Social Network Analysis and Mining (2022) 12:77
https://doi.org/10.1007/s13278-022-00906-8

ORIGINAL ARTICLE

Predicting the type and target of offensive social media posts in Marathi

Marcos Zampieri[1] · Tharindu Ranasinghe[2] · Mrinal Chaudhari[1] · Saurabh Gaikwad[1] · Prajwal Krishna[1] · Mayuresh Nene[1] · Shrunali Paygude[1]

Received: 9 March 2022 / Revised: 9 June 2022 / Accepted: 10 June 2022 / Published online: 9 July 2022

Abstract
The presence of offensive language on social media is very common motivating platforms to invest in strategies to make communities safer. This includes developing robust machine learning systems capable of recognizing offensive content online. Apart from a few notable exceptions, most research on automatic offensive language identification has dealt with English and a few other high-resource languages such as French, German, and Spanish. In this paper, we address this gap by tackling offensive language identification in Marathi, a low-resource Indo-Aryan language spoken in India. We introduce the Marathi Offensive Language Dataset v.2.0 or *MOLD 2.0* and present multiple experiments on this dataset. *MOLD 2.0* is a much larger version of *MOLD* with expanded annotation to the levels B (type) and C (target) of the popular OLID taxonomy. *MOLD 2.0* is the first hierarchical offensive language dataset compiled for Marathi, thus opening new avenues for research in low-resource Indo-Aryan languages. Finally, we also introduce *SeMOLD*, a larger dataset annotated following the semi-supervised methods presented in SOLID (Rosenthal et al. in SOLID: a large-scale semi-supervised dataset for offensive language identification. In: Findings of ACL, 2021).

Keywords Offensive language identification · Hate speech · Machine learning · Deep learning · Low-resource languages

1 Introduction

The widespread of offensive content online such as hate speech and cyber-bullying is a global phenomenon. This has sparked interest in the AI and NLP communities motivating the development of various systems trained to automatically detect potentially harmful content (Ridenhour et al. 2020). Even though thousands of languages and dialects are widely used in social media, the clear majority of these studies consider English only. This is evidenced by the creation of many offensive language resources for English such as annotated datasets (Rosenthal et al. 2021), lexicons (Bassignana et al. 2018), and pre-trained models (Sarkar et al. 2021).

More recently researchers have turned their attention to the problem of offensive content in other languages such as Arabic (Mubarak et al. 2021), French (Chiril et al. 2019), Greek (Pitenis et al. 2020), and Portuguese (Fortuna et al. 2019), to name a few. In doing so, they have created new datasets and resources for each of these languages. Competitions such as OffensEval (Zampieri et al. 2020) and TRAC (Kumar et al. 2020) provided multilingual datasets compiled and annotated using the same methodology. The availability of multilingual has made it possible to explore data augmentation methods (Ghadery and Moens 2020), multilingual word embeddings (Pamungkas and Patti 2019), and cross-lingual contextual word embeddings (Ranasinghe and Zampieri 2020).

In this paper, we revisit the task of offensive language identification for low-resource languages, that is, languages for which few or no corpora, datasets, and language processing tools are available. Our work focus on Marathi, an Indo-Aryan language spoken by over 80 million people, most of whom live in the Indian state of Maharashtra. Even though Marathi is spoken by a large population, it is relatively low-resourced compared to other languages spoken in the region, most notably Hindi, the most similar language to Marathi. We collect and annotate data from Twitter to create the largest Marathi offensive language identification dataset

✉ Marcos Zampieri
marcos.zampieri@rit.edu

[1] Rochester Institute of Technology, Rochester, NY, USA

[2] University of Wolverhampton, Wolverhampton, UK

to date. Furthermore, we train a number of state-of-the-art computational models on this dataset and evaluate the results in detail which makes this paper the first comprehensive evaluation on Marathi offensive language online.

This paper presents the following contributions:

1. We release MOLD 2.0,[1] the largest annotated Marathi Offensive Language Dataset to date. MOLD 2.0 contains more than 3600 annotated tweets annotated using the popular OLID (Zampieri et al. 2019) three-level hierarchical annotation schema; (A) Offensive Language Detection (B) Categorization of Offensive Language (C) Offensive Language Target Identification.
2. We experiment with several machine learning models including state-of-the-art transformer models to predict the type and target of offensive tweets in Marathi. To the best of our knowledge, the identification of types and targets of offensive posts have not been attempted on Marathi.
3. We explore offensive language identification with cross-lingual embeddings and transfer learning. We take advantage of existing data in high-resource languages such as English and Hindi, to project predictions to Marathi. We show that transfer learning can improve the results on Marathi which could benefit a multitude of low-resource languages.
4. Finally, we investigate semi-supervised data augmentation. We create *SeMOLD*, a larger semi-supervised dataset with more than 8000 instances for Marathi. We use multiple machine learning models trained on the annotated training set and combine the scores following a similar methodology described in Rosenthal et al. (2021). We show that this semi-supervised dataset can be used to augment the training set which leads to improves results of machine learning models.

