not transfer to different hate speech target types.

Language	Tweet	Ground Truth	XLM-Separate	XLM-cross	XLM-OLID-SOLID-OFF
Arabic	اللعنه بعدد من ماتوا بسبب تحريضك . (O black face O Swedan. And your tricks and you say eh. Curse the number of people who died because of your incitement.)	*OFF*	*NOT*	*OFF*	*OFF*
	يا كذبك يا عطوي خاف الله (Oh your lie, oh Atwi, fear God)	*OFF*	*NOT*	*NOT*	*NOT*
Danish	ryger du hash. ??? (do you smoke cannabis. ???)	*OFF*	*NOT*	*NOT*	*OFF*
	Tosse. Skulle have fløjet med helikopter. Han skulle have lov til at gå (Tosse. Should have flown by helicopter. He should be allowed to go)	*OFF*	*NOT*	*NOT*	*NOT*
Greek	ρε ποσο μαλακες ειμαστε εμεις του ιδιωτικου τομεα .... αλλος πληρωνε και πληρωνει και αλλος τα παιρνει ..... (oh how soft are we in the private sector .... another pays and pays and another takes them .....)	*OFF*	*OFF*	*OFF*	*OFF*
	κυρ Απόστολε το έχουμε χάσει το παιχνίδι. Οι λευκές γυναίκες γουστάρουν να πηδιούνται με μαύρους. Θυμήσου τη μάνα του Οβαμα που παντρεύτηκε τον Κενυάτη, την παράτησε σε δύο χρόνια και μετά ξαναπαντρεύτηκε έναν ασιάτη που την έδερνε (Mr. Apostole we have lost the game. White women like to jump with blacks. Remember Obama's mother who married a Kenyan, divorced her in two years and then remarried an Asian who was beating her)	*NOT*	*NOT*	*NOT*	*NOT*
Turkish	Hayal kurmaya devam et benim shipim onaylı sen böyle laf et dur (Keep dreaming, my ship is approved, you talk like that)	*OFF*	*NOT*	*NOT*	*NOT*
	Boş konuşuyorsunuz icraat yok sizde (You are talking nonsense, you have no action)	*OFF*	*OFF*	*OFF*	*OFF*

Table 13: Error Analysis of Cross-lingual XLM-R Hate Speech Detector for Arabic, Danish, Greek and Turkish with baselines and proposed approaches. The Google Translations are shown in the brackets for each of the tweets. Above examples shows the flaws of the annotated data and also corroborate the difficulty to approach to identify offensive language because of context-dependency. For example, the second Arabic tweet, both the Danish tweets, and first Greek tweet are all labeled as *OFF* but, in my personal opinion and experience, I won't consider them as offensive or hateful. Obviously, these can be offensive in different cultures which is difficult to verify in this work. Here *OFF*: offensive, *NOT*: Not Offensive.

8.5 Tasks B and C

Task B: We found that the misclassified example for this task were highly subjective. In other words, most examples might even be misclassified by human annotators themselves. One pattern amongst examples that get misclassified as being targeted (as opposed to untargeted) is the presence of singular pronouns in the sentences. However, due to the lack of a larger dataset for this task, it's hard to make any concrete conclusions about the model's failure modes. For this task, there were 18 false positives (classified as targeted insult while actually being untargeted) and only 4 false negatives.

Task C: Similar to Task B, most labels in for this task were highly subjective. Most errors for this task were due to the model wrongly classifying group insults as being individual insults. There were 17 such classified examples in the dataset. We also found that most errors in the dataset pertain to examples labels as "other" (OTH). Although the true label was OTH, there were 17 and 12 insults which were misclassified as group and individual, respectively. We believe this is due to

the lack a crisp classification border between the first two classes (individual and group) versus the other class (OTH) in the dataset.

All in all, we believe that the datasets for Tasks B & C should be more balanced and also annotated with higher quality labels. To this end, we recommend using multiple annotators per example and calculating inter-annotator agreement to arrive at more objective class labels.

General Analysis Moreover, we noticed that some hate speech labels are very subjective, making this problem even more challenging. For example, "Anyway this game sucks :)" might or might not be considered offensive but it is labelled as offensive.

9 Contributions of group members

List what each member of the group contributed to this project here. For example:

- Mudit Chaudhary: Developed, trained, and evaluated Prompt-XLM-RoBERTa. Crawled SOLID dataset.
- Siddhant Garg: Worked on Sections 1 and 3. Trained all the baseline models in Section 5. Worked on cross-lingual hate speech detection using XLM-RoBERTa in Section 6.2. Error analysis of Section 8.4.
- Sridhama Prakhya: Worked on training, evaluating, and error-analyzing language-specific BERT models for all tasks (A, B & C) across all five languages in the dataset.
- Priyanka Gohil: Trained and evaluated Language detection model head for language-specific BERT models, preprocessed OLID data for the same, and performed error analysis

10 Conclusion

We conducted several experiments for hate speech detection in five different languages. First we proposed a pipeline based method, to first detect a language and then fed it to a BERT model that is pretrained on that specific language. To scale up the solution, we used cross-lingual XLM-R models and proposed various strategies to address the class imbalance issue and data augmentation using only English tweets. We also did zero-shot experiments, we did soft-prompt tuning and zero-shot evaluation of XLM-R model that were trained on only English language.

