An Online Multilingual Hate speech Recognition System

1st Neeraj Vashistha  
Queen Mary University of London  
London, United Kingdom  
ec1947@qmul.ac.uk

2nd Arkaitz Zubiaga  
Queen Mary University of London  
London, United Kingdom  
a.zubiaga@qmul.ac.uk

Abstract—The exponential increase in the use of the Internet and social media over the last two decades has changed human interaction. This has led to many positive outcomes, but at the same time it has brought risks and harms. While the volume of harmful content online, such as hate speech, is not manageable by humans, interest in the academic community to investigate automated means for hate speech detection has increased. In this study, we analyse six publicly available datasets by combining them into a single homogeneous dataset and classify them into three classes, abusive, hateful or neither. We create a baseline model and we improve model performance scores using various optimisation techniques. After attaining a competitive performance score, we create a tool which identifies and scores a page with effective metric in near-real time and uses the same as feedback to re-train our model. We prove the competitive performance of our multilingual model on two langauges, English and Hindi, leading to comparable or superior performance to most monolingual models.

Index Terms—social media, hate speech, classify, near-real time

I. INTRODUCTION

Hate Speech is a characterisation of communication which is 'hateful', controversial and generates intolerance and in some forms is divisive and demeaning. There is no legal definition of hate speech, but on several accounts accepted meaning of the term deals with communication in speech, behaviour or writing, remarks which are pejorative or discriminatory with reference to a person or a group of persons, either directly or indirectly. Such remarks are based on their religion, ethnicity, nationality, descent, race, colour, gender or other identity factor Cortese 2006.

Many countries have adopted frameworks and new laws have been constituted but due to the pervasive nature of online communications, only 3% of malicious communication offenders are charged Abusive and Offensive Online Communications: A Scoping Report 2018. This is because of lack of clarity and certainty in this area. Several of these frameworks prohibit the incitement to discrimination, hostility and violence rather than prohibiting hate speech. There is a need for quick identification of such remarks and an automatic system which can identify and take measures to prevent the instigation and incitement.

There are several examples of hate speech such as

Reminding feminists world-wide that they are only mad because they’re ugly...

Good morning WHITE people! F* you N*

which either implicitly or explicitly target an individual or a group of individuals and inflict mental pain that may eventually cause of social revolts or protests.

In this work, we study the existing methods designed to tackle hate speech. We have curated the data from the most prominent social media platform, Twitter. We also combine the existing datasets into one, as these are pre-annotated and previous researchers have used these to target several cases of hate speeches. Individually, these are small datasets but after we combine the existing datasets into one dataset, we get a large corpus of hate speech sequences. This dataset is a combination of the hateful content of almost 5 different categories, such as sexual orientation, religion, nationality, gender, and ethnicity. We create models which classify the content of text as hateful, abusive, or neither. The total size of the dataset is around seventy six thousand samples and due to the high variance and low bias we are avoiding sub-categorisation of hate classes.

We also study some state-of-the-art models which claim to give superior accuracy. Since these models are built on specific languages (English or Hindi), we utilise the work and propose our model which is multilingual. Finally, we use our optimised models to create a simple tool which identifies and scores a page if hateful content is found and uses the same as the feedback to re-train the model. While the vast majority of previous works have investigated the development of hate speech detection models for specific languages Vidgen and Derczynski 2020, here we propose a multilingual model which we experiment in two languages, English and Hindi, leading to competitive performance and superior to most monolingual models.

The main contribution of this work are as follows:

- Creating a system which is trained on a sufficiently large corpus of hate speech text.
- A system which is multilingual, can work on different languages or another language be added and trained quickly using our transfer learning models.
• A system which can learn new form of hateful remark or zipfian nature of language by re-training in an online environment.
• The resulting system has models which are simple, lightweight and are optimised to be used in an near-real-time online environment.

In this way the system holds a potential as a product for safer social media utilization as well as reduces the need for human annotators and moderators to tag disturbing online messages.

The rest of the paper is organised as follows: an overview of hate speech datasets and existing models is provided in the Related Work section. In the Model section, we critically analyse the existing models and discuss their limitations. We also discuss our model and several optimisations we performed to achieve a desirable performance score. Furthermore, we talk about the feedback mechanism for the optimised model and its usage in an online environment.

II. RELATED WORK

Here we discuss related work on hate speech datasets, hate speech detection models and the different approaches. This serves as a motivation for our work and helps us to bridge our study with existing research available.

