Hate-Speech and Offensive Language Detection in Roman Urdu

Hammad Rizwan, Muhammad Haroon Shakeel, Asim Karim
Department of Computer Science,
Lahore University of Management Sciences (LUMS),
Lahore, Pakistan
{hammad.rizwan, m.shakeel, akarim}@lums.edu.pk

Abstract

The task of automatic hate-speech and offensive language detection in social media content is of utmost importance due to its implications in unprejudiced society concerning race, gender, or religion. Existing research in this area, however, is mainly focused on the English language, limiting the applicability to particular demographics. Despite its prevalence, Roman Urdu (RU) lacks language resources, annotated datasets, and language models for this task. In this study, we: (1) Present a lexicon of hateful words in RU, (2) Develop an annotated dataset called RUHSOLD consisting of $10,012$ tweets in RU with both coarse-grained and fine-grained labels of hate-speech and offensive language, (3) Explore the feasibility of transfer learning of five existing embedding models to RU, (4) Propose a novel deep learning architecture called CNN-gram for hate-speech and offensive language detection and compare its performance with seven current baseline approaches on RUHSOLD dataset, and (5) Train domain-specific embeddings on more than 4.7 million tweets and make them publicly available. We conclude that transfer learning is more beneficial as compared to training embedding from scratch and that the proposed model exhibits greater robustness as compared to the baselines.

1 Introduction

In the last decade, online social media platforms have become extremely popular, with users growing exponentially. These platforms provide users with the freedom to express their opinions and ability to interact with people of diverse groups. On one hand, this has resulted in exchanges of ideas and fostered relationships, while on the other, it is exploited to spread, incite, promote, or justify hatred, violence, and discrimination against users based on their gender, religion, race, affiliation with certain groups, and views related to certain events or subjects (e.g., politics) through hateful, offensive, derogatory, or obscene language. If such content is left unaddressed, it has known to lead to acts of violence and conflicts on a broader scale, resulting in problems for the protection of human rights, the rule of law, and freedom of speech, which are essentials for the development of an unprejudiced democratic society.

Most of the social media platforms address this issue by employing techniques such as reporting and manual review by humans, which is limited by the reviewer's speed, ability to understand the evolution of slang, jargon, and familiarity with multilingual content. Apart from these issues, this process generally takes 24 hours, and by that time, the intended damage has already taken place. Moreover, the manual process also poses problems related to the subjective notions of what constitutes hate-speech and offensive language, which might result in misuse of this process to silence minorities and to suppress criticism of official policies, political opposition, or religious beliefs. Thus, an automated system to detect hate speech and offensive language is inevitable.

In the last few years a string of events in Pakistan, such as the lynching of a student due to online anti-religious propaganda against him [1], smearing campaigns against famous political leaders and social media personalities, women being regularly targeted and harassed for sharing their viewpoints online, and the targeting of religious minorities to hurt their religious sentiment has prompted the government to make legislation against online hate-speech such as "National Action Plan" and "The Prevention of Electronic Crime Bill". Such measures speak volumes for the problems related to online hate-speech faced in the country and the need for automated systems to help counter such content.

The majority of the initial research on hate-

[1] https://www.bbc.com/news/world-asia-42970587

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 2512–2522,
November 16–20, 2020.

speech and offensive language detection is mainly focused on the English language. Although English is the official language of Pakistan, Urdu is treated as the National language. People tend to write Urdu using Latin scripts and code-switch between two languages in the same conversation (i.e., alternative use of RU and English languages within the same speech, clause/sentence or constituent/element) (Noor et al., 2015; Fatima et al., 2018).

This unique and informal dialect of communication is known as Roman Urdu (RU). It is a significantly more challenging language to model as compared to formal languages (i.e. languages that follow proper grammatical structure and standard dictionary) due to factors such as colloquial verbiage, improper grammar, spelling variations, self-made abbreviations, and code-switching (Shakeel and Karim, 2020). It is known that the nature of hate-speech content changes with demographics, thus, language resources, labeled datasets, and models for multiple languages are crucial to facilitate the research in this area (Mandl et al., 2019). However, despite its prevalence, RU is under-resourced in this context. To this end, we make following contributions.

- First, we provide a lexicon base of 621 hateful words for the RU language.
- Second, we develop a gold-standard dataset, called Roman Urdu Hate-Speech and Offensive Language Detection (RUHSOLD), from tweets in RU with binary coarse-grained as well as multi-class fine-grained labels.
- Third, we explore the transfer learning capabilities of five existing multilingual embedding models to RU language through extensive experiments.
- Fourth, we propose a novel deep learning model called Convolutional Neural Network n-gram (CNN-gram) and compare its performance with seven baseline models on the RUHSOLD dataset. In our presentation, we demonstrate that CNN-gram displays a greater robustness across both coarse-grained as well as fine-grained classification tasks.
- Fifth, to exhibit contrast with transfer learning of embedding models, we train domain-specific embeddings called "RomUrEm" on more than 4.7 million tweets and compare its performance with five existing pre-trained embeddings in terms of macro F1-score on both tasks of RUHSOLD dataset.

