A TRANSFORMER BASED MULTI-TASK LEARNING APPROACH LEVERAGING TRANSLATED AND TRANSLITERATED DATA TO HATE SPEECH DETECTION IN HINDI

Prashant Kapil and Asif Ekbal

Department of Computer Science and Engineering, IIT Patna, India

ABSTRACT

The increase in usage of the internet has also led to an increase in unsocial activities, hate speech is one of them. The increase in Hate speech over a few years has been one of the biggest problems and automated techniques need to be developed to detect it. This paper aims to use the eight publicly available Hindi datasets and explore different deep neural network techniques to detect aggression, hate, abuse, etc. We experimented on multilingual-bidirectional encoder representations from the transformer (M-BERT) and multilingual representations for Indian languages (MuRIL) in four settings (i) Single task learning (STL) framework. (ii) Transfering the encoder knowledge to the recurrent neural network (RNN). (iii) Multi-task learning (MTL) where eight Hindi datasets were jointly trained and (iv) pre-training the encoder with translated English tweets to Devanagari script and the same Devanagari scripts transliterated to romanized Hindi tweets and then fine-tuning it in MTL fashion. Experimental evaluation shows that cross-lingual information in MTL helps in improving the performance of all the datasets by a significant margin, hence outperforming the state-of-the-art approaches in terms of weighted-F1 score. Qualitative and quantitative error analysis is also done to show the effects of the proposed approach.

KEYWORDS

M-BERT, MuRIL, Weighted-F1, RNN, cross-lingual.

1. INTRODUCTION

The emergence of social media platforms like Facebook and Twitter has led to an exponential increase in user-generated content. The identification of hate speech within a large volume of posts on social media has posed a challenge and thus is a growing research area. There is a growing need to develop an automated classifier to detect different forms of hate speech such as offensive, profanity, abusive, and aggression that are prevalent on different social media platforms. The offensive posts which create social disability need to be restricted alongside maintaining the right to freedom of speech.

These incidents create mental and psychological agony for the users resulting in deactivating the account or in some cases committing suicide **[1].** While research in this area is gaining momentum, there is a lack of research in the Hindi language. In multilingual societies like India usage of code-mixed languages is common for conveying any opinion. Code-mixing is a phenomenon of embedding linguistic units such as phrases, words, or morphemes of one language into an utterance of another **[2].** In social media, Hindi posts are generally present in

David C. Wyld et al. (Eds): ARIA, SIPR, SOFEA, CSEN, DSML, NLP, EDTECH, NCWC - 2022
pp. 191-207, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121516

either Devanagari script or Hindi-English code mixed pattern. To build an efficient classifier supervised learning on the labeled dataset is the most common approach. In India, native vernacular languages are spoken by a majority of the population. Mixed code language like Hinglish is most prevalent in social media conversations. **[3]** reported in 2015 that India ranked fourth in the social hostilities index with an index value of 8.7 out of 10., indicating the need to solve this problem of hate speech. Code-switched language presents challenges of randomized spelling variations in explicit words due to foreign script and ambiguity arising due to various interpretations of words in different contextual situations **[4]**. The detection of hate speech is very important for lawmakers and social media platforms to curb any wrong activity. Table 1 consists of the definition followed to collect the different sub class of hate. Table 2 enlists the laws on hate speech in some of the countries.

The significant contributions of this work are as follows:

Dataset: We utilized eight benchmark datasets related to the hate domain, aggressiveness, offensiveness, abuse, etc. To add cross-lingual information we also translated eleven English data to Devanagari script by leveraging Google Translate. The Devanagari tweets were also transliterated to the Roman script by using Indic-trans **[19].**

Model: We investigated the various state-of-the-art models such as M-BERT and MuRIL to design eight models. The first set of model is based on single task learning paradigm. The knowledge from the transformer encoder is transferred to the bidirectional long short term memory (Bilstm) in our second set of models. The third set of model is based on MTL paradigm and in the fourth set the encoder is first pre-trained with translated and transliterated data followed by leveraging the MTL on eight data.

Error Analysis: The results and errors on the experimented models were analyzed by presenting qualitative and quantitative analysis to highlight some of the errors that need to be rectified to improve the system performance.

The remaining structure of this paper is as follows.

A brief overview of the related background literature is presented in Section 2. In Section 3, the datasets used for the experiments are discussed. Section 4 discusses in detail the proposed methodology, experimental setup. Section 5 reports the evaluation results and comparisons to the state-of-the-art, and Error analysis containing qualitative and quantitative analysis of the obtained results. Finally, the conclusion and directions for future research are presented in Section 6.

Table 1. Definition of hate speech

Authors	**Definition**
[26]	The post contains hate, offensive, or profane content.
[25]	The posts contain covertly and overtly aggressive messages.
[4]	The tweets were labeled as hate speech if they satisfied one or more of the conditions: (i) tweet using sexist or racial slur to target a minority, (ii) undignified stereotyping or (iii) supporting a problematic hashtags such as #ReligiousSc*m.
[7]	It is a bias-motivated hostile speech aimed at a person or group of people with intentions to injure, dehumanize, harass, degrade and victimize targeted groups based on some innate characteristics.
[8]	It is defined as abusive speech containing a high frequency of stereotypical words.