A. Related Datasets

As mentioned earlier, the six publicly available datasets Mandl et al. [2019], Davidson et al. [2017], ElSherief et al. [2018], Ousidhoum et al. [2019], Basile et al. [2019] and Mathur et al. [2018] are manually curated and are of modest size. In the below text, we discuss the characteristics and the generation process of each of these datasets. Using these annotated datasets we create our own resource of hate speech sequences.

1) Data Gathering process:

The data in online diaspora can be curated from various technology companies and social media platform providers such as Facebook, Google, Internet Service Provider Association, Oath, Twitter, Snap Group Limited etc. Twitter’s Public API, with its easy to use, highly availability and accessibility, it is one of the most sought and targeted online social media platforms. We can set up a mechanism and relevant data for hate speech can be sourced. The dataset we are using in this study comes from Twitter and is labelled by the annotators into pre-specified categories and subcategories of hate speech.

The examples sourced from Twitter are based on keyword matching with terms which have strong connection with hate. These are annotated and tagged as abusive or hateful. This task has to be done manually and is very time consuming. The same process has been adopted for code-mixed and pure Hindi language. Since the sample space of Mathur et al. [2018] and Mandl et al. [2019] is small, we have tried to update these datasets with our own curated samples.

To describe the multilingual aspect, the dataset consists of three different types of target language, English, Hindi and Code Mix Hindi. Below 1 describe various attributes of the dataset. Although there are some limitations such as small sample text size of 160 characters or under representation of all forms of hate speech sequences but this does not, in any way, limit the use of data. Other issues related to this kind of dataset is the filtering of unwanted details like email address, URL, date, phone number, username etc due to which textual and contextual information is lost. We also try that the collation of different dataset doesn’t impact the attributes. As these datasets are tagged by different authors as per their own perception and with the daily evolution of slangs and jargons, evaluation becomes a tedious task.

2) Existing Datasets:

HASOC2019: The dataset presented in (ibid.) shared task on hate speech and offensive content identification in Indo-European language contain around 7005 post in English and 9330 post in Hindi language, with 36% and 52% abusive content in respective languages. The task is focused on three sub-tasks,

• if the tweet text is either hate or offensive or neither
• if tweet text is either hate, offensive or profane
• if the tweet text is targeted towards an individual/group or is untargeted

For this study we are only considering the data presented for the first sub-task.

TDavidson et al: The dataset presented in Davidson et al. [2017] research includes tweets labeled by crowdflower workers into three categories: hate speech, offensive, or neither. There are 24,802 English twitter text instances with only 6% abusive content.

ElSherief et al: The dataset presented in ElSherief et al. [2018] research is procured from twitter and it is categorised into 7 different categories of Hate. The data is being sought from Waseem and Hovy [2016] and Davidson et al. [2017] with their own annotation from No Hate Speech Movement (1) and Hatebase (2). The total number of samples is 10760 with total abusive content around 10%. This dataset is in English language.

Ousidhoum et al: Here the author has categorised twitter data into 46 different sentiments of hateful, offensive, fearful, abusive, disrespectful, normal and different combinations of each. The dataset presented Ousidhoum et al. [2019] is in English and it also captures the annotators sentiments, directness, intended targets and groups. These attributes are important but to create a homogeneous dataset these have been left out from this study. The total size of this dataset is around 5647

---

1. www.nohatespeechmovement.org
2. hatebase.org
instances which are reorganised for our purpose into three categories, hateful, abusive, or neither, with abusive content of around 26%.

**SemEval 2019 Task 5**: The dataset presented Basile et al. [2019] comprises 12906 text instances of which 42% are abusive in nature. Again the data is sourced from twitter domain. This is a shared task with three sub-task in it. The first task is a binary task to classify if the text instance is hate or not, which is what we will consider for our study. The other tasks are within hate, if the text is directed towards a group or individual and within hate, if the text is aggressive or not. This dataset is hate speech sequences against immigrants and women in English language.

