Mayank Kumar et al. International Journal of Recent Research Aspects ISSN: 2349-7688, Special Issue: Recent Innovation in Science, Engineering & Management, May 2022, pp. 71-82

Hate Speech Detection on Social Media using Artificial Intelligence

Mayank Kumar*, Nitin Sharma*, Rohit Kumar*, Satyam Kumar Roy*, Yash Sharma*, Dr. Parli B. Hari**

***Student, BCA VI Semester, DPGITM, Gurugram, Haryana (India)**
****Associate Professor, Computer Science, DPGITM, Gurugram, Haryana (India)**

ABSTRACT

The emergence of social media is a blessing for society. Despite all the social media policies in place against "hate towards women and children", it continues to persist more than ever. In this paper, we propose a method to perform text style transfer addressing the problem of child and women safety in India. Towards this end, we introduce a standard dataset for combating the hate against women and children in English and Hinglish collected from Twitter as these are very common means of communication for the majority of people in India. Following this, we aim to solve the problem of normalizing offensive tweets to non-offensive one based on the concept of style transfer. Our proposed model is based on an encoder-decoder network combined with attention mechanism to normalise the hate tweets and preserve their context. We also employ a Bert Capsule classifier to classify hate based tweets. Our proposed model exceeds several classification and style transfer based baseline models with overall accuracy and BLEU score of 83.37% and 0.5417 respectively.

INTRODUCTION

Social media encompasses online technology and methods through which people can share content, personal opinion, swap different perspectives and insights into world issues, and generally discuss the evolution of media itself. Social media is expanding very briskly which tends to bring the users especially kids and adolescents in jeopardy as they spend a lot of time on social networking sites to communicate with each other, share information and knowledge. While social media users get benefited by utilizing social media, likewise they are also at risk of being prone to a lot of offensive content in different formats including image, video, and text. As youths are bound to be adversely influenced by one-sided and hurtful content than adults, identifying and normalizing offensive content for the online well-being of users turns into a critical assignment.

In a survey conducted by business standards [1] where they considered a sample size of 14000 girls aged between 15 to 25 years shows 58% females faced online harassment and abuse. This resulted in one in five (19%) girls leaving social media or reducing the amount of time they were spending on social media significantly. Also, one in ten (12%) females changed their way of expressing themselves. Driving youngsters out of online spaces is massively weakening in this progressive world which harms their capacity to be seen, acknowledged and heard. Also according to the analysis done by Hindustan times [2], social media has an impact in the radicalization of kids and youth as they are very much sensitive to online exploitation, illegal behaviour or sexual abuse. According to The Hindu [3], India tops the list when it comes to online child sexual abuse with 11.7% of the total reports or at 19.87 lakh reports.

Style transfer is one of the significant problems in numerous subfields of Artificial Intelligence (AI) like Natural Language Processing (NLP), Computer Vision (CV) etc., as it mirrors the capability to produce novel contents. Especially, in Natural Language Generation (NLG), style transfer of texts is considered as an important component as it complements numerous NLP applications [4].

In social media networks, use of offensive language is a very familiar problem of derogatory behaviour. Numerous researchers in the past have approached this problem using different techniques for distinguishing the offensive language from the non offensive ones [5] [6] [7] [8] [9] [10] [11]. However these works were often based on the theory that it is enough to detect and remove the complete offensive post. Instead, the offensive tweets should be converted to a no offensive one keeping the context intact and this is known as text style transfer. Text style transfer is basically a method in which the source or the original sentence of one style is converted into a target sentence of another style by preserving the content and context of the source sentence. Let us consider an original offensive tweet: "you know you fucked up when you been entertaining a nigga and then he starts calling you his wife

shit". Its corresponding target non-offensive tweet: "you know you messed up when you been entertaining a dude and then he starts calling you his wife shoot".

In this paper, we propose a method to perform text style transfer addressing the problem of child and women safety in India as this area is yet to be explored where the Indian slang cannot be generalized with the global ones. Thus, we aim to solve the problem of normalizing offensive tweets to non-offensive based on the concept of style transfer. For this, we introduce a India centric dataset collected from Twitter containing offensive tweets which is specifically focused on women and child safety in India. Our proposed approach is based on an encoder-decoder network along with attention [12]. The transferred tweets use the vocabulary that is common in a hate speech domain. Also, most of the previous work done in classifying the offensive tweet in text style transfer has used didn't explore Bidirectional Encoder Representations from Transformers i.e. BERT based models. Here, we employ a transformer based capsule network, namely, BERT-Caps [13] for classifying the tweets. The performance of the proposed model is evaluated using several quantitative metrics such as classification accuracy, BLEU score [14].

The key contribution of this paper are as follows:

- A new hate speech dataset is introduced for normalizing hate speech against women and children in India collected from Twitter.
- Our model performs style transfer by normalizing the hate tweet to not-hate one while preserving the context using encoder-decoder network with attention. To classify each tweet, Bert-Capsule classification model is employed.
- The proposed approach outperforms several baseline models in terms of classification and style transfer based models.

RELATED WORK

Background

There are numerous works in literature intended for text style transfer discussed below.