The development MOLD 2.0 and SeMOLD open exciting new avenues for research in Marathi offensive language identification. With these two resources, we aim to answer the following research questions:

- **RQ1**: To which extent is it possible to identify types and targets of offensive posts in Marathi?
- **RQ2**: Our second research question addresses data scarcity, a known challenge for low-resource NLP. We divide it in two parts as follows:
  - **RQ2.1**: How does data size influences performance in Marathi offensive language identification?
  - **RQ2.2**: Do available resources from resource-rich languages combine with transfer-learning techniques aid the identification of types and targets in Marathi offensive language identification?

Previous work Gaikwad et al. (2021) has addressed the identification of offensive posts in Marathi, but the types and targets included in offensive posts, the core part of the popular OLID taxonomy (Zampieri et al. 2019), have not been addressed for Marathi. Finally, with respect to data size and transfer learning, we draw inspiration on recent work that applied cross-lingual models for low-resource offensive language identification (Ranasinghe and Zampieri 2020, 2021) applying it to Marathi.

2 Related work

The problem of offensive content online continues to attract attention within the AI and NLP communities. In recent studies, researchers have developed systems to identify whether a post or part thereof is considered offensive (Ranasinghe et al. 2021) or to predict whether conversations will go awry (Zhang et al. 2018). Popular international competitions on the topic have been organized at conferences such as HASOC (Mandl et al. 2019; Modha et al. 2021), HatEval (Basile et al. 2019), OffensEval (Zampieri et al. 2020), and TRAC (Kumar et al. 2018, 2020). These competitions attracted a large number of participants and they provided participants with various of important benchmark datasets.

A variety of computing models have been proposed to tackle offensive content online ranging from classical machine learning classifiers such as SVMs with feature engineering (Dadvar et al. 2013; Malmasi and Zampieri 2017) to deep neural networks combined with word embeddings (Aroyehun and Gelbukh 2018; Hettiarachchi and Ranasinghe 2019). With the recent development of large pre-trained transformer models such as BERT and XLNET (Devlin et al. 2019; Yang et al. 2019), several studies have explored the use of general pre-trained transformers (Liu et al. 2019; Ranasinghe and Hettiarachchi 2020) while others have worked on fine-tuning models on offensive language corpora such as fBERT (Sarkar et al. 2021).

In terms of languages, due to the availability of suitable datasets, the vast majority of studies in offensive language identification use English data (Yao et al. 2019; Ridenhour et al. 2020). In the past few years, however, more offensive language dataset have been for languages other than English such as Arabic Mubarak et al. (2021), Dutch (Tulkens et al. 2016), French (Chiril et al. 2019), German (Wiegand et al. 2018), Greek (Pitenis et al. 2020), Italian (Poletto et al. 2017), Portuguese (Fortuna et al. 2019), Slovene (Fišer et al.

[1] Dataset available at: https://github.com/tharindudr/MOLD.

Table 1 MOLD v2.0—distribution of label combinations

A	B	C	Training	Test	Total
OFF	TIN	IND	503	51	554
OFF	TIN	OTH	80	56	136
OFF	TIN	GRP	157	51	208
OFF	UNT	–	327	102	429
NOT	–	–	2034	250	2,284
All			3101	510	3,611

2017), Turkish (Çöltekin 2020), and many others. To the best of our knowledge, the only Marathi dataset available to date is the aforementioned Marathi Offensive Language Dataset (MOLD) (Gaikwad et al. 2021), a manually annotated dataset containing nearly 2500 tweets. Our work builds on MOLD by applying the same data collection methods to expand it in terms of both size and annotation.

Finally, multilingual offensive language identification is a recent trend that takes advantage of large pre-trained cross-lingual and multilingual models such as XLM-R (Conneau et al. 2019). Using this architecture, it is possible to leverage available English resources to make predictions in languages with less resources helping to cope with data scarcity in low-resource languages (Ranasinghe and Zampieri 2020; Ranasinghe et al. 2021).

3 Data collection

MOLD 2.0 builds on the research presented in Gaikwad et al. (2021) which introduced MOLD 1.0. The annotation of both MOLD 1.0 and MOLD 2.0 follows the OLID annotation taxonomy which includes three levels (labels in brackets):

- **Level A**: Offensive (OFF)/Non-offensive (NOT).
- **Level B**: Classification of the type of offensive (OFF) tweet—Targeted (TIN)/Untargeted (UNT).
- **Level C**: Classification of the target of a targeted (TIN) tweet—Individual(IND)/Group(GRP) or Other(OTH).

Our initial dataset (MOLD 1.0) consisted of nearly 2,500 tweets. As shown in Table 1, we collected 1,100 additional instances for MOLD 2.0 resulting in a dataset of 3,611 tweets according to the same methodology described in Gaikwad et al. (2021). Data collection was carried out with a data extraction script which utilized the Tweepy[2] library along with the API provided by Twitter.

[2] Tweepy Python library documentation is available on https://www.tweepy.org/.

Table 2 Four tweets from the dataset, with their labels for each level of the annotation schema

Tweet	A	B	C
(Who is your favorite?)	NOT	–	–
(Stupid, what else?)	OFF	UNT	–
(Damn slut)	OFF	TIN	IND
(This is a government of thankless sick-heads)	OFF	TIN	GRP
(These artists and media without standards have sold themselves to work for the state government)	OFF	TIN	OTH

English translations are inside brackets

As MOLD 1.0 was only annotated on OLID Level A, in MOLD 2.0 we expand the annotation to the full three-level OLID taxonomy annotating Level B and Level C. Examples from the dataset along with English translation are presented in Table 2. The annotation was carried out by the 3 native speakers of Marathi. The annotators were a mix of male (1) and female (2) Master's students working in the project. We provided the annotators with guidelines on how to annotate the data and supervised the process with periodic meetings to make sure they were correctly following the guidelines. We report an inter-annotator agreement of 0.79 Cohen's kappa (Carletta 1996) on the three levels.