**PMathur et al.**: The dataset presented in the Mathur et al. [2018] contains tweets of offensive nature along with a profanity word list in code-mixed Hindi. This is a phonetic translation of a Hindi word written in English. These tweets are annotated into three different categories of normal, abusive and Hate. For our study, the profanity word list has been updated with new words and with Hindi dialect references. Using this profanity word list, new tweets have been curated from twitter for both Hindi and code-mixed Hindi text. These new tweets have been manually annotated following the guidelines mentioned by the original author, details of which are given in Model Section. The resulting size of profanity word list is around 226 words in code mix Hindi, with the meaning of each in English, proficiency score and devanagari script Hindi word. The size of the dataset is around 3189 with 55% tweets of abusive nature.

| Dataset Name          | Total Records | Hate % | Abuse % |
|-----------------------|---------------|--------|---------|
| HASOC2019 - EN        | 7605          | 36.38  | -       |
| SEMOC2019 - HI        | 9330          | 52.92  | -       |
| TDDavidson et al      | 24783         | 77.43  | 5.77    |
| ElSherif et al        | 10760         | 90.94  | -       |
| Ousidhoum et al       | 5647          | 65.66  | 22.63   |
| SemEval 2019 Task 5   | 12906         | 42.15  | -       |
| PMathur et al         | 3189          | 55.34  | 9.5     |
| Total                 | 76403         | 61.84  | 4.55    |

The data from all the sources have been collated and sorted into a homogeneous form, the details of which are illustrated in the Model section.

**B. Related Models**

Davidson et al. [2017] in his work followed a very simple approach and created a very simple model. In this work, the author created simple feature vectors for each sample instance. The feature vector consisted of Part of Speech(POS), term frequency–inverse document frequency (TFIDF) vectors amongst others. Final model was a logistic regression and linear support vector machine (SVM) which claimed to achieve an overall precision 0.99, recall 0.90 and F1 score of 0.90.

In another work by Ousidhoum et al. [2019] the author has created two types of models. Bag-of-word (BOW) a feature based logistic regression model and a deep learning based model containing Bi-directional Long Short Term Memory (BiLSTM) layer with one hidden layer. The author claims, due to the small size of the dataset, deep learning based models performed poorly. The author suggests to use Sluice network Ruder et al. [2017] as they are suitable for loosely related tasks such as the annotated aspects of the corpora. State-of-art performance scores are obtained using Sluice networks.

The author also claimed that for a single task single language the model achieved an accuracy of 94% , this is in contrast against our dataset. We found that the model only achieved 68% accuracy for English language. We could not use Hindi or code-mixed Hindi due to the fact that the embedding used by the author, Babylon multilingual word embeddings Smith et al. [2017] and MUSE Lample et al. [2017] were not compatible with our Hindi or code-mixed Hindi dataset.

One of the initial research on hate speech detection from Hindi-English tweets was done by Bohra et al. [2018]. The research was based on 4575 code-mix Hindi tweets. The author worked on features like character n-gram, emoticon count, word n-gram, punctuation count and applied dimension reduction techniques. An accuracy of 71.7% on SVM and 66.7% on Random forest was obtained.

The most relevant research done for code-mixed Hindi dataset was done by Mathur et al. [2018]. The author claimed that term frequency, inverse document frequency (TF-IDF) and BoW features with SVM model gave peak performance when compared with other configurations of baseline supervised classifiers. The other features such as Linguistic Inquiry and word count features (LIWC) Sawhney et al. [2018] proficiency vector Jay and Janschewitz [2008] with glove and twitter embeddings gave best results against the baseline model as argued by the author. Also, a transfer learning based approach called as Multi Channel Transfer Learning based Model (MTLM) achieved on its best configuration a precision of 0.851, recall of 0.905 and F1 score of 0.893 . Similar scores have been obtained in our study as well.

Santosh and Aravind [2019] applied deep learning based approaches. In the experiment the author created sub-word level LSTM model and Hierarchical LSTM model with attention based phonemic sub-words. The architecture of these models are important research in this field as attention based models perform highly in comparison to basic deep learning based models. Hierarchical LSTM with attention achieved an accuracy of 66.6% while sub-word level LSTM model scored 69.8%.

In the work of Kamble and Joshi [2018] three deep learning models using domain specific word embeddings were created. These word embeddings are available and comprise 255,309 code-mix Hindi text which the author collected from Twitter. The word embeddings are trained using gensim’s word2vec model and are used in the three deep learning models. For 1D-CNN an accuracy of 82.62% was achieved which was the highest against BiLSTM with 81.48% and LSTM with 80.21%.

In the above mentioned researches we have seen that all
the models were created to process only a single language at a time. In our project we have focused on creating a single model that will inculcate multiple languages at the same time giving it a more global approach. We found another model on similar lines was created by Sohn and Lee 2019 where the authors focus on English and Chinese. However, the author used third party translation APIs like Google translation API, to convert the raw text into English or Chinese.