Table 2: Laws of different countries on hate speech

Country	Law
USA	Hate speech is legally protected free speech under the First Amendment. However, speech that includes obscenity, speech integral to illegal conduct, and speech that incites lawless action or is likely to produce such activity are given lesser or no protection.
Brazil	According to the 1988 Brazilian constitution racism is an offense with no statute of limitations and no right to bail for the defendant.
Germany	Section 130 of the German criminal code states incitement to hatred is a punishable offense leading up to 5 years imprisonment. It also states that publicly inciting hate against some parts of the population or using insulting malicious slurs or defaming to violate their human dignity is a crime.
India	Article 19(1) of the constitution of India protects the freedom of speech and expression. However, article 19(2) states that to protect sovereignty, integrity, and security of the state, to protect decency and morality, defamation and incitement to an event, some restrictions can be imposed
Japan	The Hate speech act of 2016 does not apply to groups of people but covers threats and slander to protect.
New Zealand	Their Hate speech act follows Section 61 of the Human Rights Act 1993 that asserts that threatening, abusive content in any form, words that are likely to create hostility against a group of people based on race, color, or ethnicity is unlawful.

2. RELATED WORK

2.1. Hate speech detection in low resource languages

This section summarizes the works done on hate speech detection for low-resource languages.

Arabic: **[5]** investigated the religious hate speech detection on 6000 labeled data in Arabic from Twitter. They created and published three lexicons of religious hate terms. They investigated three different approaches namely lexicon-based, n-grams-based, and gated-recurrent unit (GRU) -based neural networks with word embeddings provided by AraVec **[6]**. **[7]** presented 3353 Arabic tweets tagged for five classification tasks. They analyzed the difficulties of collecting and annotating the Arabic data and determined 16 target groups like women, gay, Asians, Africans, immigrants, refugees, etc. The experiments showed that deep learning settings outperform the BOW (Bag of words) based method in all five tasks. **[8]** introduced the first Levantine Hate speech and abusive (L-HSAB) data comprising 5846 tweets tagged into three categories: normal, hate, and abusive. The results indicated the outperformance of naive Bayes (NB) over support vector machines (SVM) in both binary and multi-class classification experiments.

French: **[9]** described CONAN: as the first large-scale, multilingual, and expert-based hate speech/counter-narrative dataset for English, French and Italian. The data consist of other meta-data features such as expert demographics, hate speech sub-topic, and counter-narrative type. **[7]** created a hate speech dataset in English, French, and Arabic annotated for the five classification tasks: the directness of the speech, the hostility type of the tweet, the discriminating target attribute, the target group, and the annotator's sentiment.

German: **[10]** developed a dataset containing offensive posts by including their target. They implemented a two-step approach to detect the offending statements. The first step is a binary classification between offensive and not offensive. The second step classifies offensive into severity = 1 and severity =2. **[11]** released the pilot edition of the GermEval shared task on the Identification of Offensive Language comprising 8000 posts annotated for two layers. The first layer is the coarse-grained binary classification between offensive and other. The second layer is fine-grained 4-way tagging of the offensive post between profanity, abuse, insult, and others. The popular features leveraged to solve the task were word embeddings, character n-grams, and lexicons of offensive words.

Italian: **[12]** created an Italian twitter corpus of 6000 tweets annotated for hate speech against immigrants and designed a multi-layer annotation scheme to annotate the post's intensity, aggressiveness, offensiveness, irony, and stereotypes. **[13]** proposed a shared task to solve the Hate Speech detection (HaSpeeDe) on Italian Twitter and Facebook. The teams utilized traditional machine learning approaches, such as support vector machine (SVM), logistic regression (LR), random forest (RF), and deep learning techniques such as convolution neural network (CNN), gated recurrent unit (GRU), and multi-layer perceptron (MLP), etc. The results also confirmed the difficulty of cross-platform hate speech detection.

There is little work done for other low-resource languages, which include Spanish (**[14]**,**[38]**), Polish **[15]**, Portuguese **[45]**, Slovene **[16]**, Turkish **[17]** and Indonesian **[18]**.

2.2. Hate speech classification in Hindi

There has been little effort to solve the Hate speech detection in a low-resource language such as Hindi due to the scarcity of labeled data. The cost of generating labeled data is often time-consuming and tedious, limiting the further development of machine learning approaches. In recent years, shared tasks have been organized for low-resource languages, such as Hindi to solve the task of aggressive identification or hate classification. **[20]** released 15000 aggression annotated Facebook posts and comments in Hindi (Roman and Devanagari script). **[21]** conducted experiments with deep neural network models of varying complexity ranging from CNN, LSTM, BiLSTM, CNN-LSTM, LSTM-CNN, CNN-BiLSTM, and BiLSTM-CNN. To improve over the baseline, they also utilized data augmentation, pseudo labeling, and sentiment score as the feature. **[22]** explored the combination of passive-aggressive (PA) and SVM classifiers with character-based n-gram (1-5) TF-IDF for the feature representations. **[23]** uses LSTM, and CNN initialized with fast text word embeddings, and **[24]** uses BiLSTM with glove embeddings to solve the problem.

In recent times multi-layer annotated data to cover the different facets of a post has been released. **[25]** presented a shared task featuring two tasks: first is aggression identification to discriminate overtly, covertly, and non-aggressive posts and the second is gendered aggression identification. The approaches used by different teams were mostly based on neural networks such as CNN, LSTM, and BiLSTM initialized with word embeddings. The utility of M-BERT, XLM-RoBERTa, DistilRoBERTa, and transfer learning techniques based on universal sentence encoder (USE) embedding were also explored to solve the task. **[26]** and **[27]** developed 2 layer annotated data. The first is classified between Hate and Offensive (HOF) and non-hate (NOT). The second task is a fine-grained classification of HOF into hate, offensive, and profanity. **[28]** pre-trained the word vectors by 0.5 million in-domain unlabeled data to obtain task-specific embeddings. This knowledge is then transferred to CNN for classification. They observed that CNN outperforms LSTM when transfer learning through word vectors is utilized. **[29]** released the DHOT dataset in Devanagari script and developed a classifier based on FastText embeddings to classify offensive and non-offensive tweets. **[30]** explored IndicBERT, RoBERTa Hindi, and neural space BERT Hindi to solve the binary classification between Hate and Offensive (HOF) and NOT. **[31]** proposed to enhance the hate speech detection of code mixed Hind-English by incorporating social media-based features along with capturing profanity features into the model. They also proposed a novel bias elimination algorithm to mitigate any bias from the model. **[32]** experimented with two architectures, namely the sub-word level LSTM model and hierarchical