Finally, following the same methodology described for MOLD 2.0, we collected an additional 8000 instances from Twitter to create SeMOLD, a larger dataset with semi-supervised annotation presented in Sect. 6.

4 Experiments and evaluation

We experimented with several machine learning models trained on the training set, and evaluated by predicting the labels for the held-out test set. As the label distribution is highly imbalanced, we evaluate and compare the performance of the different models using macro-averaged $F1$-score. We further report per-class Precision (P), Recall (R), and $F1$-score ($F1$), and weighted average. Finally, we compare the performance of the models against simple majority and minority class baselines.

4.1 SVC

Our simplest machine learning model is a linear support vector classifier (SVC) trained on word unigrams. Before the emergence of neural networks, SVCs have achieved

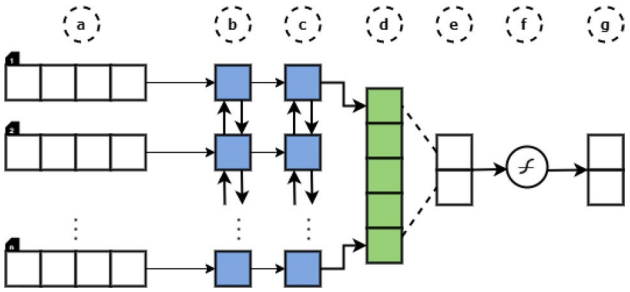

Fig. 1 The BiLSTM model for Marathi offensive language identification. The labels are **a** input embeddings, **b, c** two BiLSTM layers, **d, e** fully connected layers; **f** softmax activation, and **g** final probabilities

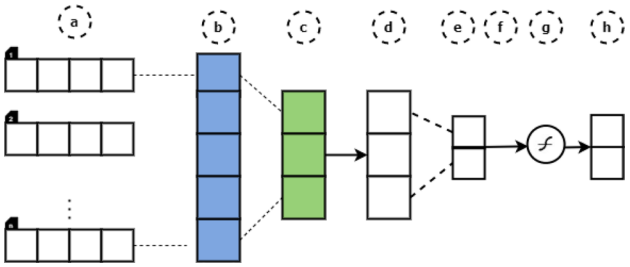

Fig. 2 CNN model for Marathi offensive language identification. The labels are **a** input embeddings, **b** 1DCNN, **c** max pooling, **d, e** fully connected layer; **f** with dropout, **g** softmax activation, and **h** final probabilities

state-of-the-art results for many text classification tasks (Schwarm and Ostendorf 2005; Goudjil et al. 2018) including offensive language identification (Zampieri et al. 2019; Alakrot et al. 2018). Even in the neural network era, SVCs produce an efficient and effective baseline.

4.2 BiLSTM

As the first embedding-based neural model, we experimented with a bidirectional long short-term-memory (BiLSTM) model, which we adopted from a pre-existing model for Greek offensive language identification (Pitenis et al. 2020). The model consists of (i) an input embedding layer, (ii) two bidirectional LSTM layers, and (iii) two dense layers. The output of the final dense layer is ultimately passed through a softmax layer to produce the final prediction. The architecture diagram of the BiLSTM model is shown in Fig. 1. Our BiLSTM layer has 64 units, while the first dense layer had 256 units.

4.3 CNN

We also experimented with a convolutional neural network (CNN), which we adopted from a pre-existing model for English sentiment classification (Kim 2014). The model

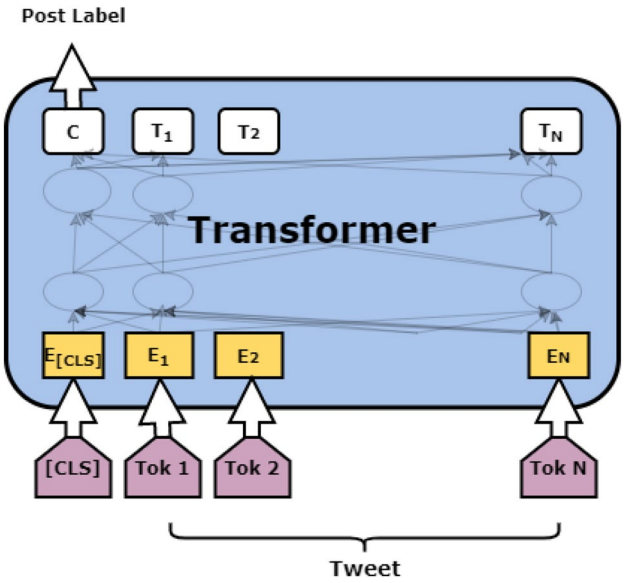

Fig. 3 Transformer model for Marathi offensive language identification (Ranasinghe and Zampieri 2020)

consists of (i) an input embedding layer, (ii) 1-dimensional CNN layer (1DCNN), (iii) max pooling layer and (iv) two dense layers. The output of the final dense layer is ultimately passed through a softmax layer to produce the final prediction (Fig. 2).