In the following sections we propose a Logistic regression model and two Deep Learning based models for both English and Hindi (code-mixed Hindi) languages. The model architecture for all the languages is the same, however the feature building process is different for different languages.

### III. Model

In this section we discuss the model building procedure. We are building a system that learns from the feedback, is multilingual and is trained on large hate sequence text. We also provide optimisation carried out and performance verification.

#### A. Data Preprocessing

Since the data in our use case is sourced from various datasets Basile et al. 2019, Mandl et al. 2019, Davidson et al. 2017, Ousidhoum et al. 2019, Mathur et al. 2018, ElSherief et al. 2018 it becomes essential to organise it into a homogeneous form. To achieve this, we consider the text and the class type of the existing datasets. In many datasets, as seen in [I], the data is classified as hate or not, and abuse class is not present. The original text is taken without any formatting. The class types (if necessary) are converted from original to either, hate, abusive or normal. In order to identify the source of the text, we have labeled each record instance with specific dataset identifiers.

Samples for Hindi and code-mixed Hindi are sourced from Mathur et al. 2018 and Mandl et al. 2019. This dataset is relatively small in comparison to English language dataset. Thus, we have updated this data sample by adding new tweet samples to Hindi and code-mix Hindi dataset by following the below procedure.

1) We have added and updated the profane word list provided by Mathur et al. 2018 with more words in code-mix Hindi language.
2) We have used this profane word list and added corresponding Hindi devnagari scripted text against each code-mix Hindi profane word.
3) In order to assign classes to newly curated text from Twitter Public API, we have followed the below guideline.

   a) Tweets involving sexist or racial slur to target a minority, these tweets may contain abusive word are annotated as Hate,
   b) Tweets which represent undignified stereotypes are marked as Hate
   c) Tweets which contain problematic hashtags are marked as Hate.

d) Tweets which contain only abusive words are tagged as Abusive.

e) Other tweets are marked as normal.

Typically few instances of our seventy six thousand dataset looks something like in the table [III].

| Dataset Type     | Example                                      | Class Type |
|------------------|----------------------------------------------|------------|
| English-EN       | I am literally too mad right now a ARAB     | Hateful    |
|                  | won #MissAmerica                            |            |
| English-EN       | Black on the bus                             | Hateful    |
| English-EN       | I can not just sit up and HATE on another   | Abusive    |
|                  | b** .. I got too much sh** going on!        |            |
| Hindi-HI         | वे लिखनी की गीड़ में लिखन की गाव डालता हूँ अभी विद्यालय गए हैं    | Hateful    |
| Hindi Code-Mixed | Main jut Punjabi hoon aar paka N league.    |            |
|                  | Madarchod Imran ki Punjab say nafrat clear  |            |
|                  | hai.                                         |            |

One of the major issues related to twitter text and to which other studies have indicated as the potential cause for degrade in performance is the small sequence of text followed by ever evolving use of unknown words such as slang and hashtags words. As seen in Table [I] the Max Seq. Len. denotes the maximum sequence length of the various language datasets, the sequence length for all the languages is very small (less than 100), but the vocabulary size (vocab. size) is way too high. Chen et al. 2019 describes the limitations of unusually large vocabulary leading to poor performance.

In our research, when we performed exploratory data analysis (EDA), we too concluded that due to above mentioned issues, the performance of our classifier was getting affected. Thus during preprocessing of text we decided to use ekphrasis Baziotis, Pelekos, and Doulkeridis 2017. This preprocessor is used in the following way:

- Normalise - url, email, percent, money, phone, user, time, date, number in twitter text
- Annotate - hashtag, allcaps, elongated, repeated, emphasis, censored, words in the text
- Word Segmentation
- Convert Hashtags to unpacked word list (if possible)
- Convert English contractions such as “can’t”, “we’ll” into “cannot” and “we will”
- Tokenise the sentences into words
- Convert Emoticons and Slangs into actual expression/phrase.

The output from this pre-processing technique contains more information than previously studied research. Although, ekphrasis works really well for English language but it does not have a multilingual aspect built into it. Therefore, we are leveraging Kunchukuttan 2020 Indic NLP and Loper and Bird 2002 NLTK library support for Hindi Language. During the EDA, we observed some issues related to unpacking of
slang, emoticons and English contractions in ekphrasis and
tokenisation issues in Kunchukuttan 2020. These are now
being rectified by raising an issue on Github and by making
an open source contribution to the original work.

B. Model Building

In previous research, we saw that BoW or TFIDF based
feature vectors along with other features tends to work best
with Logistic regression based models and they generally give
a fine baseline model performance. Similar approach has been
carried out in our study.