Rest of the paper is organized as follows. Section 2 presents a discussion on the background of the problem. In Section 3, details of the RUHSOLD dataset, it's annotation process, and definition of labels is discussed. Section 4 presents the experimental design and details of baseline models. The proposed model is introduced in Section 5 while we present the results and discussion in Section 6. Finally, we give concluding remarks in Section 7.

2 Background

Research in automatic hate-speech detection has been evolving rapidly over the last five years. Much of the existing research consists of diverse yet related tasks. For instance Waseem and Hovy (2016); Waseem (2016) focus on detection of racism and sexism on Twitter, Davidson et al. (2017) work on differentiating offensive language from hate-speech on Twitter, and de Gibert et al. (2018) focused on hateful and non-hateful speech in a white supremacy forum. Such a diverse set of terminologies has given a rise to problems such as duplication of research, absence of interrelationships, and the lack of re-usability across different strands of the hate-speech and offensive language detection tasks (Kumar et al., 2018). To address this issue Founta et al. (2018) studies these terminologies to find interrelationship between them and provide a selection of labels that eliminate ambiguities of perceivable overlap between them.

In an effort to develop resources, datasets, and models for hate-speech and offensive language detection in multiple languages, a shared task called Hate-Speech and Offensive Content Identification (HASOC) in Indo-European Languages is organized under Forum for Information Retrieval Evaluation (FIRE) (Mandl et al., 2019). This task focuses on Hindi, German and English languages. In the last couple of years, several datasets have been made public in languages such as German (Wiegand et al., 2018), Polish (Ptaszynski et al., 2019), Portuguese (Fortuna et al., 2019), Indonesian (Ibrohim and Budi, 2019), Hindi (Kumar et al., 2018; Mathur et al., 2018), etc. However, despite its prevalence, there is no publicly available dataset for RU to the best of our knowledge.

With regards to language models for hate-speech and offensive language detection, Davidson et al. (2017) have used features such as POS tags, tf-idf vectors, emotion lexicon, and n-grams with multiple classifiers such as logistic regression, naive Bayes, SVM, Decision Tree, and Random Forest. Such approaches rely on local information and are therefore unable to capture context and long-term dependencies in texts where hate-speech is subtle and cannot be judged without taking the entire span of the text into account. With the advent of larger datasets, researchers have shifted to data-hungry deep learning based approaches which are better at learning semantics, contexts, and long-term dependencies (Badjatiya et al., 2017; Agrawal and Awekar, 2018).

Lee et al. (2018) performed a comparative study for machine learning and deep learning models and concluded that deep learning models are more accurate. They also highlighted the fact that different features are important for each hate-speech label, all of which cannot be captured by a single model. Thus, ensemble methods have been used by studies such as Park and Fung (2017), who have used a hybrid-Convolutional Neural Network (CNN) which combines word and character level CNN, Pitsilis et al. (2018), who have used an ensemble of Long Short-term Memory (LSTM) classifiers with majority voting and confidence based aggregation, and Mahata et al. (2019), who used an ensemble of CNN and LSTM based classifiers to capture both salient local information and long term contexts. However, ensembles are carefully selected task-specific combinations which might not generalize well and are computationally expensive. Thus a single model with a greater robustness and generalization is desirable.

The rising success of transfer learning in other deep learning domains such as computer vision and the success of transformer models in many Natural Language Processing (NLP) domains has led to its adoption by many of the researchers who took part in the recent HASOC track at FIRE 2019. These models have outperformed other modeling techniques in five out of eight sub-tasks for different languages. Their success can mainly be attributed to either using ensemble models or performing transfer learning using a pre-trained multilingual transformer embedding model called BERT (Devlin et al., 2019). Keeping these developments in view, we also examine the transfer learning capabilities of five existing embedding models.

3 Roman Urdu Hate-Speech and Offensive Language Detection (RUHSOLD) Dataset

In the literature, researchers create hate-speech datasets by extracting hateful content from online resources by means of a collection of language-specific lexicons of hateful words (Waseem and Hovy, 2016; Basile et al., 2019; Davidson et al., 2017). Despite `Hatebase.org` having the largest collection of multilingual hateful words, it lacks such lexicon base for RU. To this end, we have constructed our own lexicon of hateful words (by searching for such keywords online and interviewing people). this lexicon consists of abusive and derogatory terms along with slurs or terms pertaining to religious hate and sexist language. Using this lexicon along with a separate collection of RU common words, we search and collect $20,000$ tweets and perform a manual preliminary analysis to find new slang, abuses, and identify frequently occurring common terms. The choice to add common RU words is made in order to extract random inoffensive tweets and the tweets that are offensive but do not contain any offensive words e.g.,

> **Tweet**: *Aap apni behan. Beti.. maan ... aur bivi ka march karwa do phir*
>
> **Translation**: *Then do a march of you sister, daughter, mom and wife.*

The tweet is offensive as it targets close relations and tries to demean them but does not contain any hateful/offensive terms/lexicon.