For the BiLSTM and CNN models presented above, we set three input channels for the input embedding layers: pre-trained Marathi FastText embeddings[3] (Bojanowski et al. 2017), Continuous Bag of Words Model for Marathi[4] (Kumar et al. 2020) as well as updatable embeddings learned by the model during training.

4.4 Transformers

Finally, we experimented with several pre-trained transformer models. With the introduction of BERT (Devlin et al. 2019), transformer models have achieved state-of-the-art performance in many natural language processing tasks (Devlin et al. 2019) including offensive language identification (Ranasinghe and Zampieri 2020; Ranasinghe et al. 2021; Sarkar et al. 2021; Ranasinghe et al. 2021). From an input sentence, transformers compute a feature vector $\boldsymbol{h} \in \mathbb{R}^d$, upon which we build a classifier for the task. For this task, we implemented a softmax layer, i.e., the predicted probabilities are $\boldsymbol{y}^{(B)} = \text{softmax}(W\boldsymbol{h})$, where $W \in \mathbb{R}^{k \times d}$ is

[3] Marathi FastText embeddings are available on https://fasttext.cc/docs/en/crawl-vectors.html.

[4] Marathi word embeddings are available on https://www.cfilt.iitb.ac.in/~diptesh/embeddings/.

Table 3 Results for offensive language detection (level A)

Type	Model	OFF			NO			Weighted			Macro-*F*1
		P	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1	
Traditional	SVM	0.72	0.68	0.70	0.84	0.80	0.82	0.80	0.76	0.78	0.74
BiLSTM	CBOW	0.75	0.71	0.73	0.87	0.83	0.85	0.83	0.81	0.80	0.77
	fastText	0.75	0.72	0.74	0.88	0.83	0.86	0.84	0.82	0.81	0.78
	Self-learned	0.73	0.69	0.71	0.85	0.81	0.83	0.79	0.76	0.77	0.76
CNN	CBOW	0.77	0.73	0.75	0.89	0.85	0.86	0.85	0.83	0.82	0.80
	fastText	0.78	0.74	0.76	0.90	0.86	0.87	0.87	0.86	0.83	0.81
	Self-learned	0.76	0.72	0.74	0.88	0.83	0.84	0.83	0.81	0.82	0.80
Transformers	mBERT	0.79	0.75	0.77	0.91	0.87	0.88	0.88	0.86	0.84	0.82
	XLM-R	0.81	0.77	0.79	0.93	0.89	0.90	0.90	0.88	0.86	0.84
	IndicBERT	0.83	0.79	0.81	0.95	0.91	0.91	0.91	0.89	0.88	**0.85**
Baseline	All OFF	1.00	0.51	0.67	0.00	0.00	0.00	1.00	0.51	0.67	0.33
	All NOT	0.00	0.00	0.00	1.00	0.49	0.65	1.00	0.49	0.65	0.33

We report precision (*P*), recall (*R*), and *F*1 for each model/baseline on all classes (OFF, NOT), and weighted averages. Macro-*F*1 is also listed (best in bold)

the softmax weight matrix and k is the number of labels. In our experiments, we used three pre-trained transformer models available in HuggingFace model hub (Wolf et al. 2020) that supports Marathi; mBERT (Devlin et al. 2019), xlm-roberta-large (Conneau et al. 2019) and IndicBERT (Kakwani et al. 2020). The implementation was adopted from the *DeepOffense* Python library.[5] The overall transformer architecture is available in Fig. 3.

For the transformer-based models, we employed a batch-size of 16, Adam optimizer with learning rate 2e−5, and a linear learning rate warm-up over 10% of the training data. During the training process, the parameters of the transformer model, as well as the parameters of the subsequent layers, were updated. The models were evaluated while training using an evaluation set that had one fifth of the rows in training data. We performed early stopping if the evaluation loss did not improve over three evaluation steps. All the models were trained for three epochs.

4.5 Offensive language detection

The performance on discriminating between offensive (OFF) and non-offensive (NOT) posts is reported in Table 3. We can see that all models perform better than the majority baseline. As expected, transformer-based models outperform other machine learning models. From the transformer models, IndicBERT model (Kakwani et al. 2020) outperforms general multilingual transformer models such as mBERT Devlin et al. (2019) and xlm-roberta-large (Conneau et al. (2019)) providing 0.85 Macro-*F*1 score on the test set.

[5] DeepOffense is available as a pip package in https://pypi.org/project/deepoffense/.

4.6 Categorization of offensive language

In this set of experiments, the models were trained to discriminate between targeted insults and threats (TIN) and untargeted (UNT) offenses. The performance of various machine learning models on this task is shown in Table 4. Similar to level A, transformer models outperformed other machine learning models in this set of experiments too. Furthermore, IndicBERT model (Kakwani et al. 2020) performs best from the transformer model with providing 0.74 Macro-*F*1 score.

4.7 Offensive language target identification

Here, the models were trained to distinguish between three targets: a group (GRP), an individual (IND), or others (OTH). In Table 5, we can see that all the models achieved similar results, far surpassing the random baselines, with a slight performance edge for the transformer models. Similar to the previous levels, IndicBERT performed best in this level too providing 0.65 Macro-*F*1 score.