We created a Logistic regression model as the baseline
model. We applied an L2 penalty giving equal class weights
to three classes. To comprehend the sheer volume of data and
to make the model converge, the model’s hyper-parameter,
maximum iteration is set to 5000 iterations for English, 3000
for Hindi and code-mix Hindi. Other hyper-parameters were
obtained in grid search, with 5 fold cross validation, and the
performance scores are described in Table IV.

| Dataset Type | f1-Neither | f1-Hate | f1-Abuse | Acc  |
|--------------|------------|---------|----------|------|
| EN           | 0.68       | 0.47    | 0.80     | 0.687|
| HI           | 0.95       | 0.96    | -        | 0.956|
| HI-Code Mix  | 0.84       | 0.62    | 0.91     | 0.855|

After building a baseline model, we explore the possibility
of building a Hierarchical Deep Neural Network by combining
several Convolutional Neural Network (CNN) filters into Bi-
directional Long Short Term Memory network. During the
EDA we discovered valuable sequential information in twitter
text for each class. When we apply a BiLSTM layer on top of
a CNN layer, we are able to capture the sequential information
as well as the low lying textual representational. Thus we
improve the simple contextual BiLSTM classification with use
of CNN layers. This model has the following architecture:

1) Word Embedding - to capture and convert text into
sequences.
2) CNN - to capture low lying information using different
filter size
3) Bidirectional LSTM - to capture contextual information
of the sequence.
4) Fully-Connected output

This model is inspired by the work of Zhang and LeCun
2015, Jozefowicz et al. 2016 and Kim et al. 2015. In these
researches, with use of CNN, the character level property of
the text is explored.

Figure 1 describes the architecture and different layers. The
word embedding layer converts sparse representation of word
sequences into dense vector representation. The 3 CNN layers
consist of 3 parallel convolutional, 2 dimensional filters, of size
3, 4 and 5 with batch normalisation and max pooling layer.
This CNN layer is time distributed over the LSTM layer, which
captures the contextual information and finally the model is
fed forwarded to a dense layer, where sigmoid activation layer
performs the classification. This model is built in order to
analyse the use of deep neural networks and, if there can be
any improvement in the previous benchmarks with respect to
models performance.

The above network is trained on all three datasets and after
optimisations the results obtained are relatively similar to the
Logistic regression model.

| Dataset Type | f1-Neither | f1-Hate | f1-Abuse | Acc  |
|--------------|------------|---------|----------|------|
| EN           | 0.74       | 0.63    | 0.72     | 0.78 |
| HI           | 0.86       | 0.86    | -        | 0.85 |
| HI-Code Mix  | 0.83       | 0.62    | 0.70     | 0.86 |

Table IV describes the performance for three language
datasets. This model is trained on GPU thus the model
convergence time is way quicker in comparison to the logistic
regression model.

We use random search to find the parameters and hyper-
parameters. Table V describes the parameters which resulted
in the best performance.

| Dataset Type | Epochs | Batch Size | Optimiser | Learning Rate | Dropout | Hidden Size |
|--------------|--------|------------|-----------|---------------|---------|-------------|
| EN           | 22     | 30         | sgd       | 0.001         | 0.2     | 64          |
| HI           | 20     | 30         | sgd       | 0.001         | 0.1     | 32          |
| HI-Code Mix  | 22     | 30         | adam      | 0.001         | 0.2     | 64          |

In the previous model, the word embeddings are learned
from data that have low dimension and are more dense than
TFIDF features. But the model seemed to be less performant
than the logistic regression model built previously. Therefore,
in the next model we incorporate Bidirectional Encoder Rep-
resentations from Transformers (BERT) Devlin et al. 2019 to
leverage contextual word embeddings in place of CNN layers
we previously added to the CNN LSTM model. This model
has the following architecture:

1) Pre-trained BERT Embedding - to capture contextual
word embeddings
2) Bidirectional LSTM - to capture contextual information
of the sequence.
The models developed in this study are oriented to be used in an online environment. Therefore, it becomes important that we pursue a state of art model which can be re-trained in a given frame of time within a resource constrained environment. It is critical to analyse the prediction time as well as the time taken by the model to train.

| Model Type | Preprocess Time | Training Time | Inference Time | System Req. | Model size | Test Acc. |
|------------|-----------------|---------------|----------------|-------------|------------|-----------|
| LR-EN      | 1453sec         | 4192sec       | 0.7sec         | CPU         | 161.2KB    | 85%       |
| LR-HI      | 1453sec         | 4192sec       | 0.7sec         | GPU         | 402.2KB    | 85%       |
| CNN-LSTM-EN| 218sec          | 210sec        | 0.4sec         | GPU         | 13.9MB     | 78%       |
| CNN-LSTM-HI| 712.2MB         | 14927sec      | 0.9sec         | GPU         | 712.3MB    | 78%       |

The table [IX] summaries some of the important performance features of the models built so far.