Shen [15], Zhao [16], Melnyk [17], Fu [18] presented an adversarial training based method to segregate the content and the style in a sentence. For instance, multi-decoder and style embedding are the two models that Fu et al. 2017 came up with and each one of them learns a representation for the input sentence that only contains the content information. Following this, all the decoders that are present in the multi decoder are assigned to each style which generates a sentence in their respective style. On the other hand, style embedding is opposite to multi-decoder as it learns the style embedding along with the content representation. Similarly, Shen [12] also encodes the original sentence into a vector form but unlike Fu [15], their discriminator makes use of Recurrent Neural Network (RNN) decoder's hidden states. The objective of the content encoder here is to trick the style discriminator when adversarial training is applied on unsupervised text style transfer by separating the context information from content embedding. Moreover, the non-differentiability of discrete word tokens makes the generator hard to improve. Consequently, most models try to utilize Reinforcement Learning (RL) Sutton [19] to fine tune their trained models Yu [20] or use educator driving technique to transfer the style. To get rid of the style of the original sentence Prabhumoye [4] proposed a back translation method to restructure the sentence and then give rise to a sentence which is specifically based on the content using a distinct style generator. They begin with learning the representation of the grounded source sentence so as to preserve the context of the source sentence and bring down the style properties. To match the required output style they used adversarial generation techniques based evaluation on different style transformations, i.e., political slant, gender and sentiment. They showed an improvement in both automatic evaluation of style transfer and manual evaluation of preserving the context and fluency of the sentence. However, in this method the generated sentence failed to preserve the detailed content information present in the source.

A perception was made that text styles are frequently set apart by unique sentences (e.g. "a great fun") in Li et al. (2018). By removing the phrases which are correlated with source style of the sentences, the authors proposed to first fetch the content words. Secondly, the new phrases were recaptured related with the target style and then used a neural network model to merge them into final output smoothly. However, in (Tetreault and Pavlick, 2016), authors demonstrated that on some instances it was difficult to clearly distinguish the content and style using only the phrases boundaries.

Yang [21] proposed to use a language model as discriminator where it utilizes a target language model from the domain to give token-level assessment during learning. The language model was trained as a discriminator in a way that real sentence gets higher probability, replacing standard binary classifier. Hu [22] presented a model where the objective was to generate sentences with controllable styles by latent representations. In their work, they integrate variation auto encoder (VAEs) with holistic attribute discriminators to impose smooth and effective structure on semantic

representation. Also, alternatively their model utilized wake-sleep algorithms intensify variation encoder for using fake data as training samples. Santos [23] is most relevant to our work. The objective of their model was to not only classify the offensive post on social media like Twitter and Reddit but to restructure the post from offensive to non offensive ones. They proposed a methodology for training encoder-decoder with non-parallel data which is incorporated with a collaborative classifier.

Motivation

These are the following observation we made after going through the work that was done in the field of style transfer which motivated us to study the task of hate speech detection and normalization.

- Assumption behind most of the work that are done for hate speech is to detect and remove the complete offensive text.
- Not a single work is done on detecting and normalizing the hate against women and children in an Indian context. • Classifier that were used in most of the work is based traditional Convolutional Neural Network (CNN).

DATASET

In this work, we introduce a new hate speech dataset which is India centric and focuses on women and child safety to identify hate against women and children. Also, a brief overview of existing hate speech datasets is given in Table I

Data Collection

Tweets were collected from Twitter streaming and search API which helped us fetch the real time and past tweets based on specific hashtags, keywords, accounts etc. We fetched approximately 53k tweets from August, 2020 to

TABLE I
AN OVERVIEW OF EXISTING HATE SPEECH DATASET

Dataset	Availability	labels	Distributions
BURNAP [24]	NO	Sexual Race Disability Religion	-
WASEEM [25]	YES	Racism Sexism Neither	11.69% 20.00% 68.33% 16k tweets
DAVIDSON [26]	YES	Hate Speech Offensive Neither	5.77% 77.43% 16.80% 24k tweets
FOUNTA [27]	YES	Abusive Hateful Spam Normal/None	11% 7.5% 22.5% 59% 80k tweets
WARNER [28]	NO	Anti-Semitic Anti-Black Anti-Asian Anti-Women Anti-Muslim Anti-Immigrant Other hate	- 9k paragraph
DJURIC [29]	NO	Hate Speech Clean	5.91% 94.08% 951k Comments
NOBATA [30]	NO	Abusive Clean	7%+16% 3.4% +10% of F and N
QIAN (Reddit) [31]	YES	Hate Speech Not-Hate Speech	23% 76.5% 22l comments
QIAN(Gab) [31]	YES	Hate Speech Not-Hate Speech	43.2% 51.8% 33k Comments
ROSS [32]	YES	6 point Likert Scale	- 541 tweets
BENIVOKA [33]	YES	Hate Speech Not Hate Speech	33% 67% 36 tweets
TULKENS [34]	NO	No hate Weak Hate Strong hate	- 6k Comments
VIGNA [35]	NO	Racist Non-Racist	- 7k Comments
MMHS150K [36]	YES	Racist Sexist Homophobic Religion Others	- 150k tweets

December, 2020 using some specific hashtags like Hathras, MeToo, KathuaRapeCase, Women-Empowerment, DomesticViolence, Rape, HathrasHorrorShocksIndia, Nirbhya, TrippleTalaq, childLabour etc.