We discard words or terms for which the number of extractable tweets are too few.

Using this updated lexicon we search and collect $50,000$ new tweets. From this updated tweet base, around $10,000$ tweets are randomly sampled for annotations. To avoid issues related to user distribution bias as highlighted by Arango et al. (2019), we restrict a maximum of 120 tweets per user.

To create a gold-standard, the data is manually labeled by three independent annotators and is called Roman Urdu Hate-Speech and Offensive Language Detection (RUHSOLD) dataset. During the annotation process, all conflicts are resolved by a majority vote among three annotators. Tweets on which a consensus cannot be reached or that are reckoned to provide insufficient information for labeling are discarded and replaced by new randomly sampled

Tweet	Translation	Target Label
randi ke bache tu apne hashar ki fikar kar	you son of a prostitute, you should worry for what will happen to you.	Abusive/Offensive
Hindu bhenchod hi ki gaand ma hi keerra hota hay Tum hindu ho hi harami tumhara kabhi 1 baap nhi hota	There are always insects in asses of Hindu sisterfu**kers. These hindus have multple fathers instead of 1	Religious Hate
No wonder you can't make it to First Lady. At least you managed to grab the title of FIRST RANDDI	No wonder you can't make it to First Lady. At least you managed to grab the title of FIRST PROSTITUTE	Sexism
bahria central park karachi forms sold out in two days. Abhi tax maango bhenchodo ka rona shru hojayega	bahria central park karachi forms sold out in two days. Now ask them for tax these motherf**kers start crying.	Profane
pakistan me ptv news or ptv parliment ne hi mulk k liye acha kam kia	in pakistan, only ptv news and ptv parliment has done good work for the country	Neutral

Table 1: Samples of tweets for each label from RUHSOLD dataset

tweets from the data collection. We develop the gold-standard for two sub-tasks. First sub-task is based on binary labels of *Hate-Offensive content* and *Normal content* (i.e., inoffensive language). These labels are self-explanatory. We refer to this sub-task as "coarse-grained classification". Second sub-task defines Hate-Offensive content with four labels at a granular level. These labels are the most relevant for the demographic of users who converse in RU and are defined in related literature. We refer to this sub-task as "fine-grained classification". The objective behind creating two gold-standards is to enable the researchers to evaluate the hate-speech detection approaches on both easier (coarse-grained) and challenging (fine-grained) scenarios. All labels and their definitions are summarized as follows:

Abusive/Offensive: Profanity, strongly impolite, rude or vulgar language expressed with fighting or hurtful words in order to insult a targeted individual or group (Nobata et al., 2016; Founta et al., 2018; Mathur et al., 2018).

Sexism: Language used to express hatred towards a targeted individual or group based on gender or sexual orientation (Waseem and Hovy, 2016; Waseem, 2016; Warner and Hirschberg, 2012).

Religious Hate: Language used to express hatred towards a targeted individual or group based on their religious beliefs or lack of any religious beliefs and the use of religion to incite violence or propagate hatred against a targeted individuals or group (Albadi et al., 2018; Warner and Hirschberg, 2012). e.g

Profane: The use of vulgar, foul or obscene language without an intended target (Davidson et al., 2017; Mandl et al., 2019).

Lablel	Tweet Count
Abusive/Offensive	2,402
Sexism	839
Religious Hate	782
Profane	640
Normal	5,349
Total	10,012

Table 2: Tweet counts with respect to labels for fine-grained classification task

Normal: This contains text that don't fall into the above categories.

Table 1 shows sample tweets for each of the previously described labels along with their English translation.

Religious Hate and Sexism can be combined under the umbrella of the single "Hate-Speech" tag as defined in (Davidson et al., 2017; Golbeck et al., 2017). However, in our case, this defeats the purpose of identifying hate-speech and offensive content at a granular level while at the same time differentiating between the subject matter of abusive content. Thus we refrain from merging any labels.

Table 2 shows the tweet labels and their respective counts. The mode, mean, max, and min length of the tweets are 42, 18, 73, and 1 respectively.

We split the data in train, test, and validation sets with 70,20,10 split ratio using stratification based on fine-grained labels. The use of stratified sampling is deemed necessary to preserve the

same labels ratio across all splits. This way, train split contains 7,209 tweets while test and validation splits have 2003, and 801 tweets respectively. These standard splits along with RU lexicon base is made publicly available to further the research in this direction [2].

4 Experimental Design

In this section, we describe the details of the experiments designed to evaluate the performance of different embeddings, baseline models, and the proposed model for both tasks (i.e., coarse-grained and fine-grained classification).