LSTM model with attention, based on phonemic sub-words for hate speech detection on social media code-mixed text. **[2]** presented an annotated corpus of 4575 tweets in Hindi-English code mixed text. To build the classification system, they utilized features such as character n-grams, punctuations, negation words, word n-grams, punctuations, negation words, and a hate lexicon.

3. Data Sets

In this section, we will briefly describe all 8 Hindi datasets related to the hate domain used in this paper. The statistics of all the hate-related data are in Table 3.

Data 1 (D1) [29]: A lexicon of abusive words in Hindi were built. The 20 abusive terms collected serve as keywords that were assigned to a data acquisition program. The tweets were also mined from popular Twitter hashtags of viral topics, and popular public figures like politicians, sports personalities, and movie actors. The annotation of DHOT tweets is done by three language experts. The average value of cohen kappa for the inter-annotator agreement is 84%.

Data 2 (D2) [26]: The authors followed the heuristics approach to search for hate speech in an online forum by identifying the topics for which hate speech can be expected. Different hashtags and keywords were used to sample the posts from Twitter and Facebook. The inter-annotator agreement score obtained is 36%.

Data 3 (D3) [27]: The sampling of the dataset was done during the extremely hard COVID-19 second wave in India. Therefore during the sampling process, major topics in social media are influenced by COVID-19. To obtain potential hateful tweets, a weak classifier based on an SVM classifier with n-grams features to predict weak labels on the unlabeled corpus is used. The trending hashtags used to sample the tweets were *#resignmodi*, *#TMCTerror*, *#chinesevirus*, *#islamophobia*, *#covidvaccine*, *#IndiaCovidcrisi s*, etc. The inter-annotator agreement score is 69%.

Data 4 (D4) [25]: The data is crawled from the public Facebook pages and Twitter. For Facebook, more than 40 pages were crawled which included news websites, web-based forums, political parties, student organizations, etc. For Twitter, the data was collected using some of the popular hashtags such as beef ban, election results, etc. The complete dataset contains 18K tweets and 21K Facebook comments annotated with aggression and discursive effects. The inter-annotator agreement for the top level is 72%.

Data 5 (D5) [46]: The dataset is collected from various social media platforms namely Facebook, Twitter, and Youtube. The actual sources of information ranged from public posts, tweets, videos, news coverage, etc. The annotation of data involves multiple human interventions and constant deliberations over the justification of assigned tags.

Data 6 (D6) [33]: They collected posts from various social media platforms like Twitter, Facebook, Whatsapp, etc. To collect hate speech data, the tweets encouraging violence against minorities based on race and religious beliefs were sampled. The timeline of the users with significant hate-related posts was also analyzed. The offensive posts are crawled by Twitter search API by employing the list of swear words used in the Hindi language released by **[29]**. The posts related to the defamation category are collected from viral news articles where people or a group are publicly shamed due to misinformation. The topic-wise search is performed to collect defamation tweets.

Data 7 (D7) [4]: The tweets were mined from popular Twitter hashtags of viral topics across the news feed. The tweets were collected from the Twitter handles of sportspersons, political figures, news channels, and movie stars. The annotation of tweets was done by three annotators having a background in NLP research. The Cohen kappa inter-annotator agreement score is 83%.

Data 8 (D1) **[2]**: presented an annotated corpus of 4575 tweets in Hindi-English code mixed text. To build the classification system, they utilized features such as character n-grams, punctuations, negation words, word n-grams, and a hate lexicon. The Kappa score is 98.20%.

Table 3. Statistics of dataset

Datasets	**Labels and train/test set**
D1	HOF: 403/81, NOT: 1200/316
D2	HOF: 2469/605, NOT: 2196/713
D3	HOF: 1433/483, NOT: 3161/798
D4	OAG: 6072/362, CAG: 6115/413 NAG: 2813/195
D5	OAG: 1118/669, CAG: 1040/215 NAG: 2823/316
D6	Hostile: 3054/780, Non-Hostile:3485/873
D7	Abusive: 1765 , Hate: 303 Neutral: 1121
D8	Hate: 1299, Neutral: 2249

Table 4: Translated and Transliterated sample

S1	We dont trust these n****s all these bitch.
Translation	हम इन सभी काले लोगों पर भरोसा नहीं करते हैं
Transliteration	ham in sabhi kutiya par bharosa nahi karte.
S2	your grammar is trash.
Translation	आपका व्याकरण कचरा है
Transliteration	aapka vyakaran kachra hai.
S3	you are irrelevent b***h.
Translation	तुम्हारे तुम अप्रासंगिक हो
Transliteration	aap aprasangik kutiya hai.

3.1 Cross-lingual Data

As there are abundant data available for English, we aim to determine if knowledge from one language can be used to improve the performance of another language. We utilized eleven English data **[34]**, **[35]**, **[20]**, **[36]**, **[37]**, **[26]**, **[27]**, **[38]**, **[39]**, **[40]**, and **[41]**.

Translated Data: The Google translate API is used to translate approximately 2,50,000 tweets to the Devanagari script. We have selected 100 random samples to analyze the translation of the original post. The human evaluation found the translation to be satisfactory.