Data Pre-processing

Pre-processing is a much needed step before approaching any Natural language processing task because tweets in raw form are not very useful and it needs to be processed which will not only provide some meaning to it but will also make the annotation work smooth and for that we built a channel which will perform the following things on raw tweets:

- Eliminate all unwanted links.
- Striped symbols like ¡", "¿", "@", "RT" and emoticons.
- Eliminate Non-Ascii characters
- Removed extra, ending and leading spaces
- Put space in between punctuation marks
- Removed the tweets having length less than 5
- Transformed every words in lowercase
- Expanded the contracted words

Data Annotation

After Pre-processing the dataset is given to an annotator having proficiency in both Hindi and English to perform the annotation task and provide a Hate score ranging from 0 to 2 based on the context of the tweets. The tweets having a score 0(Not Hate) denotes there is no shred of evidence that tweet is spreading hate, Score 1(Hate) means tweets is spreading hate but it is not that extreme. And the tweets with score 2(Aggressive) denotes the percentage of hate against women or children is severe. For better annotation we provide some tweets with true labels along with explanation. Annotator understood the task clearly but also worried because of its subjectivity. This is undoubtedly a subjective task, profoundly reliant on the annotator sensitivity. Few examples of the annotated tweets are shown in Table II.

Data Statistics

After pre-processing the sample size of our dataset was 9162 in which 4812 tweets have a hate score of 0, 3268 tweets have a hate score of 1 and 1082 tweets have a hate score of 2. We know class balancing is needed when we care about the minority classes. So to make our dataset balanced we merge the tweets having a hate score of 1 and hate score of 2 which then make our dataset distribution to 4812 tweets have hate score 0 i.e. Not Hate and 4350 have a hate score 1 i.e. Hate. The percentage distribution of the Not Hate and Hate in our dataset are 53% and 47% respectively shown in figure 1, whereas earlier the distribution of Not Hate, Hate and Aggressive was 53%, 36% and 11% respectively. Statistic based on Not Hate and Hate is shown in Fig 2.

TABLE II
EXAMPLES FROM THE ANNOTATED DATASET

Tweets	Hate Score	Class
T1: get fucked cunt	2	Hate
T2 : We work for the development of poor and needy peoples and to provide our best help to the acid attack victim .	1	Not Hate
T3 : Are didi ise child labour nhi kehte isse opportunity kehte hai ... so that they don't face such problems.	1	Hate
T4 : Well let me promote how deepika made fun of acid attack victim look end	1	Hate
T5 : From Islam! The religion is allowing paedophilia... Do you understand	2	Hate

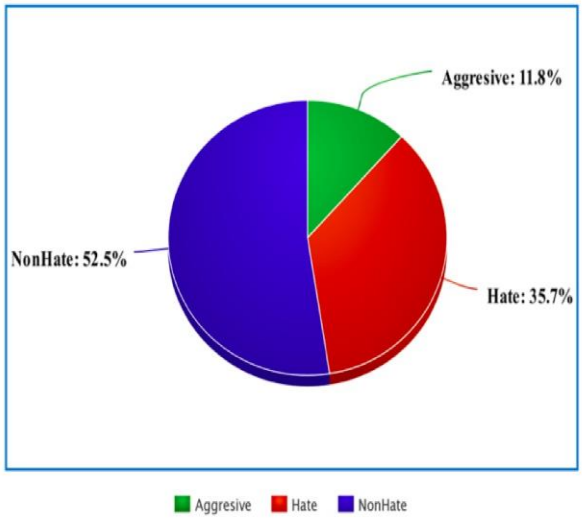

Fig. 1. Unbalanced class wise data distribution

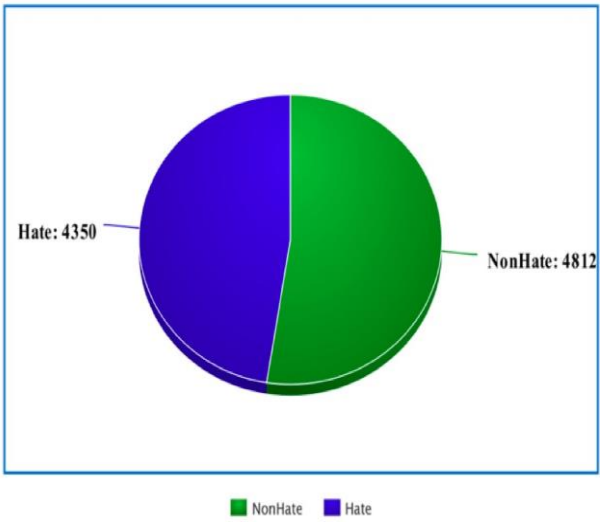

Fig. 2. Balanced class wise data distribution

PROBLEM STATEMENT

The end goal of this work is to detect hate tweets that are subjected towards women and children in India on Twitter, i.e., hate speech classification and normalizing those tweets so that it does not look offensive while preserving its context at the same time. Given a tweet X which can be of any style between c_1 and c_2, where c_1, c_2 represent Hate and Non-Hate tweets respectively, the task is to transfer the tweet having a style c_1 to c_2 while preserving the context and assign a label to X between hate and Not Hate.

METHODOLOGY

In this section, we discuss different components of our proposed methodology. Our proposed architecture of this work is show in Figure 5.