It is shown in literature that using pre-trained word embeddings for NLP tasks improves the predictive performance of the models (Shakeel et al., 2019). Although, for many years, robust pre-trained embeddings were mainly limited to the English language, in recent years, multilingual embeddings are also made publicly available. These embeddings include LASER (Artetxe and Schwenk, 2019), ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2019), and FastText (Bojanowski et al., 2017). Thus we first compare the out-of-the-box performance of these five pre-trained embeddings. We also fine-tune these embeddings on RUHSOLD dataset in order to gauge their capability of transfer learning to a different domain and language. We use tokenizers for all embedding models from HuggingFace library [3] while for BERT, "base" version is used. To exhibit the contrast to transfer learning, the domain-specific embeddings called *RomUrEm* are also trained. We use Twitter API to collect 4,770,677 random and hate-speech tweets and use word2vec (Mikolov et al., 2013) to train 200 dimensional embeddings. We set the window size to 5, minimum word count to 2, and the number of iterations in pre-training to 10. These embeddings are also made publicly available along with dataset.

Secondly, we compare the performance of seven baseline models through extensive experimentation. The comprehensive details of each model are described below.

4.1 Baseline Models

The baseline models are selected based on their reported good performance for hate-speech detection on multiple datasets. We re-implement these models from the companion code or detailed description, whichever is available.

LSTM+GBDT: Badjatiya et al. (2017) perform multiple experiments using traditional machine learning and deep learning approaches along with various pre-trained embeddings and different ensembles on sexism and racism Twitter dataset in the English Language (Waseem and Hovy, 2016). They conclude that LSTM+GBDT, which utilizes random embeddings followed by an ensemble of LSTM and Gradient Boosting Decision Tree (GBDT) outperforms 16 other models in terms of predictive performance.

FastText+CNN: Kumar et al. (2018) held an open task to model a dataset of Hindi tweets (in both Roman and Devanagari script). Team "DA-LD-Hildesheim" (Modha et al., 2018) employed Fasttext embeddings along with CNN, which outperformed 18 other submitted approaches.

Bi-LSTM with Attention: Bi-LSTM along with attention mechanism have been used consistently for hate-speech detection tasks and is able to achieve top performance on fox news comment dataset (Gao and Huang, 2017).

The rest of the models describe below are taken from HASOC track at FIRE.

BERT+LAMB: Team "3-idiots" (Mishra and Mishra, 2019) utilized pre-trained BERT embedding with LAMB optimizer and achieved top performance for tasks B and C in English language and task B (fine-grained hate speech detection) in Hindi language.

SVM+RF+AB: Team "A3-108" (Mujadia et al., 2019) utilized an ensemble of Linear SVM, Adaboost, and Random Forest along with soft and hard voting mechanism and achieved top performance on Hindi language task C (targeted/untargeted offense).

Domain Embeddings+CNN: Team "QutNocturnal" (Bashar and Nayak, 2019) utilized CNN along with embeddings trained on 494,311 random tweets in Hindi and 5,251 sarcasm tweets in "Hinglish". This architecture was able to achieve top performance on Hindi language task A (binary hate speech detection). To replicate their embeddings, we use RomUrEm.

BERT+LASER+GBDT: Team "HateMonitors" (Saha et al., 2019) utilized pre-trained multilingual BERT and LASER embedding with LGBM classifier in order to achieve top performance for German language task A.

[2] github.com/haroonshakeel/roman_urdu_hate_speech
[3] https://huggingface.co/

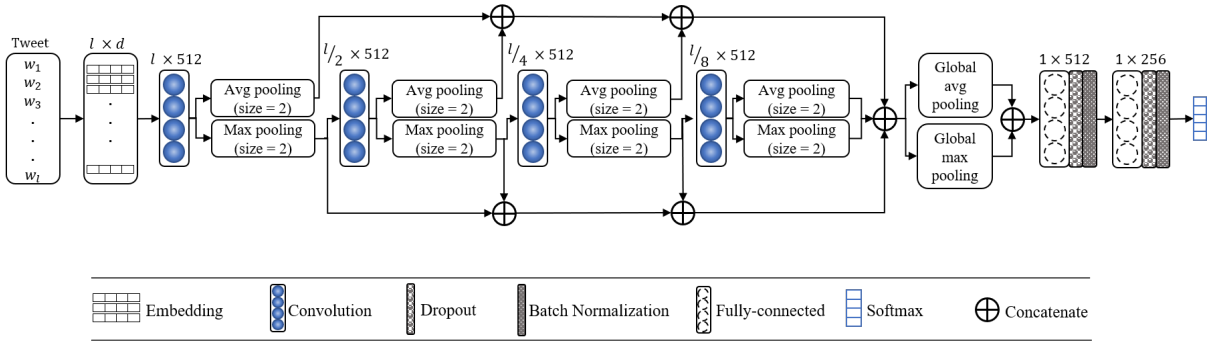

Figure 1: CNN-gram model for hate-speech and offensive language detection in Roman Urdu