Transliterated Data: After obtaining the translated Devanagari posts, it is transliterated to Romanized form by Indic-trans **[19]**. Table 4 consists of some instances of translated and transliteration.

4. METHODOLOGY

4.1. Pre-Processing

Social media posts contain a lot of noisy text which is not considered a useful feature for the classification. We perform the following steps to remove the noise, and make it ready for experiments:

1. Words are reduced to lower case so that words such as "BI**H", "bi**h" and "Bi**h" will have the same syntax and will utilize the same pre-trained embedding values.
2. Word segmentation is being done using the Python-based word segment to preserve the important features present in hashtag mentions.
3. All the emoticons were categorized into 5 categories, namely प्रेम *love*, *दुख sad*, **खुशी** *happy*, *आश्चर्यshocking*, and *गुस्सा anger*. The Unicode character of the emoticon in the text is substituted with one category.
4. All the @ (ex.@abc) mentions were replaced with the common token, i.e *user*.
5. The stop words were not removed due to the risk of losing some useful information, and this was also empirically found to be of little or no impact on the classification performance after removing them.
6. The maximum sequence length is set to 40. Post padding is done if any sentence is less than 40 and pruning is performed from the last if the sentence is greater than 40.

We experimented on 8 transformer-based approaches which are discussed in this section.

4.2. Models

Model 1(M1): **Multilingual-BERT**: **[42]** introduced {**M-BERT**} i.e Multilingual Bidirectional Encoder Representation from Transformers to pre-train deep bidirectional representations from unlabeled texts by jointly conditioning on both left and right contexts in all layers. There are two steps in the training framework: pre-training and fine-tuning. During pre-training, the model is trained on unlabeled data over different pre-training tasks. For fine-tuning, the BERT model is first initialized with the pre-trained parameters and all of the parameters are fine-tuned using labeled data from the downstream tasks. It follows two training objectives which are described as follows:

Masked language modeling (MLM): The model randomly masks 15% of the tokens from the input, and the objective is to predict the masked words based only on their context. The training data generator chooses 15% of the token positions at random for prediction. If the ith token is chosen it is replaced with (1) the [MASK] token 80% of the time (2) a random token 10% of the time (3) the unchanged ith token 10% of the time.

Next Sentence prediction (NSP): It jointly pre-trains text-pair representations, and the model is to predict whether two sentences are following each other or not.

The multi-lingual version of the BERT is capable of working with 104 languages. The first token of every sequence starts with a unique classification token ([CLS]. The final hidden state corresponding to this token is used as the aggregate sequence representation for the classification task.

Model 2(M2): **MuRIL [43]**: It is a multilingual language model specifically developed for the Indian languages by training on IN text corpora of 16 Indian languages. It utilizes two training objectives: MLM and Translation language modeling (TLM). The MLM uses monolingual text only (unsupervised), and TLM uses translated and transliterated document pairs to train the model. The maximum sequence length is 512, global batch size of 4096, and trained for 1M steps. The total trained parameters are 236M that is optimized by Adam optimizer with the learning rate of 5e-4. The general architecture of transformer encoder block is shown in Figure 1

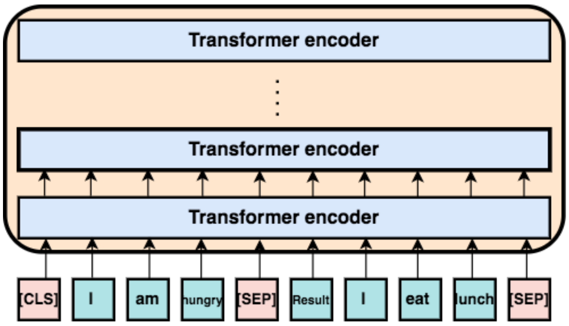

Figure 1. Transformer encoder

4.2.1. Knowledge Transfer

[42] compared different combinations of layers of BERT to conclude that the output of the last four layers combined encodes more information than only the last layer. In this work, we utilize the last 4 hidden layers output from pre-trained M-BERT and MuRIL models into Bilstm followed by the softmax activation function. Figure 2 shows the architecture.

Model 3 (M3): M-BERT-Bilstm: The concatenation of the last 4 hidden layers was passed into Bilstm.

Model 4 (M4): MuRIL-Bilstm: The concatenation of the last 4 hidden layers was passed into Bilstm.

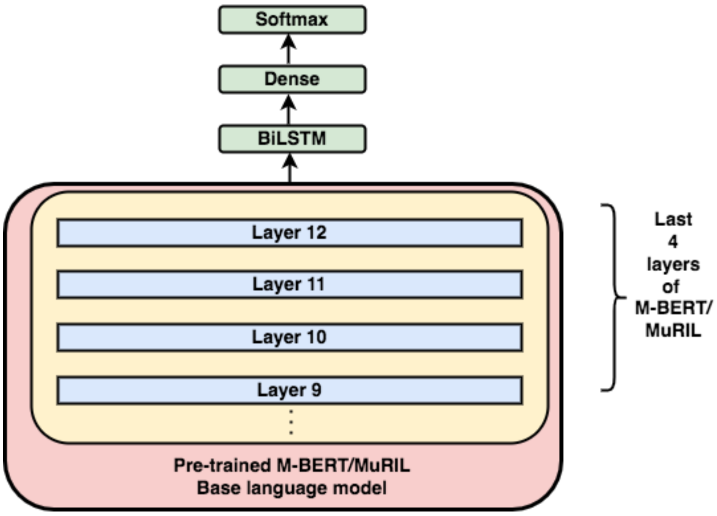

Figure 2. M-BERT/MuRIL-Bilstm architecture

4.2.2. Multi task learning (MTL)

Multi-tasking learning aims at solving more than one problem simultaneously. End-to-end deep multi-task learning has been recently employed in solving various problems of natural language processing (NLP). It enables the model by sharing representations between the related tasks and generalizing better by achieving better performance for the individual tasks.