5 Transfer-learning experiments

The main idea of the methodology is that we train a classification model on a resource-rich, typically English, using a cross-lingual language model, save the weights of the model and when we initialize the training process for Marathi, start with the saved weights from English. Previous work has shown that a similar transfer learning approach can improve the results for Arabic, Greek and Hindi (Ranasinghe and Zampieri 2020, 2021; Ranasinghe and Zampieri 2021). We only experimented transfer-learning

Table 4 Results for offensive language categorization (level B)

Type	Model	TIN			UNT			Weighted			Macro-*F*1
		P	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1	
Traditional	SVM	0.85	0.89	0.87	0.41	0.31	0.39	0.82	0.79	0.80	0.48
BiLSTM	Word2vec	0.89	0.93	0.91	0.65	0.53	0.59	0.88	0.81	0.83	0.66
	fastText	0.90	0.93	0.93	0.67	0.55	0.61	0.89	0.82	0.85	0.68
	Self-learned	0.89	0.92	0.90	0.61	0.49	0.55	0.84	0.77	0.79	0.62
CNN	Word2vec	0.91	0.93	0.92	0.66	0.55	0.61	0.89	0.83	0.85	0.69
	fastText	0.93	0.95	0.94	0.68	0.58	0.63	0.90	0.84	0.86	0.70
	Self-learned	0.89	0.93	0.91	0.62	0.50	0.56	0.85	0.78	0.80	0.64
Transformers	mBERT	0.94	0.90	0.92	0.68	0.58	0.64	0.90	0.85	0.87	0.71
	XLM-R	0.94	0.90	0.92	0.70	0.60	0.66	0.91	0.86	0.88	0.72
	IndicBERT	0.95	0.91	0.93	0.72	0.61	0.68	0.93	0.88	0.90	**0.74**
Baseline	All TIN	0.89	1.00	0.94	0.00	0.00	0.00	0.79	0.89	0.83	0.47
	All UNT	0.00	0.00	0.00	0.11	1.00	0.20	0.01	0.11	0.02	0.10

We report precision (*P*), recall (*R*), and *F*1 for each model/baseline on all classes (TIN, UNT), and weighted averages. Macro-*F*1 is also listed (best in bold)

Table 5 Results for offense target identification (level C)

Type	Model	GRP			IND			OTH			Weighted			Macro-*F*1
		P	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1	
Traditional	SVM	0.82	0.94	0.81	0.63	0.33	0.44	0.45	0.18	0.26	0.58	0.62	0.56	0.52
BiLSTM	Word2vec	0.82	0.94	0.81	0.67	0.38	0.48	0.49	0.25	0.31	0.62	0.65	0.60	0.56
	fastText	0.84	0.95	0.83	0.68	0.40	0.50	0.51	0.27	0.33	0.64	0.67	0.62	0.58
	Self-learned	0.84	0.95	0.84	0.69	0.41	0.51	0.51	0.28	0.34	0.65	0.67	0.62	0.58
CNN	Word2vec	0.84	0.96	0.83	0.69	0.40	0.50	0.51	0.27	0.33	0.64	0.67	0.62	0.58
	fastText	0.86	0.96	0.84	0.70	0.41	0.51	0.52	0.28	0.34	0.66	0.69	0.64	0.60
	Self-learned	0.84	0.96	0.83	0.69	0.40	0.50	0.51	0.27	0.33	0.64	0.67	0.62	0.58
Transformers	mBERT	0.86	0.97	0.85	0.72	0.43	0.53	0.54	0.32	0.38	0.68	0.70	0.65	0.62
	XLM-R	0.87	0.97	0.85	0.72	0.43	0.53	0.56	0.34	0.40	0.70	0.71	0.66	0.63
	IndicBERT	0.87	0.97	0.85	0.74	0.45	0.55	0.58	0.36	0.42	0.72	0.73	0.68	**0.65**
Baseline	All GRP	0.37	1.00	0.54	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.37	0.20	0.18
	All IND	0.00	0.00	0.00	0.47	1.00	0.64	0.00	0.00	0.00	0.22	0.47	0.30	0.21
	All OTH	0.00	0.00	0.00	0.00	0.00	0.00	0.16	1.00	0.28	0.03	0.16	0.05	0.09

We report precision (*P*), recall (*R*), and *F*1 for each model/baseline on all classes (GRP, IND, OTH), and weighted averages. Macro-*F*1 is also listed (best in bold)

experiments with the transformer models as they provided better results than other embedding models in Sect. 4.

We first trained a transformer-based classification model on a resource-rich language. We used different resource-rich languages for each level which we describe in the following sections. Then we save the weights of the transformer model as well as the softmax layer. We use this saved weights from the resource-rich language to initialize the weights for Marathi.