5 Proposed Model

In NLP, n-gram information can efficiently be used to learn a certain pattern from text (Attia et al., 2018). The proposed model named Convolutional Neural Network n-gram (CNN-gram) learns patterns based on unigram, bigram, trigram, and quadgram. Complete model architecture is illustrated in Figure 1. Each tweet is first converted to $l \times d$ embedding matrix, where l represents the number of words in the tweet while d is the embedding vector dimension for each word. Four CNN layers are then employed to learn the feature maps. The first CNN layer uses a kernel size of 1 with *ReLU* activation function to learn unigram features followed by max-pooling and average-pooling layers with a pool size of 2. Max-pooling is utilized to drop low activation values from learned representations, which also acts as dimensionality reduction by downsampling the output. The average-pooling is utilized to capture average activations of features. The max-pooled output is then forwarded to another CNN layer which uses kernel size of 2 to learn bigram patterns. This is followed by another set of max-pooling and average-pooling layers identical to the first layer. Similarly, third and fourth CNN layers are used to learn trigram and quadgram patterns respectively. Note that these are not bigram, trigram, and quadgram patterns in "true" sense as one of the two activations is dropped during max-pooling process with a pool size of 2 after every CNN layer. However, on forwarded high activations, the notion of bigram, trigram, and quadgram holds true. Outputs of all four max-pooling and average-pooling layers are concatenated followed by a global average-pooling and global max-pooling layers in parallel, which takes the average and maximum value as the feature corresponding to each filter. These average and maximum feature values are concatenated and are forwarded to a small fully-connected network with two fully-connected layers to squash the information to smaller dimensions. Dropout and batch-normalization after each fully-connected layer is also utilized to avoid feature co-adaptation, followed by *softmax* activation function for final prediction of the label. The categorical cross-entropy is used as the loss function.

All the implementation is done in Python using Keras library with Tensorflow backend running on Nvidia 1080Ti GPU. All weights of the networks are initialized randomly and to mitigate the effect of randomness, random seed is fixed across all experiments. In each of the experiments, the model is trained for 200 epochs. A checkpoint of the learned weights is saved at epoch with best predictive performance on the validation split and is later used to evaluate the test split. The training is stopped if validation error does not decrease for 15 epochs.

5.1 Hyper-parameters Tuning

In the proposed model, the choices of the number of convolutional filters and the number of units in dense layers are made empirically. Figure 1 shows these choices for CNN-gram. The rest of the hyper-parameters were selected by performing a grid search on validation split and utilizing RomUrEm embeddings without finetuning. For available choices of $[0.1, 0.2, 0.3, 0.4, 0.5]$ for dropout rate, 0.5 was found to be most optimal. While for optimizer, "Adam" was chosen over "Adadelta" and "SGD". Finally, 0.002 learning rate turned out to be the optimal choice among $[0.001, 0.002, 0.003.0.004, 0.005]$.

5.2 Evaluation Metrics

We employed the standard metrics that are widely adopted in the literature for measuring the classifi-

Embedding	Without Fine-tuning				With Fine-tuning			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
LASER	0.74	0.74	0.74	0.74	0.76	0.76	0.76	0.76
ELMo	0.80	0.80	0.80	0.80	0.79	0.79	0.79	0.79
BERT	0.68	0.70	0.68	0.67	**0.89**	**0.90**	**0.89**	**0.89**
XLM-RoBERTa	0.53	0.27	0.50	0.35	0.85	0.85	0.85	0.85
FastText	0.74	0.75	0.73	0.73	0.88	0.88	0.88	0.88
RomUrEm	**0.85**	**0.84**	**0.84**	**0.84**	0.88	0.88	0.88	0.88

Table 3: Out-of-the-box performance of different embeddings for coarse-grained classification

	Accuracy	Precision	Recall	F1-score
LSTM+GBDT	0.54	0.58	0.51	0.38
BERT+LASER+GBDT	0.89	0.89	0.89	0.89
FastText+CNN	0.87	0.87	0.87	0.87
SVM+RF+AB	**0.90**	**0.90**	**0.90**	**0.90**
BERT+LAMB	**0.90**	**0.90**	0.89	0.89
Domain Embeddings+CNN	0.88	0.89	0.88	0.88
BiLSTM with Attention	0.86	0.86	0.85	0.85
BERT+CNN-gram	**0.90**	**0.90**	**0.90**	**0.90**
XLM-RoBERTa+CNN-gram	0.88	0.88	0.88	0.88
FastText+CNN-gram	0.81	0.81	0.80	0.80
RomUrEm+CNN-gram	0.89	0.89	0.89	0.89

Table 4: Comparisons of the proposed approach with baseline models on coarse-grained classification

cation performance involving imbalanced dataset. These metrics are *accuracy, macro precision, macro recall,* and *macro F1-score* (Attia et al., 2018). In RUHSOLD dataset, "Normal" class is the dominant while other classes are underrepresented. Thus, it is prudent to use macro-averaging to reflect the model performances as it is insensitive to skewness in class distribution.

6 Results and Discussion

In this section, we discuss the findings of our experiments. Subsequent two subsections present the results on test split of RUHSOLD dataset for coarse-grained and fine-grained classification respectively.