[47] developed two forms of MTL, namely Symmetric multi-task learning (SMTL) and Asymmetric multi-task learning (AMTL). The former is joint learning of multiple classification tasks, which may differ in data distribution due to temporal, geographical, or other variations, and the latter refers to the transfer of learned features to a new task to improve the new task's learning performance.

[48] discussed the two most commonly used ways to perform multi-task in deep neural networks.

(i) **Hard Parameter Sharing**: Sharing the hidden layers between all tasks with several task-specific output layers.

(ii) **Soft Parameter Sharing**: Each task has its specific layers with some sharable parts.

Model 5(M5): This model leverages the M-BERT trained in the MTL paradigm.

Model 6(M6): This model leverages the MuRIL trained in the MTL paradigm.

The architecture of the MTL-DNN is shown in Figure 3. The lower layers are shared across all the tasks, while the top layers represent task-specific outputs. In our experiment, all the tasks are classified. The input X is a word sequence (either a sentence or a pair of sentences packed together) represented as a sequence of embedding vectors, one for each word in l1. Then the transformer encoder captures the contextual information for each word via self-attention and generates a sequence of contextual embedding in l2. The shared semantic representation is trained by the multi-task objectives. In the following, we will describe the model in detail.

Lexicon Encoder (l1): The input X = {*x1,x2,....xm*} is a sequence of tokens of length *m*. Following **[42]** the first token x1 is always the {CLS} token. If X is packed by a sentence pair (X1, X2), we separate the two sentences with a special token [SEP]. The lexicon encoder maps X into a sequence of input embedding vectors, one for each token, constructed by summing the corresponding word, segment, and positional embeddings.

Transformer Encoder (l2): It consists of a multi-layer bidirectional Transformer encoder **[49]** to map the input representation vectors (l1) into a sequence of contextual embedding vectors *C* belongs to *R(d*m)*. This will be the shared representation across different tasks. MT-DNN learns the representation using multi-task objectives, in addition to pre-training.

Single-Sentence Classification Output: Suppose that x is the contextual embedding (l2) of the token [CLS] that can be viewed as the semantic representation of input sentence X. The probability that X is labeled as class c is predicted with softmax.

$$P_r(c|X) = \text{softmax}(\mathbf{W}_{SST}^{\top} \cdot \mathbf{x}), \qquad (1)$$

The training procedure of MT-DNN consists of two stages: pre-training and multi-task learning.

In the multi-task learning stage, mini-batch-based stochastic gradient descent (SGD) is used to learn the parameters of our model. In each epoch, a mini-batch bi is selected among all the tasks

For the classification tasks, the loss function used is categorical cross-entropy loss.

$$-\sum_{c} \mathbb{1}(X, c) \log(P_r(c|X)), \tag{2}$$

Where 1(X, c) is the binary indicator (0 or 1) if class label c is the correct classification for X

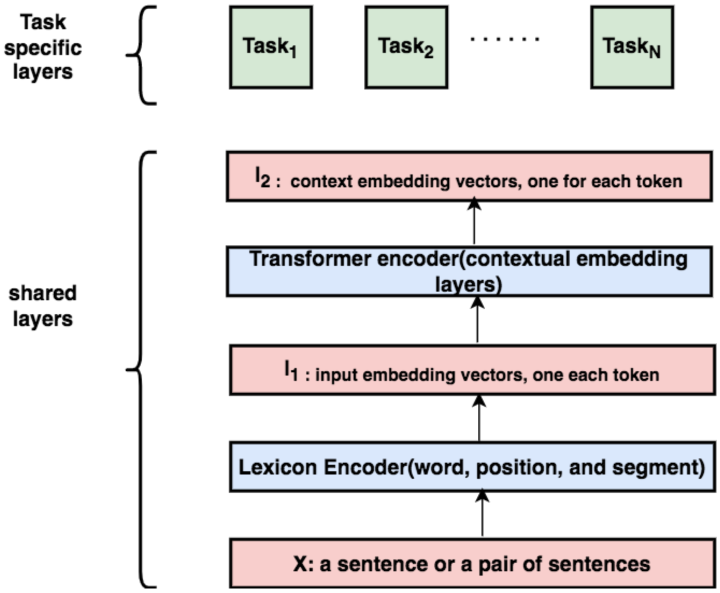
Task specific layers
Task_1
Task_2
......
Task_N
shared layers
l_2 : context embedding vectors, one for each token
Transformer encoder(contextual embedding layers)
l_1 : input embedding vectors, one each token
Lexicon Encoder(word, position, and segment)
X: a sentence or a pair of sentences

Figure 3. Multi task learning architecture with BERT/ALBERT as shared encoder

4.2.3. Pre-training with Cross lingual Information

.**Step 1**: We utilize the 250K translated data and 250K transliterated data to pre-train the M-BERT and MuRIL.

Step 2: The trained parameters is used to initialize the weight of the shared encoder.
Step 3: The same procedure as of multi task learning is followed as in Figure 3.

Model 7 and **Model 8** utilizes the cross lingual information.

4.2.4. Experimental Setup

All the deep learning models were implemented using Keras, a neural network package **[50]** with Tensorflow **[51]** as the backend. Each dataset is split into an 80:20 ratio to use 80% in grid-search to tune the batch size and learning epochs using 5-fold cross-validation experiments and test the optimized model on 20% held-out data. The results are the mean of 5 runs with the same setup. For some data with a separate test set, the model is trained on train data, and performance is evaluated using test data. Categorical cross-entropy is used as a loss function, and Adam **[52]** optimizer is used for optimizing the network.