Encoder-Decoder Network

Encoder-decoder model is a way of using recurrent neural networks for sequence-to-sequence prediction problems and as the name suggest it consist of two parts. First is the encoder which encodes the input text token wise at each time stamp, fetch the information by processing it and forward it to the decoder. The decoder then predicts the output at each time step. Similarly, in our model encoder takes c_i and a_s^i i.e Encoder(c_i , a_s^i) as input where c_i = (c_1 or c_2). c_1 represents the context of the tweets that are "hate", c_2 represents the context of tweets that are "not hate" and a_s^i represent the sentence s of context c_i which is sampled randomly from our dataset. The output from the encoder is the sequence of the hidden states which act as an input to the decoder along with the desired style label c_j and outputs a sentence. $a^{(i->j)}_s$ that denotes the source sentence of context c_i is transferred to the style of a target sentence c_j. When the target sentence successfully preserves the context of the source sentence, i.e., i == j, then it is represented as $a^{(i->i)}_s$. Otherwise i! = j which is denoted as $a^{(i->j)}_s$.

Attention Mechanism

Attention mechanism found in neural networks is somewhat similar to the one found in humans. They focus on a certain part of input while the rest of the input is ignored because not each and every word is important at the given point of time. So Bahdanau [12] proposed an idea shown in figure 3 that along with considering all the input words at once to generate a context vector, an importance should also be assigned to each word so the moment model generates a sentence it searches for a position where the important information is available in the hidden states. Similarly, in our work the encoder takes c_i and a_s^i as input, i.e., Encoder(c_i , a_s^i) and generates an output B_s^i which is basically a sequence of hidden states. This B_s^i act as input to the decoder in the attention mechanism, where importance is assigned to a certain section of input tweets which at the end improve the quality of the normalized non hate tweet.

BERT Capsule

Bert Capsule [13] shown in Figure 4 is built on top of BERT. BERT is a bidirectional encoder representation from the transformer as the architecture forms after stacking the encoding layer of the transformer. It is pre trained on Book Corpus having 800M word and English Wikipedia having 2500M words and is fine tuned according to the task specific dataset which in this case is hate speech classification of tweet. To generate the representation for each token of a given tweet, Bert-Capsule does not consider [CLS] and [SEP] tokens and this generated representation first passes through the CNN which fetch and learn the abstract characteristic from n-gram and provides feature maps. These feature maps then act as an input to bidirectional LSTM which encodes these feature maps into the hidden states sequentially and leans semantic dependency based feature. Then this semantic dependency

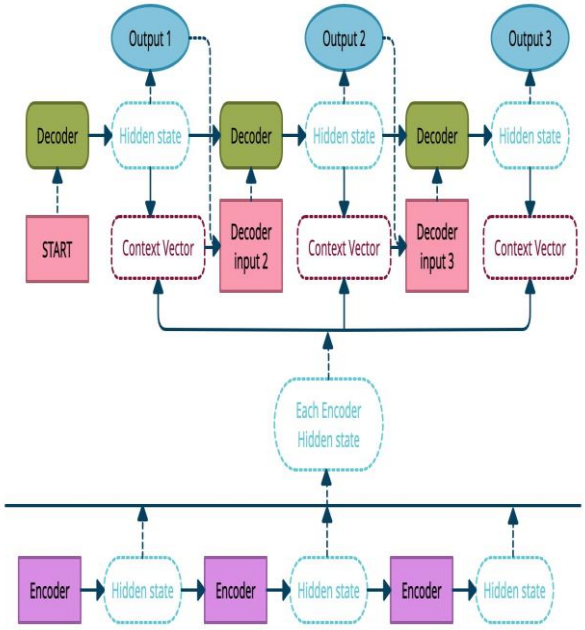

Fig. 3. Bahdanau Attention Mechanism

based feature is passed through a capsule layer which fragments the instantiated parts with another convolution operation and the local ordering of the words in tweets along with semantic is captured. Next, the capsule layer encompasses dynamic routing. The final capsule consists of a number of class capsules which depend on the number of class labels that exist in the dataset. In our work, we use this as a classifier which take the output from the decoder. $a_s^{i->j}$ signifies the target tweet is not able to preserve the context of the source tweet, i.e., $i! = j$ or $a_s^{i->i}$ representing the target tweet is able to preserve the context of the source tweet successfully, i.e., i == j is fed as input and outputs the probability over style labels representing how good the tweets are classified into hate and not hate. The target sentence that fails to preserve the context is transferred back to the source sentence. The reconstruction loss $a_s^{i->j->i}$ is computed again. Along with this, the accuracy is also controlled using the classifier.

Reconstruction loss is computed in both forward pass and the backward pass. In forward pass reconstruction loss is calculated between the encoded input sentence a_s^i and $a_s^{i->i}$ and this loss will signify how well the decoder constructed the target sentence. Reconstruction loss for the forward pass is calculated as below equation 1:

$$L_{fwd_recons} = Ea_s^i \; A[-logPG(a_s^i, c_i), c_i] \quad (1)$$

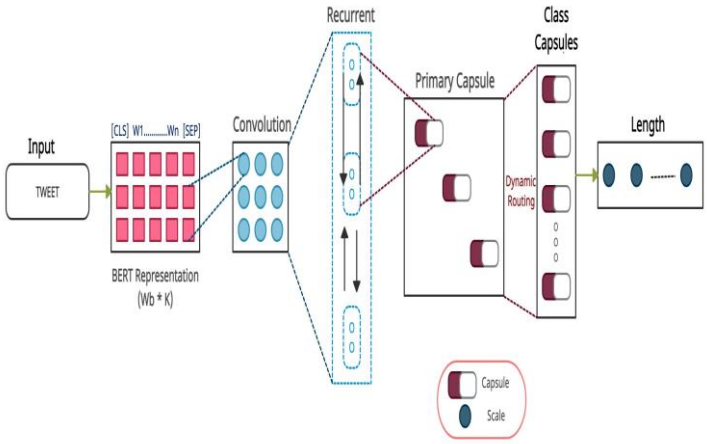

Fig. 4. BERT Capsule

Reconstruction loss in the backward pass is calculated when the target sentence as $a_s^{i->j}$ is transferred back to the original sentence represented by $a_s^{i->j->i}$ and then compared with a_s^i to preserve the context of the source sentence. Formally it can be represented as in Eqn 2

$$L_{bwd_recons} = Ea_s^i \; A[-logPG(a_s^i|E(a_s^{(i->j)}, c_j), c_j)] \quad (2)$$