6.1 Coarse-grained Classification

Table 3 summarizes the results for the out-of-the-box predictive performance of the pre-trained embeddings (without utilizing any downstream model). This experiment is intended to highlight the ability of pre-trained embeddings to be adapted for different domains, languages, and tasks. We evaluate the predictive performance in terms of macro F1-score. However, for completeness sake, accuracy, macro precision, and macro recall are also given. We show the results of both variants i.e., without and with fine-tuning (i.e. transfer learning). In case where fine-tuning is not allowed, the domain-specific RomUrEm embeddings, that are trained on a parallel corpora (recall section 4), outperform all other pre-trained embeddings by a significant margin. It yields an F1-score of 0.84, which is followed by ELMo with an F1-score of 0.80. LASER and FastText show comparable performance with F1-scores of 0.74 and 0.73 respectively. The XLM-RoBERTa embeddings yield the poorest performance among all the embeddings with an F1-score of 0.35. The highest performance of RomUrEm can be attributed to the fact that it is trained on tweets having both random and hateful content, while the other embeddings are trained on common texts. Thus, to make a fair comparison, we perform fine-tuning for all embeddings on RUHSOLD dataset in order to perform transfer learning from one domain to the other.

With fine-tuning, the highest F1-score of 0.89 is shown by BERT, which is closely followed by an F1-score of 0.88 of FastText and RomUrEm embeddings. It is worthwhile to note that by allowing fine-tuning, the F1-score is increased from 0.67 to 0.89 for BERT. The highest improvement, however, is shown by XLM-RoBERTa for which, fine-tuning boosts F1-score from 0.35 to 0.85, which is a significant increment of 0.50. These results lead us to deduce that BERT, FastText, and XLM-RoBERTa have a higher potential for transfer learning, thus, are plausible candidates to be used as embeddings for any downstream model for the task of hate-speech detection. These results however, need to be interpreted with caution. As RomUrEm is already trained on a corpora of hate-speech tweets, the same cannot be concluded for this particular embedding. However, it's results can act as a standard to gauge the transfer learning potential of other embeddings. In that regard, BERT embedding has an advantage which exhibits a higher capability for domain adaptation and transfer learning.

Table 4 presents the comparison of the proposed approach with baseline models. In baseline models, LSTM+GBDT has the least F1-score of 0.38. LSTM captures long-term dependencies and order

Embedding	Without Fine-tuning				With Fine-tuning			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
LASER	0.66	0.62	0.42	0.46	0.67	0.59	0.52	0.54
ELMo	**0.70**	0.64	**0.52**	**0.56**	0.60	0.66	0.50	0.55
BERT	0.61	0.60	0.36	0.37	0.77	0.72	0.65	0.67
XLM-RoBERTa	0.53	0.11	0.20	0.14	**0.79**	0.70	**0.75**	**0.72**
FastText	0.62	0.55	0.33	0.35	0.77	0.69	0.63	0.66
RomUrEm	**0.70**	**0.69**	0.51	**0.56**	**0.79**	**0.76**	0.63	0.67

Table 5: Out-of-the-box performance of different embeddings for fine-grained classification

of the words and it is evident that this information is not rich enough for the task of hate-speech detection with respect to this dataset. It is interesting to note that complex ensemble models yield a higher F1-score. For instance, SVM+RF+AB shows an F1-score of 0.90, which is the highest amongst all the baseline approaches. This is closely followed by two other ensemble models i.e., BERT+LASER+GBDT and BERT+LAMB, with an F1-score of 0.89. Other baseline models show similar performance with a variation of ± 0.03 score. Turning now to the proposed model, we employ four embeddings for our experiments by keeping in the view results presented in Table 3. The variation with BERT embedding show the highest F1-score of 0.90, closely followed by the variations with RomUrEm and XLM-RoBERTa, which yields F1-score of 0.89 and 0.88 respectively. The least performance is achieved by using FastText embedding which gives an F1-score of 0.80. Note that the results of the best performing variation of the proposed model are identical to the baseline of SVM+RF+AB. Interestingly, BERT performs similarly to the domain specific RomUrEm embeddings which is consistent with the findings of Table 3 that BERT can be a convincing replacement for domain specific embeddings. However, it is relatively an easier task as compared to fine-grained classification. Thus, more concrete conclusions can be drawn by analyzing the results on fine-grained classification task.

6.2 Fine-grained Classification

Table 5 shows that without fine-tuning, ELMo embedding performs par with domain specific RomUrEm embeddings with an F1-score of 0.56. This is followed by LASER embedding that yields 0.46 F1-score. Other pre-trained embeddings, however, show a poorer performance. For instance, BERT, FastText, and XLM-RoBERTa yield an F1-score of 0.37, 0.35, and 0.14 respectively. Conversely, allowing fine-tuning makes the XLM-RoBERTa top performer amongst the bunch. This is closely followed by BERT and RomUrEm. In general, all embedding models, except ELMo, benefit from fine-tuning. Much like coarse-grained classification, highest improvement with fine-tuning is shown by XLM-RoBERTa with a difference of 0.58. It is interesting to note that RomUrEm is able to achieve an F1-score of 0.67 on this challenging task, which is identical to BERT. These results strengthen our confidence in BERT and XLM-RoBERTa that these models are able to capture the complexity of natural language semantics to a greater or equal extent of domain-specific embedding trained from scratch.