We use a learning rate of 2e-5 for the transformer models. The batch size of 30 is used to train the shared encoder and an epoch of 2 is found to be optimal. The value for bias is randomly initialized to all zeros, the relu activation function is employed at the intermediate layer, and Softmax is utilized at the last dense layer. The transformers library is loaded from Hugging Face. It is a python library providing a pre-trained and configurable transformer model useful for various NLP tasks.

5. RESULTS, COMPARISON AND ANALYSIS

We report the weighted-F1 score of all the eight datasets in Table 5. Table 6 enlists comparison with the state-of-the-art approaches and the proposed approach over the weighted-F1 score. From the results it can be seen that pre-training with the translated and transliterated data followed by training in MTL outperformed the other methods. We are also presenting the statistical significance results between best and worst model in Table 7. We also did some qualitative analysis and presented different patterns of hate posts detected by the best model.

Table 5. Weighted-F1 scores of eight data sets

Data	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**
D1	91.43	94.41	92.94	94.53	94.94	94.89	**95.19**	94.99
D2	74.47	81.64	77.11	82.57	79.55	82.27	**82.78**	82.39
D3	77.09	80.41	79.14	82.12	82.82	81.23	**83.11**	82.94
D4	61.80	59.80	62.67	61.80	65.63	63.89	**65.97**	64.23
D5	76.90	74.50	80.51	81.38	80.98	81.96	81.22	**82.14**
D6	82.32	80.94	83.45	81.28	85.98	83.65	**86.14**	83.78
D7	84.98	85.85	85.97	88.27	90.16	88.10	**90.98**	88.67
D8	89.10	89.10	89.58	90.51	90.10	91.78	90.86	**92.17**

5.1. Qualitative Analysis

In this section we are giving four types of hate posts with the explanation which were correctly classified by the model.

HATE IS TOXIC

GROUND VALUE: **HATE**
PREDICTED CLASS: **HATE**

1. अपनी औकात भूल गए हो तुम कुत्ते सुवर की औलाद
   TRANSLITERATION: Apni aukaat bhool gaye ho tum suwar kii aaulaad.
   TRANSLATION: You have forgotten your real worth. You son of a pig.
2. पता लगा बे हराम कौन ट्रेंड कर रहा है
   TRANSLITERATION: Pata laga be haram kaun trend kar raha hai.
   TRANSLATION: Find out you scoundrel, who the hell is trending.

EXPLANABILITY: Both the tweets consists of slang term such as s***r , and h***m. As the training data consists of large number of tweets containing these terms it detected it successfully.

.

INDIRECT REFERENCES

GROUND TAG: HATE
PREDICTED CLASS: **HATE**

1. Kaun rapper aachha gaata hai. I hate all. Bas music kaa kachara karne aaye hai sab
   TRANSLITERATION: Kaun rapper aachha gaata hai. I hate all. Bas music kaa kachara karne aaye hai sab.
   TRANSLATION : No rapper is good enough, I hate all of them as they are just making the trash of music.
2. आखिर कब तक जनता उठाएगी निकम्मे कर्मचारियों का बोझ

TRANSLITERATION: Aakhir kab tak janta uthaegii nikamme karmachariyon kaa bojha.

TRANSLATION: After all, how long will the public bear the burden of the useless employees.

EXPLANABILITY: Here Indirect attack in a softer tone is being done which the model is able to detect.

CONTEXTUAL INFORMATION

GROUND TAG: HATE
PREDICTED CLASS: **HATE**

1. जो भी हो मुझे भी लगता है । दाल में कुछ काला
   TRANSLITERATION :Jo bhi ho mujhe bhi lagta hai, daal me kuch kaala.
   TRANSLATION: Whatever, even I think there is something fishy.

2. @INCINDIA ISHLIYE CORRUPTION KE JARIYE SAB KI KHOON CHOOS RAHE HEINE
   TRANSLITERATION: Isliye corruption ke jariye sabi kii khoon choos rahe hain.
   TRANSLATION: Thats why, sucking everyone’s blood through corruption

EXPLANABILITY: These two tweets also needs the contextual information to get the true sentiment. As the model is also learning the cross-lingual information it is able to detect it.

HATE IS SARCASTIC

1. अभी तो कबीर सिंह फिल्म की वजह ये लोग पागल हो रखे है, जब RX100 का रीमेक आएगा तबतो चूड़ियां तोड़ेगी ये फेमिनिस्ट.

TRANSLITERATION: abhi to kabir singh film kii wajah ye log pagal ho rakhe hai, jab RX100 kaa remake aaega tab to churiyaan torengi ye feminist..

TRANSLATION: Right now these people are going crazy because of Kabir Singh movie, when the remake of RX100 comes, then these feminists will break bangles.

2. BOLLYWOOD FILM DEKHNE KE SAMAY LOGIC GHAR MEIN CHORKE ANA PARTA HAIN. PLEASE LOGIC MAT GHUSAO

TRANSLITERATION:. Bollywood film dekhne ke samay logic ghar mein chorke ana parta hai. Please logic mat gusao.

TRANSLATION: you have to leave your brain behind before watching any Bollywood movie. Please don’t use any logic.

EXPLANABILITY: These tweets are sarcastic in nature. But as the encoder consists of all types of features, it is able to distinguish it.