5.1 Offensive language detection

For the level A, we used several datasets as the resource-rich language. As the first resource-rich language, we used English which can be considered as the language with highest resources for offensive language identification. We specifically used the OLID (Zampieri et al. 2019) level A tweets which is similar to the level A of MOLD 2.0. Also, in order to perform transfer learning from a closely related language to Marathi, we utilized a Hindi dataset used in the HASOC 2020 shared task (Mandl et al. 2020). Both the English and Hindi datasets we used for transfer learning experiments

Table 6 Transfer learning results for offensive language identification ordered by Macro (M)-*F*1 for MOLD 2.0

Language	Model	M *F*1	W *F*1
Hindi	XLM-R	0.87	0.89
English	XLM-R	0.86	0.88
–	IndicBERT	0.85	0.88
Hindi	IndicBERT	0.85	0.88
English	IndicBERT	0.84	0.87
–	XLM-R	0.84	0.86
Hindi	mBERT	0.84	0.86
English	mBERT	0.83	0.85
–	mBERT	0.82	0.84

We also report weighted (W) *F*1 scores. For the comparison purpose, we also report the results for XLM-R, mBERT and IndicBERT when trained from scratch too

Table 7 Transfer learning results for categorization of offensive language ordered by Macro (M)-*F*1 for MOLD 2.0

Language	Model	M *F*1	W *F*1
English	XLM-R	0.75	0.91
–	IndicBERT	0.74	0.90
English	mBERT	0.73	0.88
English	IndicBERT	0.72	0.89
–	XLM-R	0.72	0.88
–	mBERT	0.71	0.87

We also report weighted (W) *F*1 scores. For the comparison purpose, we also report the results for XLM-R, mBERT and IndicBERT when trained from scratch too

contain Twitter data making them in-domain with respect to *MOLD 2.0*. The results are shown in Table 6.

As can be seen, the use of transfer learning substantially improved the monolingual results for mBERT and XLM-R. However, the IndicBERT model which performed best in the monolingual experiments did not improve with the transfer learning approach. We believe that this can be due to the fact that the IndicBERT model is not cross-lingual. The best cross-lingual results were shown by the XLM-R model. From the two languages that we performed transfer learning, Hindi outperformed the results obtained using the English dataset suggesting that language similarity played a positive role in transfer learning.

5.2 Categorization of offensive language

For level B, we used the OLID level B as the initial task to train the transformer-based classification model. However, as far as we know, there are no datasets equivalent to MOLD 2.0 level B in related languages to Marathi such as Hindi and Bengali. Therefore, for level B, we only used English OLID level B as the initial task. The results are shown in Table 7.

Table 8 Transfer learning results for offensive language target identification ordered by Macro (M)-*F*1 for MOLD 2.0

Language	Model	M *F*1	W *F*1
English	XLM-R	0.74	0.90
–	IndicBERT	0.74	0.89
English	IndicBERT	0.73	0.88
English	mBERT	0.72	0.88
–	XLM-R	0.72	0.88
–	mBERT	0.71	0.87

We also report weighted (W) *F*1 scores. For the comparison purpose, we also report the results for XLM-R, mBERT and IndicBERT when trained from scratch too

Table 9 Semi-supervised data augmentation results for offensive language identification in MOLD 2.0

Model	MOLD	SeMOLD + MOLD
mBERT	0.82	0.82
XLM-R	0.84	0.84
IndicBERT	0.85	0.84

We report the Macro-*F*1 scores with the augmented data from SeMOLD in **SeMOLD + MOLD** column. For the comparison purpose, we report the results without data augmentation in **MOLD** column

As can be seen in the results, transfer learning improved the results for level B in XLM-R and mBERT. Similar to level A, IndicBERT performance was not improved with transfer learning. XLM-R with transfer learning provided the best results with 0.75 Macro-*F*1 score.

5.3 Offensive language target identification

As there are no equivalent datasets similar to level C in MOLD 2.0 in related languages, we only used OLID level C as the initial dataset.

As can be seen in Table 8 transfer learning improved the results of XLM-R and mBERT. However, in this level too, transfer learning did not improve the performance of IndicBERT. Overall, XLM-R with transfer learning provided the best result with 0.74 Macro-*F*1 score.

6 SeMOLD: semi-supervised data augmentation

For the semi-supervised experiments, we collected additional 8,000 Marathi, using the same methods described in Sect. 3. Rather than labeling them manually we followed a semi-supervised approach described to annotate SOLID Rosenthal et al. (2021). We first selected the three best machine learning classifiers we had from Sect. 4:

Table 10 Semi-supervised data augmentation results for categorization of offensive language in MOLD 2.0

Model	MOLD	SeMOLD + MOLD
mBERT	0.71	0.73
XLM-R	0.72	0.74
IndicBERT	0.74	0.76

We report the Macro-*F*1 scores with the augmented data from SeMOLD in **SeMOLD + MOLD** column. For the comparison purpose, we report the results without data augmentation in **MOLD** column

Table 11 Semi-supervised data augmentation results for offensive language target identification in MOLD 2.0

Model	MOLD	SeMOLD + MOLD
mBERT	0.62	0.64
XLM-R	0.63	0.65
IndicBERT	0.65	0.68

We report the Macro-*F*1 scores with the augmented data from SeMOLD in **SeMOLD + MOLD** column. For the comparison purpose, we report the results without data augmentation in **MOLD** column

mBERT, XLM-R and IndicBERT. Then, for each instance in the larger dataset, we saved the labels from each machine learning model. We release this larger dataset as SeMOLD: Semi-supervised Marathi Offensive Language Dataset. We use filtered SeMOLD instances to augment the training set. We only performed the data augmentation experiments for the transformer models.

6.1 Offensive language detection

In the data augmentation process for level A, we augmented instances from SeMOLD, where at least two machine learning models predicted the same class in level A. For the level A, as can be seen in Table 9, when training with MOLD+SeMOLD, the results did not improve for the transformer models.