	Accuracy	Precision	Recall	F1-score
LSTM+GBDT	0.53	0.20	0.20	0.15
BERT+LASER+GBDT	0.80	0.73	0.70	0.71
FastText+CNN	0.78	0.70	0.67	0.68
SVM+RF+AB	0.77	0.73	0.62	0.67
BERT+LAMB	0.80	0.72	0.73	0.72
Domain Embeddings+CNN	0.72	0.63	0.52	0.55
BiLSTM with Attention	0.76	0.67	0.63	0.65
BERT+CNN-gram	**0.82**	**0.75**	**0.74**	**0.75**
XLM-RoBERTa+CNN-gram	0.81	0.74	0.71	0.72
FastText+CNN-gram	0.66	0.45	0.41	0.42
RomUrEm+CNN-gram	0.75	0.68	0.61	0.64

Table 6: Comparisons of the proposed approach with baseline models on fine-grained classification

Let us now look at the results of baseline and the proposed model shown in Table 6. As this is a difficult task as compared to coarse-grained classification, it reflects the true learning capabilities of the models. It is evident from the results that LSTM+GBDT is the poorest performer among all the baselines with an F1-score of 0.15, which is in line with the results of coarse-grained classification task. This is followed by Domain Embeddings+CNN, which yield an F1-score of 0.55. These results reflect the difficulty that simpler models face in the identification of hate-speech at fine-grained level. The more complex models such as BERT+LASER+GBDT and BERT+LAMB yield a higher F1-score of 0.71 and 0.72 respectively. We note from Table 4 and 6 that all baseline models utilizing BERT embeddings show a consistent

Tweet [Translation]	Ground Truth	BERT+ LAMB	BERT+ CNN-gram
you people just used shehnaz in whole season fck off bc	Abus./Offen.	Profane	Profane
baqwas band kr tu bhi chali ja. [shut up you get lost]	Abus./Offen.	Normal	Abus./Offen.
tu meri ha ye bat ab teri maa ko batani parygi [you're mine I'll have to tell this to your mom]	Normal	Abus./Offen.	Abus./Offen.
ye chutiya myth hi faila raha hai [this fu**er is spreading myths]	Abus./Offen.	Sexist	Abus./Offen.
na na is moty ko aur dj ko bhenchod samny lao ikthy urao [no no bring this fat and dj fu**er in front and shoot them together]	Abus./Offen.	Profane	Profane

Table 7: Fine-grained classification predictions of best performing baseline and the proposed model

performance across both tasks.

As far as the performance of the variants of the proposed model is concerned, the BERT+CNN-gram has the highest F1-score of 0.75, which is an improvement over the baseline. This result corroborates with the result of coarse-grained classification task presented in Table 4. XLM-RoBERTa based variation exhibit the second highest score of 0.72, which is identical to the baseline BERT+LAMB model. The lowest performance is shown by FastText+CNN-gram with an F1-score of 0.42 while RomUrEm based variant has the F1-score of 0.64. These results substantiate the findings on coarse-grained classification task which suggest that instead of training embeddings, using existing pre-trained embeddings by fine-tuning them on the task in hand is a more perceptive choice. However, a carefully tailored model on top of these embeddings is advantageous. The results of both coarse-grained and fine-grained classification experiments support this conclusion.

We show some examples of fine-grained classification predictions in Table 7 for best performing baseline and proposed model variation to showcase challenges faced with respect to classification at the granular level. It is observed that the models are more "confused" between Abusive/Offensive and Profane as compared to other labels. It shows the limitation of the models with respect to intricacies of human language for subtle differences between profane language and targeted abuse or offensive language.

7 Conclusion and Future Work

In this work, we presented a dataset in Roman Urdu for the task of hate-speech detection in social media content, annotated with five fine-grained labels. We also make publicly available domain-specific embeddings trained on a parallel corpora of more than 4.7 million tweets. Furthermore, an extensive experimentation with respect to multiple embeddings, their power of transfer learning, and comparison with existing baseline models is carried out. As a future research, semantically challenging cases at fine-grained level with respect to complexities of Abusive/Offensive (targeted) and Profane (untargeted) language demand further investigation.

References

Sweta Agrawal and Amit Awekar. 2018. Deep learning for detecting cyberbullying across multiple social media platforms. In *European Conference on Information Retrieval, (ECIR)*, pages 141–153.

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2018. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, (ASONAM)*, pages 69–76.

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 45–54.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics, (ACL)*, 7:597–610.