Table 6. Comparison to the state-of-the-art systems and the proposed approach

Best Model (Weighted-F1)	**Comparison (Weighted-F1)**
D1 (**95.19**)	**[2] 92.20**
D2 (**82.78**)	**[26] 80.30**
D3 (**83.11)**	**[27] 77.97,[27] 77.48**
D4 (**65.97**)	**[25] 60.81**
D5 (**82.14**)	**[46] 80.0**
D6 (**86.14**)	**[33] 84.11, [33] 83.98**
D7 (**90.98**)	**[4] 89.50,[4] 89.30**
D8 (**92.17**)	**[29] 80**

5.2. Statistical Significance Test

We also determine whether a difference between the M-BERT in STL (M1) and Model 7 is statistically significant (at p<=0.05), for this we run a bootstrap sampling test on the predictions of two systems. The test takes 3 confusion matrix out of 5 at a time and compares whether the better system is the same as the better system on the entire dataset. The resulting (p-) value of the bootstrap testis thus the fraction of samples where the winner differs from the entire data set.

Table 7. Bootstrapping Test

Data	**Sample taken**	**p-value**
D1	60%	<=0.03
D2	60%	<=0.01
D3	60%	<=0.03
D4	60%	<=0.05
D5	60%	<=0.03
D6	60%	<=0.04
D7	60%	<=0.05
D8	60%	<=0.05

6. CONCLUSIONS AND FUTURE WORK

In this paper, we leverage a deep multi-task learning framework to leverage the useful information of multiple related tasks. To deal with the data scarcity problem we utilize a multi-task learning approach that enables the model by sharing representations between the related tasks and generalize better by achieving better performance for the individual tasks. Detailed empirical evaluation shows that the proposed multi-task learning framework achieves statistically significant performance improvement over the single-task setting.

We have leveraged the labeled corpora for each tasks and experimented on single task learning and multi-task learning paradigm. The plausible extensions include the inclusion of more affective phenomenon correlated to hate speech such as sarcasm/irony **[53]**, "big five" personality traits **[54]**, and emotion role labeling **[55].**

REFERENCES

[1] Patchin, Justin W., and Sameer Hinduja. "Cyberbullying and Online Aggression Survey." (2015).

[2] Bohra, Aditya, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. "A dataset of Hindi-English code-mixed social media text for hate speech detection." In *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media*, pp. 36-41. 2018.

[3] Liu, Joseph. "Religious hostilities reach six-year high." (2014).

[4] Mathur, Puneet, Ramit Sawhney, Meghna Ayyar, and Rajiv Shah. "Did you offend me? classification of offensive tweets in hinglish language." In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pp. 138-148. 2018.

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

[6] Soliman, Abu Bakr, Kareem Eissa, and Samhaa R. El-Beltagy. "Aravec: A set of arabic word embedding models for use in arabic nlp." *Procedia Computer Science* 117 (2017): 256-265.

[7] Ousidhoum, Nedjma, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. "Multilingual and multi-aspect hate speech analysis." *arXiv preprint arXiv:1908.11049* (2019).

[8] Mulki, Hala, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. "L-hsab: A levantine twitter dataset for hate speech and abusive language." In *Proceedings of the third workshop on abusive language online*, pp. 111-118. 2019.

[9] Chung, Yi-Ling, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. "CONAN--COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech." *arXiv preprint arXiv:1910.03270* (2019).

[10] Bretschneider, Uwe, and Ralf Peters. "Detecting offensive statements towards foreigners in social media." In *Proceedings of the 50th Hawaii International Conference on System Sciences*. 2017.

[11] Wiegand, Michael, Melanie Siegel, and Josef Ruppenhofer. "Overview of the germeval 2018 shared task on the identification of offensive language." (2018): 1-10.

[12] Sanguinetti, Manuela, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. "An italian twitter corpus of hate speech against immigrants." In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*. 2018.

[13] Bosco, Cristina, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. "Overview of the evalita 2018 hate speech detection task." In *EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, vol. 2263, pp. 1-9. CEUR, 2018.

[14] Álvarez-Carmona, Miguel Á., Estefanıa Guzmán-Falcón, Manuel Montes-y Gómez, Hugo Jair Escalante, Luis Villasenor-Pineda, Verónica Reyes-Meza, and Antonio Rico-Sulayes. "Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets." In *Notebook papers of 3rd sepln workshop on evaluation of human language technologies for iberian languages (ibereval), seville, spain*, vol. 6. 2018.

[15] Ptaszynski, Michal, Agata Pieciukiewicz, and Paweł Dybała. "Results of the poleval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in polish twitter." (2019).

[16] Ljubešić, Nikola, Tomaž Erjavec, and Darja Fišer. "Datasets of Slovene and Croatian moderated news comments." In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pp. 124-131. 2018.

[17] Çöltekin, Çağrı. "A corpus of Turkish offensive language on social media." In *Proceedings of the 12th language resources and evaluation conference*, pp. 6174-6184. 2020.

[18] Alfina, Ika, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. "Hate speech detection in the Indonesian language: A dataset and preliminary study." In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pp. 233-238. IEEE, 2017.

[19] Bhat, Irshad Ahmad, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. "Iiit-h system submission for fire2014 shared task on transliterated search." In *Proceedings of the Forum for Information Retrieval Evaluation*, pp. 48-53. 2014.

[20] Kumar, Ritesh, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. "Aggression-annotated corpus of hindi-english code-mixed data." *arXiv preprint arXiv:1803.09402* (2018).

[21] Aroyehun, Segun Taofeek, and Alexander Gelbukh. "Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling." In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pp. 90-97. 2018.

[22] Arroyo-Fernández, Ignacio, Dominic Forest, Juan-Manuel Torres-Moreno, Mauricio Carrasco-Ruiz, Thomas Legeleux, and Karen Joannette. "Cyberbullying detection task: the ebsi-lia-unam system (elu) at coling'18 trac-1." In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, pp. 140-149. 2018.