This is similar to the previous experiments in data augmentation (Rosenthal et al. 2021) where the results do not improve when the machine learning classifier is already strong. We can assume that the transformer models are already well trained for MOLD and adding further instances to the training process would not improve the results for the transformer models.

6.2 Categorization of offensive language

For the level B, the MOLD training set is smaller, and the task is also more complex than the level A. Therefore, the machine learning models can benefit from adding more data. As can be seen in Table 10, all of the transformer models improve with data augmentation from SeMOLD instances. IndicBERT model performed best with the data augmentation and provided 0.76 Macro-*F*1 score.

6.3 Offensive language target identification

Finally for level C, the manually annotated OLID dataset is even smaller, and the number of classes increases from two to three. As can be seen in Table 11, all the models improve with the data augmentation process. IndicBERT model performed best after the data augmentation process scoring 0.68 Macro-*F*1 score.

7 Conclusion and future work

We presented a comprehensive evaluation of Marathi offensive language identification along with two new resources: MOLD 2.0 and SeMOLD. MOLD 2.0 contains over 3600 tweets annotated with OLID's three-level annotation taxonomy making it the largest manually annotated Marathi offensive language dataset to date. SeMOLD is a larger dataset of 8,000 instances annotated with semi-supervised methods. Both these results open exciting new avenues for research on Marathi and other low-resource languages.

Our results show that it is possible to identify types and targets of offensive posts in Marathi with a relatively small size dataset (answering **RQ1**). With respect to **RQ2**, we report that (2) the use of the larger dataset (SeMOLD) combined with MOLD 2.0 results in performance improvement particularly for levels B and C where less data are available in MOLD (answering **RQ2.1**); and (2) transfer learning techniques from both English and Hindi result in performance improvement for Marathi in the three tasks (identification, categorization, and target identification) (answering **RQ2.2**). We believe that these results shed light on offensive language identification applied to Marathi and low-resource languages as well, particular Indo-Aryan languages.

In future work, we would like to extend MOLD's annotation to a fine-grained token-level annotation. This would allow us to jointly model both instance label and token annotation as in MUDES (Ranasinghe et al. 2021). Finally, we would like to use the knowledge and data obtained with our work on Marathi and expand it to closely related Indo-Aryan languages such as Konkani.

References

Alakrot A, Murray L, Nikolov NS (2018) Towards accurate detection of offensive language in online communication in arabic. Procedia Comput Sci 142:315–320

Aroyehun ST, Gelbukh A (2018) Aggression detection in social media: using deep neural networks, data augmentation, and pseudo labeling. In: Proceedings of TRAC

Basile V, Bosco C, Fersini E, Nozza D, Patti V, Pardo FMR, Rosso P, Sanguinetti M (2019) Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In: Proceedings of SemEval

Bassignana E, Basile V, Patti V (2018) Hurtlex: a multilingual lexicon of words to hurt. In: Proceedings of CliC-It

Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Trans Assoc Comput Linguist 5:1

Carletta J (1996) Assessing agreement on classification tasks: the kappa statistic. Comput Linguist 22(2):249–254

Chiril P, Benamara Zitoune F, Moriceau V, Coulomb-Gully M, Kumar A (2019) Multilingual and multitarget hate speech detection in tweets. In: Proceedings of TALN

Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V (2019) Unsupervised cross-lingual representation learning at scale. In: Proceedings of ACL

Çöltekin c (2020) A Corpus of Turkish Offensive Language on Social Media. In: Proceedings of LREC

Dadvar M, Trieschnigg D, Ordelman R, de Jong F (2013) Improving dyberbullying detection with user context. In: Proceedings of ECIR,

Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL

Fišer D, Erjavec T, Ljubešić N (2017) Legal framework, dataset and annotation schema for socially unacceptable on-line discourse practices in Slovene. In: Proceedings ALW

Fortuna P, da Silva JR, Wanner L, Nunes S, et al (2019) A hierarchically-labeled portuguese hate speech dataset. In: Proceedings of ALW

Gaikwad SS, Ranasinghe T, Zampieri M, Homan C (2021) Cross-lingual offensive language identification for low resource languages: the case of Marathi. In: Proceedings of RANLP

Ghadery E, Moens M-F (2020) LIIR at semeval-2020 task 12: a cross-lingual augmentation approach for multilingual offensive language identification. Proceedings of SemEval

Goudjil M, Koudil M, Bedda M, Ghoggali N (2018) A novel active learning method using svm for text classification. Int J Autom Comput 15(3):290–298

Hettiarachchi H, Ranasinghe T (2019) Emoji powered capsule network to detect type and target of offensive posts in social media. In: Proceedings of RANLP

Kakwani D, Kunchukuttan A, Golla S, NC G, Bhattacharyya A, Khapra MM, Kumar P (2020) IndicNLPSuite: monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In: Findings of the Association for Computational Linguistics: EMNLP 2020

Kim Y (2014) Convolutional neural networks for sentence classification. In: Proceedings of EMNLP

Kumar R, Ojha AK, Malmasi S, Zampieri M (2020) Evaluating aggression identification in social media. In: Proceedings of TRAC

Kumar R, Ojha AK, Malmasi S, Zampieri M (2018) Benchmarking aggression identification in social media. In: Proceedings of TRAC