After calculating the reconstruction loss, we then check the target sentence which is transferred to original sentence has correct labels or not and it is estimated as Eqn 3.

Lbwdclass = E(a(i−>j)s
A[−logPC(ci|G(E(a(si−>j),cj),cj]

(3)

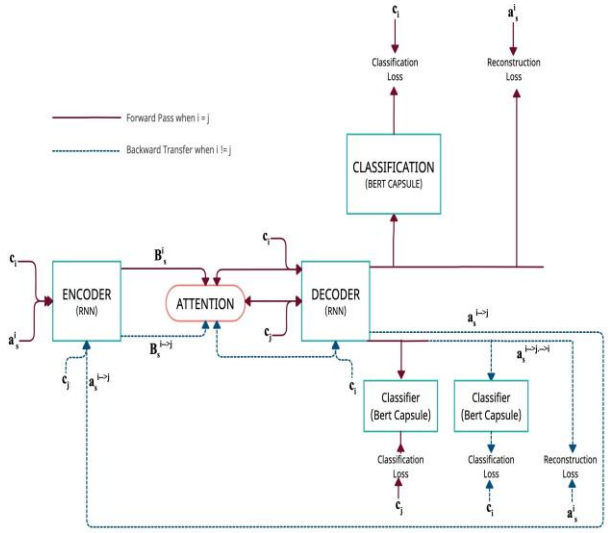

Fig. 5. Architecture of our proposed model

EXPERIMENTAL RESULTS AND ANALYSIS

This section represents the result of different models assessed on our hate speech dataset. All the experiments were executed 3 times on clutter or multiple GPUs having graphics Coprocessor NVIDIA GeForce RTX 2080 Ti and we report the average of three runs below.

Training Details

The distribution of training, validation and test set for our experiment is 80%, 10%, 10% respectively. Detailed class wise distribution of train, validation and test set are shown in Table III. For the experiments, we have used python 3.6, pytorch = 0.4.1, Torchtext =0.2.3 to implement all the deep learning models.

TABLE III
DATSET DISTRIBUTION

Distribution	Hate	Not Hate	Total
Training Set	3492	3837	7329
Validation Set	427	490	917
Test Set	431	485	916

Style Transfer baseline models

We have introduced these following models with and without attention which are thoroughly used in text processing to implement our objective of normalizing the hate tweets. Using BIEU Score [14], which is a metric for context preservation, we measured the quality of the tweets after the normalization of hate tweets.

- Recurrent Neural Network (RNN) (Baseline 1): Recurrent networks [37] are the type of artificial neural network designed to recognise the pattern in sequence of data. Input to the RNN is the output from the $t-1$ time step and a tokenized word of a sentence at time t.
- Bidirectional Long Short term Memory (Bi-LSTM) (Baseline 2): Bi-LSTM [37] [38] consists of two LSTM where one LSTM takes the input in forward path and second LSTM takes it in backward path. Bi-LSTM effectively improves the amount of information available to the network, which plays a major role in preserving the context of the input sentence.
- Bidirectional Gated Recurrent unit (Bi-GRU) (Baseline 3): Bi-GRU [38] also consists of two GRU where one GRU takes the input in forward path and second GRU takes it in backward path. The difference in Bi-GRU and Bi-LSTM is Bi-GRU has two gates, i.e., input and forget gates whereas Bi-LSTM has input, forget and output gates.
- Attention (Baseline 4): As the previous models under performed when it comes to preserving the context of input tweet after normalisation, we decided to add an attention mechanism [12] in the above mentioned model where the basic idea is not every word is important at a given point of time.

Baseline Classification model

We have introduced the following baseline classification model to compare with Bert-Capsule on our hate speech dataset and the result shown in Table V and few output samples in Table VI.

- LSTM + Glove Embedding (Baseline 1): As a baseline model, we used one layer bidirectional LSTM and the hidden size was 256. Using GloVe twitter embedding having a dimension 50, the Hate and NotHate tweets are embedded in the final layer. The LSTM cells are stacked up and connected to a softmax classifier via fully connected layer.
- BERT (Baseline 2): Here according to our hate speech dataset, we fine tune the basic BERT sequence classification model using BERT base uncased model having 12 transformer layers, 12 self attention heads with hidden size of 768.
- BERT+CNN (Baseline 3): In this model, we stack CNN over BERT and final hidden states of BERT act as input to the CNN layer. In the final layer, CNN is linked to a softmax layer through fully connected layer.
- BERT+LSTM (Baseline 4): In this model, we stack LSTM over BERT and final hidden states of BERT act as input to the LSTM layer. In the final layer,

TABLE IV
STYLE TRANSFER BASED RESULTS AND COMPARISON WITH BASELINE MODELS

Models	BLEU Score
RNN	0.3926
Bi-LSTM	0.4218
Bi-GRU	0.4371
RNN + Attention	0.4716
Bi-LSTM + Attention	0.5179
Bi-GRU + Attention	0.5417

LSTM is linked to a softmax layer through fully connected layer.