- The Logistic regression model takes an excessive amount of time for preprocessing and training. The BERT and CNN LSTM model takes much less time in comparison. This is due to the fact that the Logistic regression model requires TFIDF features on the entire vocabulary size.
- The model size for Logistic regression models are thousand times less than the BERT model and 100 times less than CNN LSTM model. Thus the inference time for the Logistic model is the least as compared to other models.
- The logistic regression model is well suited for smaller datasets as it is trained on CPU, for larger datasets GPU or TPU based CNN LSTM or BERT based models could be used to reduce the training and inference time.
- One of the reasons for high inference time for CNN LSTM models is inability to find the right set of hyper-parameters, and so we see a higher performance for BERT models as compared with other models.

IV. RESULTS AND DISCUSSIONS

We show that the Logistic regression model supplemented with TFIDF and POS features gave relatively good results in comparison to other models but the time taken for the model to converge is very long. These results presented are in line with previous research Badjatiya et al. [2017], Davidson et al. [2017], Mathur et al. [2018].

The Deep learning model, gave similar performance scores without much feature engineering and the model converged quickly too. Such results can be attributed to use of embedding layers in the neural network. We also tried to use the glove and twitter embedding layers but due to the high number of unknown words, desirable results were not obtained.

The logistic regression model has less system requirements, is very quick to infer a single sample instance and has great performance scores. Due its inability to scale on a large dataset and the time taken to build the model is very high. It can only
be used as a benchmark for other models and cannot be used in an online environment, where models are continuously retrained on the feedback loop.

The CNN LSTM model is an average model which has a mediocre performance but it’s performance can be perfected. The random search optimization used in this study is a test driven approach. Other optimization techniques like bayesian optimization can find better hyper-parameters. But the time taken to build the model will increase considerably. The CNN LSTM model requires a moderate GPU processing and infer a single sample instance in near-real time. Thus we are utilizing this model in our online web application.

The BERT model is a high performance state of the art model. It has the highest accuracy amongst all the models. If the system requirement(TPU) criteria is fulfilled, it can be used in an online environment. This model has very less training time and is very quick to infer a single sample instance. The only constraint is the use of the Tensor Processing Unit. Though the model size is 1000 times the size of a logistics regression model, memory and disk space consumption is hardly a worry these days due to their easy availability. Unlike GPU the cost of a single TPU instance is very high. Still the scope of TPU usage as the main stream processing unit for machine learning in near future is high. Thus, the research done in this study is optimistic and this novel state-of-the-art model is very much applicable to be used in the real world.

In order to give a fair comparison of our models to existing ones, we apply our models to the datasets in isolation, we find that our models outperforms on most of the datasets. Table X shows these results. The dataset was segmented in the exact proportions of test and train as it was done in the original research. Further, out of the many models we described here, we have chosen the best results (after optimising it to work on smaller datasets) of CNN LSTM model as it works faster in loading the data as well as giving inference. The results are consistent with the performance which is consistent as well across the datasets.

### TABLE X
Comparison of different dataset models with our model

| Datasets                  | Existing Best Scores | Our Model Best Scores |
|---------------------------|----------------------|-----------------------|
|                           | Precision Recall F1  | Precision Recall F1   |
| HASOC 2019 EN             | 0.9                  | 0.91                  |
| HASOC 2019 HI             | 0.9                  | 0.9                   |
| Davidson et al. 2017      | 0.9                  | 0.9                   |
| Elser et al.              | 0.8                  | 0.86                  |
| Daudshoorn et al. 2019    | 0.9                  | 0.91                  |
| Sentivul 2019 Task 6      | 0.68                 | 0.85                  |
| Mathur et al. 2018        | 0.85                 | 0.89                  |

Further, we discuss some categories of errors that were observed in the deep learning and logistic regression models:

- The noisy and repetitive nature of the data present on social media creates a skewness in the class distribution. Although it is taken care of by hyper-parameter tuning, nonetheless, it still caused overfitting in both logistic regression models and CNN LSTM models.