Language-specific pretrained BERT and separate XLM-R models, when independently trained for hate speech detection, separated for all languages, give the highest overall results where language-specific pretrained BERTs were slightly better. We also proposed various approaches to improve the cross-lingual hate speech detection by translating SOLID to other languages and balance the distribution between the class labels. But the overall performance of separate models were better. We addressed the limitations of our proposed approaches in terms of scalability and evaluation metrics based on the available datasets. Our experiments showed that cross-lingual transfer is difficult due to cultural-centric dependencies that may not translate well to other languages. We also addressed the limitations of the annotated dataset and certain biases and subjectivity that is associated with the groundtruth labels. From our results on Prompt-XLM-RoBERTa, we observe that prompt tuning without updating the language model is not practical for cross-lingual transfer in the current setting. Moreover, it is worth exploring why performance prompt-tuning on English suffers a performance drop even on English evaluation when compared to full XLM-RoBERTa finetuning. We hypothesize that it might be because of the difference in data distribution between the XLM-RoBERTa pre-training data and OLID hate speech data.

A good future direction would be to pre-train the XLM-RoBERTa on twitter hate speech data followed by prompt-tuning. To improve the performance of cross-lingual hate speech detectors, we need large scale language-specific datasets. There are several datasets available for offensive language detection for various languages, which can be used to create a well-balanced dataset.

References

Alonso, P., Saini, R., and Kovács, G. (2020). Hate speech detection using transformer ensembles on the hasoc dataset. In *International conference on speech and computer*, pages 13–21. Springer.

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

Çöltekin, Ç. (2020). A corpus of Turkish offensive language on social media. In *Proceedings of the 12th Language*

Resources and Evaluation Conference, pages 6174–6184, Marseille, France. European Language Resources Association.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Del Vigna12, F., Cimino23, A., Dell'Orletta, F., Petrocchi, M., and Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gambäck, B. and Sikdar, U. K. (2017). Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, pages 85–90.

Greevy, E. and Smeaton, A. F. (2004). Classifying racist texts using a support vector machine. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 468–469.

Hande, A., Priyadharshini, R., and Chakravarthi, B. R. (2020). Kancmd: Kannada codemixed dataset for sentiment analysis and offensive language detection. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 54–63.

He, B., Ziems, C., Soni, S., Ramakrishnan, N., Yang, D., and Kumar, S. (2021). Racism is a virus: Anti-asian hate and counterspeech in social media during the covid-19 crisis.

i Orts, Ò. G. (2019a). Multilingual detection of hate speech against immigrants and women in twitter at semeval-2019 task 5: Frequency analysis interpolation for hate in speech detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 460–463.

i Orts, Ò. G. (2019b). Multilingual detection of hate speech against immigrants and women in twitter at semeval-2019 task 5: Frequency analysis interpolation for hate in speech detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 460–463.

Koutsikakis, J., Chalkidis, I., Malakasiotis, P., and Androutsopoulos, I. (2020). Greek-bert: The greeks visiting sesame street. In *11th Hellenic Conference on Artificial Intelligence*, pages 110–117.

Kovács, G., Alonso, P., and Saini, R. (2021). Challenges of hate speech detection in social media. *SN Computer Science*, 2(2):1–15.

Kumar, R., Ojha, A. K., Malmasi, S., and Zampieri, M. (2018). Benchmarking aggression identification in social media. In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, pages 1–11.

Kumar, R., Ojha, A. K., Malmasi, S., and Zampieri, M. (2020). Evaluating aggression identification in social media. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 1–5.

Kwok, I. and Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. In *Twenty-seventh AAAI conference on artificial intelligence*.

Lester, B., Al-Rfou, R., and Constant, N. (2021). The power of scale for parameter-efficient prompt tuning.

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., and Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., and Patel, A. (2019). Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th forum for information retrieval evaluation*, pages 14–17.

Mondal, M., Silva, L. A., and Benevenuto, F. (2017). A measurement study of hate speech in social media. In *Proceedings of the 28th ACM conference on hypertext and social media*, pages 85–94.

Mubarak, H., Rashed, A., Darwish, K., Samih, Y., and Abdelali, A. (2020). Arabic offensive language on twitter: Analysis and experiments. *arXiv preprint arXiv:2004.02192*.

Nozza, D. (2021). Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914.

Pelicon, A., Shekhar, R., Martinc, M., Škrlj, B., Purver, M., Pollak, S., et al. (2021). Zero-shot cross-lingual content filtering: Offensive language and hate speech detection.

Pitenis, Z., Zampieri, M., and Ranasinghe, T. (2020a). Offensive language identification in greek. *arXiv preprint arXiv:2003.07459*.

Pitenis, Z., Zampieri, M., and Ranasinghe, T. (2020b). Offensive language identification in greek. *arXiv preprint arXiv:2003.07459*.

Rosenthal, S., Atanasova, P., Karadzhov, G., Zampieri, M., and Nakov, P. (2020). A large-scale semi-supervised dataset for offensive language identification. *arXiv preprint arXiv:2004.14454*.

Safaya, A., Abdullatif, M., and Yuret, D. (2020). KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

Sigurbergsson, G. I. and Derczynski, L. (2019). Offensive language and hate speech detection for danish. *arXiv preprint arXiv:1908.04531*.

Struß, J. M., Siegel, M., Ruppenhofer, J., Wiegand, M., Klenner, M., et al. (2019). Overview of germeval task 2, 2019 shared task on the identification of offensive language.

Vidgen, B. and Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.

Wang, B., Ding, Y., Liu, S., and Zhou, X. (2019). Ynu_wb at hasoc 2019: Ordered neurons lstm with attention for identifying hate speech and offensive language. In *FIRE (Working Notes)*, pages 191–198.

Wiedemann, G., Ruppert, E., Jindal, R., and Biemann, C. (2018). Transfer learning from lda to bilstm-cnn for offensive language detection in twitter. *arXiv preprint arXiv:1811.02906*.

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