- BERT + GRU (Baseline 5): In this model, we stack GRU cells over BERT and final hidden states of BERT act as input to the GRU layer. In the final layer, GRU is linked to a softmax layer through fully connected layer.

TABLE V
CLASSIFICATION RESULTS AND COMPARISON WITH BASELINE MODELS

Models	Classification Accuracy	BLEU Score
Bi-GRU + Attention + LSTM + Glove	67.92	0.5409
Bi-GRU + Attention + BERT	72.16	0.5411
Bi-GRU + Attention + BERT + CNN	72.81	0.5414
Bi-GRU + Attention + BERT + LSTM	77.29	0.5416
Bi-GRU + Attention + BERT + GRU	79.63	0.5417
Bi-GRU + Attention + BERT Capsule	83.37	0.5417

Table IV shows the results of baseline models for style transfer where initially we implemented different versions of encoder-decoder networks and observed the results. Later, we integrated the attention mechanism to encoder-decoder models and found out that Bi-GRU + Attention model exceeds in terms of BLEU score [14]. Since Bi-GRU + Attention provide us the best result, we then merge several classification model one by one with it to classify the tweet into Hate and Not Hate and observed the results in Table V. Results show that our proposed model i.e., Bi-GRU + Attention + Bert Capsule outperforms all others classification models with an accuracy of 83.37%.

Error Analysis

We have manually checked tweets which were misclassified by the classifier in the classification part and the tweets whose context changed after style transfer. Below are the examples.

Style Transfer: "Source Tweet [ST]: Reforming the custody courts will directly save children who are now forced to live with their abusers". Normalised Tweet [NT]: "Reforming the custody courts will directly save the children who are now love to live with their caretaker". In this example, after the style transfer the context of the target sentence completely changes when compared to the source sentence as source sentence is talking about saving the children who are living with one's who hate them. But according to the target sentence, court will save the children who love to live with the ones who care about them. The possible reason for this is the size of the dataset as we have used approximately 7000 tweet to train our model.

Classification: Tweet: "Reforming the custody courts will directly save children who are now forced to live with their abusers". The true label of this tweet is Not Hate but model predicted it as Hate. The possible reason could be the presence of the word abusers, forced, custody signifies that our model didn't get the context of sentence. The reason this happened is because of the unavailability of the contextual dataset.

TABLE VI
OUTPUT SAMPLES

Hate and Normalized Tweets
ST: she fucked up, her response was too faggot NT: she messed up, her response was overly pansy
ST: By going at JNU deepika itself prove that she is nothing but a pretty face cunt NT: By going at JNU deepika itself prove that she is a pretty face women
ST: Mumbai commissioner doesn't give a shit what that loud mouth kangana is saying NT: Mumbai commissioner doesn't give a damn what that loud mouth kangana is saying
ST: sexual assault jokes?? bitch what the fuck is this. you don't joke about rape NT: sexual assault jokes?? bimbo what the hell is this. you don't joke about rape

CONCLUSION

This work is the first step in the direction of normalizing the hate spreading against women and children in India virtually. In this paper, we propose a new dataset for fighting against hate speech by detecting and normalizing it. Following this, we propose a model based on encoder-decoder architecture with bahdanau attention for normalizing hate tweets and Bert Capsule classifier for classifying hate speech tweets. We compare our proposed approach with several normalization and classification baseline models. Our model attained an overall BLEU score and classification accuracy of 0.5417 and 83.37% respectively.

In future, attempts can be made to increase the model performance in preserving the context after normalizing the hate tweet by improving the dataset. We will also enhance our model by extending the domain which can lead to better context preservation and classification of the tweets. In particular, it would be interesting to add multi-modality concepts in which we will consider both image and text to conclude whether the post is hate or not hate because at times only text is not enough to deduce that this particular post is offensive.

REFERENCES

[1] B. Standard. (2020) 58% young females on social media have faced harassment.[Online].Available:https://wap.businessstandard.com/article-amp/current-affairs/58young-females-on-social-media-have-faced-harassment/

[2] H. Times. (2018) Why India should make online child safety a priority. [Online].Available:https://www.hindustantimes.com/analysis/whyindia-should-make-online-child-safety-a-priority/

[3] T. Hindu. (2020) Most online content on child sexual abuse from India. [Online].Available:https://www.thehindu.com/news/national/mostonline-content-on-child-sexual-abuse-from India/

[4] S. Prabhumoye, Y. Tsvetkov, R. Salakhutdinov, and A. W. Black, "Style transfer through back-translation," arXiv preprint arXiv:1804.09000, 2018.

[5] G. Xiang, B. Fan, L. Wang, J. Hong, and C. Rose, "Detecting offensive tweets via topical feature discovery over a large scale twitter corpus," in Proceedings of the 21st ACM international conference on Information and knowledge management, 2012, pp. 1980–1984.

[6] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in Proceedings of the second workshop on language in social media, 2012, pp. 19–26.

[7] I. Kwok and Y. Wang, "Locate the hate: Detecting tweets against blacks," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 27, no. 1, 2013.