- The code mixed words tend to be biased as they appear at specific locations. Singh [1985] showed that bilingual languages favor code mixing of specific words at specific locations.

- Almost all the class labels are hand annotated. Since there is no defined criteria of how one should classify, it can lead to ambiguity between classes. This is the reason for lower f1 scores between abusive and hate classes in both logistic regression and CNN LSTM models for all the languages.

### V. Online Feedback Mechanism

In this section, we discuss the use of created models in an online environment. Once the model is created it is essential to understand how the model will behave. Here, we will also discuss the complete execution of the machine learning pipeline; by adding an external feedback loop to the model. This allows the model to learn the evolution of text and textual context over time. We are aware that there are several disadvantages of doing this like the model may become biased towards one class if used over a stretch of time, but we can overcome this by adding human-in-the-loop. In this way the weight of tagging text by moderator and annotators could be reduced significantly.

To achieve this, we create a RESTful API which connects the machine learning model to an online web chat system. The web chat we create here is a live application demonstrating a chat room. The most important aspect is the API which scans the page and scores each comment by the user. It summarizes the total score of all the classes viz. normal, hateful or offensive. In the image we see the metric which understands the page content with respect to it’s hateful or abusive nature.

![An Online Multi Lingual Hate Speech Recognition System](image)

Fig. 3. Real time monitoring of content and scoring page based on in percentage if Hateful content is present.

The above use case is a passive method showcasing the capability of our model in an online application. An active use case would be if we can prevent the use of hateful/abusive comments. In order to do so, we create a notifier on the submit button of the comments. If the comment is hateful or abusive, the user gets notified and we can actively prevent users from commenting derogatory remarks. Thus, users can be made self-
aware and can be advised to refrain from sending something incinerating to the social world. Figure 4 depicts the mentioned use case.

![An Online Multi Lingual Hate Speech Recognition System](image)

Fig. 4. Proactively notifying user that their comments are hateful in nature

The above application has the capability to automatically switch between different machine learning models depending on the language of the user. If the user uses Hindi, the model automatically switches to the Hindi model.

We capture these online comments into a database and use this database to update our existing datasets. The database can be reviewed by an annotator for comments which have lower confidence of belonging to a single class. Thus a very small portion of the comments have to be reviewed. Such comments can be added to the training dataset for future addition.

### VI. Conclusion

The objective of this study was to bring to light a model which was trained on a large dataset of multilingual languages. By experimenting on an aggregated dataset combining six languages in English, Hindi and Code-mixed Hindi, we demonstrate that our models achieve comparable or superior performance to a wide range of baseline monolingual models. The model leads to competitive performance on combined data and works in an online environment in near-real time.

Further work can be done both in the space of improving the model architecture. We can add more use cases to the online system. From the Logistic Regression point of view, we can apply feature selection methods to choose most prominent features and extend the model to other code-mixed languages. We can exploit other advanced hierarchical models such as Spinal, GRU. Bayesian optimization and other fine tuning deep learning models and hyper-parameter selection are some of the areas of future research.

### References

Abusive and Offensive Online Communications: A Scoping Report (2018).

Badjatiya, Pinkesh et al. (2017). “Deep Learning for Hate Speech Detection in Tweets”. In: Proceedings of the 26th International Conference on World Wide Web Companion - WWW ’17 Companion. DOI: [10.1145/3041021.3054223](http://dx.doi.org/10.1145/3041021.3054223)

Basile, Valerio et al. (June 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. Minneapolis, Minnesota, USA: Association for Computational Linguistics, pp. 54–63. DOI: [10.18653/v1/S19-2007](https://www.aclweb.org/anthology/S19-2007)

Baziotis, Christos, Nikos Pelekis, and Christos Doulkeridis (Aug. 2017). “DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis”. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Vancouver, Canada: Association for Computational Linguistics, pp. 747–754.

Bohra, Aditya et al. (June 2018). “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. New Orleans, Louisiana, USA: Association for Computational Linguistics, pp. 36–41. DOI: [10.18653/v1/W18-1105](https://www.aclweb.org/anthology/W18-1105)

Chen, Wenhu et al. (June 2019). “How Large a Vocabulary Does Text Classification Need? A Variational Approach to Vocabulary Selection”. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, pp. 3487–3497. DOI: [10.18653/v1/N19-1352](https://www.aclweb.org/anthology/N19-1352)

Cortese, Anthony Joseph Paul (2006). Opposing hate speech. Greenwood Publishing Group.