Mohammed Attia, Younes Samih, Ali Elkahky, and Laura Kallmeyer. 2018. Multilingual multi-class sentiment classification using convolutional neural networks. In *Proceedings of the International Conference on Language Resources and Evaluation, (LREC)*, pages 635–640.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the*

26th International Conference on World Wide Web Companion, (WWWC), pages 759–760.

Md Abul Bashar and Richi Nayak. 2019. Qutnocturnal@ hasoc'19: Cnn for hate speech and offensive content identification in hindi language. In *Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation, (FIRE)*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation, (Sem-Eval)*, pages 54–63.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, (ACL)*, 5:135–146.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *11th International AAAI Conference on Web and Social Media, (ICWSM)*, pages 512–515.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT)*, pages 4171–4186.

Mehwish Fatima, Saba Anwar, Amna Naveed, Waqas Arshad, Rao Muhammad Adeel Nawab, Muntaha Iqbal, and Alia Masood. 2018. Multilingual sms-based author profiling: Data and methods. *Natural Language Engineering, (NLE)*, 24(5):695–724.

Paula Fortuna, João Rocha da Silva, Leo Wanner, Sérgio Nunes, et al. 2019. A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the 3rd Workshop on Abusive Language Online, (ALW3)*, pages 94–104.

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *12th International AAAI Conference on Web and Social Media*.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, (RANLP)*, pages 260–266.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online, (ALW2)*, pages 11–20.

Jennifer Golbeck, Zahra Ashktorab, Rashad O Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A Geller, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, et al. 2017. A large labeled corpus for online harassment research. In *Proceedings of the ACM on Web Science Conference*, pages 229–233.

Muhammad Okky Ibrohim and Indra Budi. 2019. Multi-label hate speech and abusive language detection in indonesian twitter. In *Proceedings of the 3rd Workshop on Abusive Language Online, (ALW3)*, pages 46–57.

Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11.

Younghun Lee, Seunghyun Yoon, and Kyomin Jung. 2018. Comparative studies of detecting abusive language on twitter. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 101–106.

Debanjan Mahata, Haimin Zhang, Karan Uppal, Yaman Kumar, Rajiv Shah, Simra Shahid, Laiba Mehnaz, and Sarthak Anand. 2019. Midas at semeval-2019 task 6: Identifying offensive posts and targeted offense from twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 683–690.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 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, (FIRE)*, pages 14–17.

Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018. Detecting offensive tweets in hindi-english code-switched language. In *Proceedings of the 6th International Workshop on Natural Language Processing for Social Media, (NLPSM)*, pages 18–26.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the International Conference on Neural Information Processing Systems, (NIPS)*, pages 3111–3119.

Shubhanshu Mishra and Sudhanshu Mishra. 2019. 3idiots at hasoc 2019: Fine-tuning transformer neural networks for hate speech identification in indoeuropean languages. In *Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation, (FIRE)*.

Sandip Modha, Prasenjit Majumder, and Thomas Mandl. 2018. Filtering aggression from the multilingual social media feed. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, (TRAC-2018)*, pages 199–207.

Vandan Mujadia, Pruthwik Mishra, and Dipti Misra Sharma. 2019. Iiit-hyderabad at hasoc 2019: Hate speech detection. In *Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation, (FIRE)*.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web, (WWW)*, pages 145–153.

Mehwish Noor, Dr Anwar, Fakharh Muhabat, Bahram Kazemian, et al. 2015. Code-switching in urdu books of punjab text book board, lahore, pakistan. *Communication and Linguistics Studies*, 1(2):13–20.

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on twitter. In *Proceedings of the First Workshop on Abusive Language Online, (ALW1)*, pages 41–45.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT)*, pages 2227–2237.

Georgios K Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Effective hate-speech detection in twitter data using recurrent neural networks. *Applied Intelligence*, 48(12):4730–4742.

Michal Ptaszynski, Agata Pieciukiewicz, and Paweł Dybała. 2019. Results of the poleval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in polish twitter. *Proceedings of the PolEval Workshop*, page 89.

Punyajoy Saha, Binny Mathew, Pawan Goyal, and Animesh Mukherjee. 2019. Hatemonitors: Language agnostic abuse detection in social media. In *Working Notes of Forum for Information Retrieval Evaluation, FIRE*, volume 2517, pages 246–253.

Muhammad Haroon Shakeel and Asim Karim. 2020. Adapting deep learning for sentiment classification of code-switched informal short text. In *The 35th ACM/SIGAPP Symposium on Applied Computing, (ACM-SAC), online event*, pages 903–906.

Muhammad Haroon Shakeel, Asim Karim, and Imdadullah Khan. 2019. A multi-cascaded deep model for bilingual sms classification. In *International Conference on Neural Information Processing, (ICONIP)*, pages 287–298.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the 2nd Workshop on Language in Social Media*, pages 19–26.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the 1st Workshop on NLP and Computational Social Science*, pages 138–142.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Student Research Workshop*, pages 88–93.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language.