[8] W. Wang, L. Chen, K. Thirunarayan, and A. P. Sheth, "Cursing in English on twitter," in Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing, 2014, pp. 415–425.

[9] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in Proceedings of the 25th international conference on world wide web, 2016, pp. 145–153.

[10] T. Davidson, D. Warmsley, M. Macy, and I. 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, 2017.

[11] A. M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis, "A unified deep learning architecture for abuse detection," in Proceedings of the 10th ACM conference on web science, 2019, pp. 105–114.

[12] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.

[13] T. Saha, S. R. Jayashree, S. Saha, and P. Bhattacharyya, "Bert-caps: A transformer-based capsule network for tweet act classification," IEEE Transactions on Computational Social Systems, vol. 7, no. 5, pp. 1168– 1179, 2020.

[14] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[15] T. Shen, T. Lei, R. Barzilay, and T. Jaakkola, "Style transfer from non-parallel text by cross-alignment," arXiv preprint arXiv:1705.09655, 2017.

[16] J. Zhao, Y. Kim, K. Zhang, A. Rush, and Y. LeCun, "Adversarially regularized autoencoders," in International conference on machine learning. PMLR, 2018, pp. 5902–5911.

[17] I. Melnyk, C. N. d. Santos, K. Wadhawan, I. Padhi, and A. Kumar, "Improved neural text attribute transfer with non-parallel data," arXiv preprint arXiv:1711.09395, 2017.

[18] Z. Fu, X. Tan, N. Peng, D. Zhao, and R. Yan, "Style transfer in text: Exploration and evaluation," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, 2018.

[19] R. S. Sutton, D. A. McAllester, S. P. Singh, Y. Mansour et al., "Policy gradient methods for reinforcement learning with function approximation." in NIPs, vol. 99. Citeseer, 1999, pp. 1057–1063.

[20] L. Yu, W. Zhang, J. Wang, and Y. Y. SeqGAN, "Sequence generative adversarial nets with policy gradient. arxiv e-prints, page," arXiv preprint arXiv:1609.05473, 2016.

[21] Z. Yang, Z. Hu, C. Dyer, E. P. Xing, and T. Berg-Kirkpatrick, "Unsupervised text style transfer using language models as discriminators," arXiv preprint arXiv:1805.11749, 2018.

[22] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing, "Toward controlled generation of text," in International Conference on Machine Learning. PMLR, 2017, pp. 1587–1596.

[23] C. N. d. Santos, I. Melnyk, and I. Padhi, "Fighting offensive language on social media with unsupervised text style transfer," arXiv preprint arXiv:1805.07685, 2018.

[24] P. Burnap and M. L. Williams, "Us and them: identifying cyber hate on twitter across multiple protected characteristics," EPJ Data science, vol. 5, pp. 1–15, 2016.

[25] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter," in Proceedings of the NAACL student research workshop, 2016, pp. 88–93.

[26] T. Davidson, D. Warmsley, M. Macy, and I. 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, 2017.

[27] A. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis, "Large scale crowdsourcing and characterization of twitter abusive behaviour," in Proceedings of the International AAAI Conference on Web and Social Media, vol. 12, no. 1, 2018.

[28] ——, "Large scale crowdsourcing and characterization of twitter abusive behaviour," in Proceedings of the International AAAI Conference on Web and Social Media, vol. 12, no. 1, 2018.

[29] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in Proceedings of the second workshop on language in social media, 2012, pp. 19–26.

[30] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, "Hate speech detection with comment embedding's," in Proceedings of the 24th international conference on world wide web, 2015, pp. 29–30.

[31] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in Proceedings of the 25th international conference on world wide web, 2016, pp. 145–153.

[32] J. Qian, A. Bethke, Y. Liu, E. Belding, and W. Y. Wang, "A benchmark dataset for learning to intervene in online hate speech," arXiv preprint arXiv:1909.04251, 2019.

[33] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki, "Measuring the reliability of hate speech annotations: The case of the European refugee crisis," arXiv preprint arXiv:1701.08118, 2017.

[34] Malik, M., Nandal, R., Dalal, S., Jalglan, V., & Le, D. N. (2021). Driving pattern profiling and classification using deep learning. Intelligent Automation & Soft Computing, 28(3), 887-906.

[35] Jindal, U., Dalal, S., Rajesh, G., Sama, N. U., Jhanjhi, N. Z., & Humayun, M. (2021). An integrated approach on verification of signatures using multiple classifiers (SVM and Decision Tree): A multi-classification approach.

[36] Seth, B., Dalal, S., Le, D. N., Jaglan, V., Dahiya, N., Agrawal, A., ... & Verma, K. D. (2021). Secure Cloud Data Storage System Using Hybrid Paillier–Blowfish Algorithm. Computers, Materials & Continua, 67(1), 779-798.

[37] Vijarania, M., Dahiya, N., Dalal, S., & Jaglan, V. (2021). WSN Based Efficient Multi-Metric Routing for IoT Networks. In Green Internet of Things for Smart Cities (pp. 249-262). CRC Press.

[38] Goel, M., Hayat, A., Husain, A., & Dalal, S. (2021). Green-IoT (G-IoT) Architectures and Their Applications in the Smart City. In Green Internet of Things for Smart Cities (pp. 47-59). CRC Press.