Kumar S, Kumar S, Kanojia D, Bhattacharyya,P (2020) A passage to India: Pre-trained word embeddings for Indian languages. In: Proceedings of SLTU

Liu P, Li, W, Zou L (2019) NULI at SemEval-2019 task 6: transfer learning for offensive language detection using bidirectional transformers. In: Proceedings of SemEval

Malmasi S, Zampieri M (2017) Detecting hate speech in social media. In: Proceedings of RANLP

Mandl T, Modha S, Majumder P, Patel D, Dave M, Mandlia C, Patel, A (2019) Overview of the Hasoc track at fire 2019: hate speech and offensive content identification in Indo-European languages. In: Proceedings of FIRE

Mandl T, Modha S, Kumar M A, Chakravarthi BR (2020) Overview of the hasoc track at fire 2020: hate speech and offensive language identification in Tamil, Malayalam, Hindi, English and German. In: Proceedings of FIRE

Modha S, Mandl T, Shahi GK, Madhu H, Satapara S, Ranasinghe T, Zampieri M (2021) Overview of the HASOC Subtrack at FIRE 2021: hate speech and offensive content identification in English and Indo-Aryan languages and conversational hate speech. In: Proceedings of FIRE

Mubarak H, Rashed A, Darwish K, Samih Y, Abdelali A (2021) Arabic offensive language on twitter: analysis and experiments. In: Proceedings of WANLP

Pamungkas, EW, Patti V (2019) Cross-domain and cross-lingual abusive language detection: a hybrid approach with deep learning and a multilingual lexicon. In: Proceedings ACL:SRW

Pitenis Z, Zampieri M, Ranasinghe T (2020) Offensive language identification in Greek. In: Proceedings of LREC

Poletto F, Stranisci M, Sanguinetti M, Patti V, Bosco C (2017) Hate speech annotation: analysis of an Italian twitter corpus. In: Proceedings of CLiC-it

Ranasinghe T, Zampieri M (2021) An evaluation of multilingual offensive language identification methods for the languages of india. Information 12(8):1

Ranasinghe T, Zampieri M (2020) Multilingual offensive language identification with cross-lingual embeddings. In: Proceedings of EMNLP

Ranasinghe T, Zampieri M (2021) Multilingual offensive language identification for low-resource languages. ACM transactions on asian and low-resource language information processing (TALLIP)

Ranasinghe T, Zampieri M (2021) MUDES: multilingual detection of offensive spans. In: Proceedings of NAACL

Ranasinghe T, Hettiarachchi H (2020) BRUMS at SemEval-2020 task 12: transformer based multilingual offensive language identification in social media. In: Proceedings of SemEval

Ranasinghe T, Sarkar D, Zampieri M, Ororbia A (2021) WLV-RIT at SemEval-2021 task 5: a neural transformer framework for detecting toxic spans. In: Proceedings of SemEval

Ridenhour M, Bagavathi A, Raisi E, Krishnan S (2020) Detecting online hate speech: approaches using weak supervision and network embedding models. arXiv preprint arXiv:2007.12724

Rosenthal S, Atanasova P, Karadzhov G, Zampieri M, Nakov P(2021) Solid: a large-scale semi-supervised dataset for offensive language identification. In: Findings of ACL

Sarkar D, Zampieri M, Ranasinghe T, Ororbia A (2021) fbert: a neural transformer for identifying offensive content. In: Findings of the association for computational linguistics: EMNLP 2021, pp 1792– 1798

Schwarm SE, Ostendorf M (2005) Reading level assessment using support vector machines and statistical language models. In: Proceedings of ACL

Tulkens S, Hilte L, Lodewyckx E, Verhoeven B, Daelemans W (2016) A dictionary-based approach to racism detection in Dutch Social Media. In: Proceedings of TA-COS

Wiegand M, Siegel M, Ruppenhofer J (2018) Overview of the GermEval 2018 shared task on the identification of offensive language. In: Proceedings of GermEval

Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac,P, Rault T, Louf R, Funtowicz M, Davison J, Shleifer S, von Platen P, Ma C, Jernite Y, Plu J, Xu C, Le Scao T, Gugger S, Drame M, Lhoest Q, Rush A (2020) Transformers: state-of-the-art natural language processing. In: Proceedings of EMNLP

Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV (2019) XLNet: generalized autoregressive pretraining for language understanding. In: Proceedings of NeurIPS

Yao M, Chelmis C, Zois D-S (2019)Cyberbullying ends here: towards robust detection of cyberbullying in social media. In: Proceedings of WWW

Zampieri M, Malmasi S, Nakov P, Rosenthal S, Farra N, Kumar R (2019) Predicting the type and target of offensive posts in social media. In: Proceedings of NAACL

Zampieri M, Nakov P, Rosenthal S, Atanasova P, Karadzhov G, Mubarak H, Derczynski L, Pitenis Z, Çöltekin C (2020) SemEval-2020 Task 12: multilingual offensive language identification in social media (OffensEval 2020). In: Proceedings of SemEval

Zhang J, Chang J, Danescu-Niculescu-Mizil C, Dixon L, Hua Y, Taraborelli D, Thain N (2018) Conversations gone awry: detecting early signs of conversational failure. In: Proceedings of ACL