Davidson, Thomas et al. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. arXiv: [1703.04009](https://arxiv.org/abs/1703.04009) [cs.CL]

Devlin, Jacob et al. (2018). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: arXiv preprint arXiv:1810.04805.

-- (June 2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: [10.18653/v1/N19-1423](https://www.aclweb.org/anthology/N19-1423)

ElSherief, Mai et al. (2018). Peer to Peer Hate: Hate Speech Instigators and Their Targets. arXiv: [1804.04649](https://arxiv.org/abs/1804.04649) [cs.SI]
Jay, Timothy and Kristin Janschewitz (Jan. 2008). “The pragmatics of swearing”. In: Journal of Politeness Research-language Behaviour Culture - J POLITENESS RES-LANG BEH CUL 4, pp. 267–288. doi: [10.1515/JPLR.2008.013]
Jozefowicz, Rafal et al. (2016). Exploring the Limits of Language Modeling. arXiv: [1602.02410 [cs.CL]]
Kamble, Satyajit and Aditya Joshi (2018). Hate Speech Detection from Code-mixed Hindi-English Tweets Using Deep Learning Models. arXiv: [1811.05145 [cs.CL]]
Kim, Yoon et al. (2015). Character-Aware Neural Language Models. arXiv: [1508.06615 [cs.CL]]
Kunchukuttan, Anoop (2020). The IndicNLP Library. https: //github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf.
Lample, Guillaume et al. (2017). Unsupervised Machine Translation Using Monolingual Corpora Only. arXiv: [1711.00043 [cs.CL]]
Loper, Edward and Steven Bird (2002). “NLTK: The Natural Language Toolkit”. In: Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics.
Mandl, Thomas et al. (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 ‘19. Kolkata, India: Association for Computing Machinery, pp. 14–17. ISBN: 9781450377508. doi: [10.1145/3368567.3368584] url: https://doi.org/10.1145/3368567.3368584.
Mathur, Puneet et al. (Oct. 2018). “Did you offend me? Classification of Offensive Tweets in Hinglish Language”. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2). Brussels, Belgium: Association for Computational Linguistics, pp. 138–148. doi: [10.18653/v1/W18-5118] url: https://www.aclweb.org/anthology/W18-5118
Ousidhoum, Nedjma et al. (Nov. 2019). “Multilingual and Multi-Aspect Hate Speech Analysis”. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, pp. 4675–4684. doi: [10.18653/v1/D19-1474] url: https://www.aclweb.org/anthology/D19-1474
Ruder, Sebastian et al. (2017). “Sluice networks: Learning what to share between loosely related tasks”. In: ArXiv abs/1705.08142.
Santosh, Tokala and K. Aravind (Jan. 2019). “Hate Speech Detection in Hindi-English Code-Mixed Social Media Text”. In: pp. 310–313. doi: [10.1145/3297001.3297048]
Sawhney, Ramit et al. (July 2018). “A Computational Approach to Feature Extraction for Identification of Suicidal Ideation in Tweets”. In: Proceedings of ACL 2018, Student Research Workshop. Melbourne, Australia: Association for Computational Linguistics, pp. 91–98. doi: [10.18653/v1/P18-3013] url: https://www.aclweb.org/anthology/P18-3013
Singh, Rajendra (1985). “Grammatical Constraints on Code-Mixing: Evidence from Hindi-English”. In: Canadian Journal of Linguistics/Revue canadienne de linguistique 30.1, pp. 33–45. doi: [10.1017/S0008413100010677]
Smith, Samuel L. et al. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. arXiv: [1702.03859 [cs.CL]]
Sohn, H. and H. Lee (2019). “MC-BERT4HATE: Hate Speech Detection using Multi-channel BERT for Different Languages and Translations”. In: 2019 International Conference on Data Mining Workshops (ICDMW), pp. 551–559.
Vidgen, Bertie and Leon Derczynski (2020). “Directions in Abusive Language Training Data: Garbage In, Garbage Out”. In: arXiv preprint arXiv:2004.01670.
Waseem, Zeerak and Dirk Hovy (2016). “Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter”. In: Proceedings of the NAACL Student Research Workshop. San Diego, California: Association for Computational Linguistics, pp. 88–93. url: http://www.aclweb.org/anthology/N16-2013
Zhang, Xiang and Yann LeCun (2015). Text Understanding from Scratch. arXiv: [1502.01710 [cs.LG]]