[23] Modha, Sandip, Prasenjit Majumder, and Thomas Mandl. "Filtering aggression from the multilingual social media feed." In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, pp. 199-207. 2018.

[24] Golem, Viktor, Mladen Karan, and Jan Šnajder. "Combining shallow and deep learning for aggressive text detection." In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pp. 188-198. 2018.

[25] Kumar, Ritesh, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. "Benchmarking aggression identification in social media." In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, pp. 1-11. 2018.

[26] Mandl, Thomas, Sandip Modha, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Prasenjit Majumder, Johannes Schäfer et al. "Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages." *arXiv preprint arXiv:2112.09301* (2021).

[27] Mandl, Thomas, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. "Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german." In *Forum for information retrieval evaluation*, pp. 29-32. 2020.

[28] Bashar, Md Abul, and Richi Nayak. "QutNocturnal@ HASOC'19: CNN for hate speech and offensive content identification in Hindi language." *arXiv preprint arXiv:2008.12448* (2020).

[29] Jha, Vikas Kumar, P. Hrudya, P. N. Vinu, Vishnu Vijayan, and P. Prabaharan. "DHOT-repository and classification of offensive tweets in the Hindi language." *Procedia Computer Science* 171 (2020): 2324-2333.

[30] Velankar, Abhishek, Hrushikesh Patil, Amol Gore, Shubham Salunke, and Raviraj Joshi. "Hate and offensive speech detection in Hindi and Marathi." *arXiv preprint arXiv:2110.12200* (2021).

[31] Chopra, Shivang, Ramit Sawhney, Puneet Mathur, and Rajiv Ratn Shah. "Hindi-english hate speech detection: Author profiling, debiasing, and practical perspectives." In *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 01, pp. 386-393. 2020.

[32] Santosh, T. Y. S. S., and K. V. S. Aravind. "Hate speech detection in hindi-english code-mixed social media text." In *Proceedings of the ACM India joint international conference on data science and management of data*, pp. 310-313. 2019.

[33] Bhardwaj, Mohit, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. "Hostility detection dataset in Hindi." *arXiv preprint arXiv:2011.03588* (2020).

[34] Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingmar Weber. "Automated hate speech detection and the problem of offensive language." In *Proceedings of the international AAAI conference on web and social media*, vol. 11, no. 1, pp. 512-515. 2017.

[35] Waseem, Zeerak, and Dirk Hovy. "Hateful symbols or hateful people? predictive features for hate speech detection on twitter." In *Proceedings of the NAACL student research workshop*, pp. 88-93. 2016.

[36] Zampieri, Marcos, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. "Predicting the type and target of offensive posts in social media." *arXiv preprint arXiv:1902.09666* (2019).

[37] Golbeck, Jennifer, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos et al. "A large labeled corpus for online harassment research." In *Proceedings of the 2017 ACM on web science conference*, pp. 229-233. 2017.

[38] Basile, Valerio, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. "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*, pp. 54-63. 2019.

[39] De Gibert, Ona, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. "Hate speech dataset from a white supremacy forum." *arXiv preprint arXiv:1809.04444* (2018).

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

[41] Bhattacharya, Shiladitya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr Ojha. "Developing a multilingual annotated corpus of misogyny and aggression." *arXiv preprint arXiv:2003.07428* (2020).

[42] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

[43] Khanuja, Simran, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam et al. "Muril: Multilingual representations for indian languages." *arXiv preprint arXiv:2103.10730* (2021).

[44] Zhang, Yu, and Qiang Yang. "A survey on multi-task learning." *IEEE Transactions on Knowledge and Data Engineering* (2021).

[45] Fortuna, Paula, Joao Rocha da Silva, Leo Wanner, and Sérgio Nunes. "A hierarchically-labeled portuguese hate speech dataset." In *Proceedings of the third workshop on abusive language online*, pp. 94-104. 2019.

[46] Kumar, Ritesh, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. "Evaluating aggression identification in social media." In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pp. 1-5. 2020.

[47] Xue, Ya, Xuejun Liao, Lawrence Carin, and Balaji Krishnapuram. "Multi-Task Learning for Classification with Dirichlet Process Priors." *Journal of Machine Learning Research* 8, no. 1 (2007).

[48] Ruder, Sebastian. "An overview of multi-task learning in deep neural networks." *arXiv preprint arXiv:1706.05098* (2017).

[49] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

[50] Chollet, François. "Keras: The python deep learning library." *Astrophysics source code library* (2018): ascl-1806.

[51] Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado et al. "Tensorflow: Large-scale machine learning on heterogeneous distributed systems." *arXiv preprint arXiv:1603.04467* (2016).

[52] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).

[53] Reyes, Antonio, Paolo Rosso, and Davide Buscaldi. "From humor recognition to irony detection: The figurative language of social media." *Data & Knowledge Engineering* 74 (2012): 1-12.

[54] Flek, Lucie. "Returning the N to NLP: Towards contextually personalized classification models." In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 7828-7838. 2020.

[55] Mohammad, Saif, Xiaodan Zhu, and Joel Martin. "Semantic role labeling of emotions in tweets." In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 32-41. 2014.

AUTHORS

Prashant Kapil is a PhD scholar in the Department of CSE at IIT Patna. The author would like to acknowledge the funding agency, the University Grant Commission (UGC) of the Government of Indiafor providing financial support in the form of UGC NET-JRF/SRF.
Research interests: AI, NLP, and ML

Asif Ekbal is an Associate Professor in the Department of CSE, IIT Patna, India.
Research interests: AI, NLP and ML.