[39] Chawla, N., & Dalal, S. (2021). Edge AI with Wearable IoT: A Review on Leveraging Edge Intelligence in Wearables for Smart Healthcare. Green Internet of Things for Smart Cities, 205-231.

[40] Dahiya, N., Dalal, S., & Jaglan, V. (2021). Efficient Green Solution for a Balanced Energy Consumption and Delay in the IoT-Fog-Cloud Computing. In Green Internet of Things for Smart Cities (pp. 113-123). CRC Press.

[41] Dahiya, N., Dalal, S., & Jaglan, V. (2021). Mobility Management in Green IoT. In Green Internet of Things for Smart Cities (pp. 125-134). CRC Press.

[42] Seth, B., Dalal, S., & Dahiya, N. (2021). Practical Implications of Green Internet of Things (G-IoT) for Smart Cities. In Green Internet of Things for Smart Cities (pp. 61-81). CRC Press.

[43] Dalal, S., Agrawal, A., Dahiya, N., & Verma, J. (2020, July). Software Process Improvement Assessment for Cloud Application Based on Fuzzy Analytical Hierarchy Process Method. In International Conference on Computational Science and Its Applications (pp. 989-1001). Springer, Cham.

[44] Seth, B., Dalal, S., Jaglan, V., Le, D. N., Mohan, S., & Srivastava, G. (2020). Integrating encryption techniques for secure data storage in the cloud. Transactions on Emerging Telecommunications Technologies.

[45] Hooda, M., & Shravankumar Bachu, P. (2020). Artificial Intelligence Technique for Detecting Bone Irregularity Using Fastai. In International Conference on Industrial Engineering and Operations Management Dubai, UAE (pp. 2392-2399).

[46] Arora, S., & Dalal, S. (2019). An optimized cloud architecture for integrity verification. Journal of Computational and Theoretical Nanoscience, 16(12), 5067-5072.

[47] Arora, S., & Dalal, S. (2019). Trust Evaluation Factors in Cloud Computing with Open Stack. Journal of Computational and Theoretical Nanoscience, 16(12), 5073-5077.

[48] Shakti Arora, S. (2019). DDoS Attacks Simulation in Cloud Computing Environment. International Journal of Innovative Technology and Exploring Engineering, 9(1), 414-417.

[49] Shakti Arora, S. (2019). Integrity Verification Mechanisms Adopted in Cloud Environment. International Journal of Engineering and Advanced Technology (IJEAT), 8, 1713-1717.

[50] Sudha, B., Dalal, S., & Srinivasan, K. (2019). Early Detection of Glaucoma Disease in Retinal Fundus Images Using Spatial FCM with Level Set Segmentation. International Journal of Engineering and Advanced Technology (IJEAT), 8(5C), 1342-1349.

[51] Sikri, A., Dalal, S., Singh, N. P., & Le, D. N. (2019). Mapping of e-Wallets With Features. Cyber Security in Parallel and Distributed Computing: Concepts, Techniques, Applications and Case Studies, 245-261.

[52] Seth, B., Dalal, S., & Kumar, R. (2019). Hybrid homomorphic encryption scheme for secure cloud data storage. In Recent Advances in Computational Intelligence (pp. 71-92). Springer, Cham.

[53] Seth, B., Dalal, S., & Kumar, R. (2019). Securing bioinformatics cloud for big data: Budding buzzword or a

glance of the future. In Recent advances in computational intelligence (pp. 121-147). Springer, Cham.

[54] Jindal, U., & Dalal, S. (2019). A hybrid approach to authentication of signature using DTSVM. In Emerging Trends in Expert Applications and Security (pp. 327-335). Springer, Singapore.

[55] Le, D. N., Seth, B., & Dalal, S. (2018). A hybrid approach of secret sharing with fragmentation and encryption in cloud environment for securing outsourced medical database: a revolutionary approach. Journal of Cyber Security and Mobility, 7(4), 379-408.

[56] Sikri, A., Dalal, S., Singh, N. P., & Dahiya, N. (2018). Data Mining and its Various Concepts. Kalpa Publications in Engineering, 2, 95-102.

[57] Sameer Nagpal, S. (2018). Analysis of LrMu Power Algorithm in the Cloud Computing Environment using CloudSim Toolkit. International Journal of Research in Electronics and Computer Engineering (IJRECE), 6(3), 1175-1177.

[58] Nagpal, S., Dahiya, N., & Dalal, S. (2018). Comparative Analysis of the Power Consumption Techniques in the Cloud Computing Environment. Journal Homepage: http://www. ijmra.us, 8(8), 1.

[59] Kumar, N., Dalal, S., & Dahiya, N. (2018). Approach of Lion Optimization Algorithm for Efficient Load Balancing in Cloud Computing. Journal Homepage: http://www. ijmra.us, 8(8), 1.

[60] Sameer Nagpal, S. (2018). Comparison of Task Scheduling in Cloud Computing Using various Optimization Algorithms. Journal of Computational Information Systems, 14(4), 43-57.

[61] Arora, S., & Dalal, S. (2018). Hybrid algorithm designed for handling remote integrity check mechanism over dynamic cloud environment. International Journal of Engineering & Technology, 7(2.4), 161-164.

[62] Kukreja, S., & Dalal, S. (2018). Modified drosophila optimization algorithm for managing re-sources in cloud environment. International Journal of Engineering & Technology, 7(2.4), 165-169.

[63] Jindal, U., Dalal, S., & Dahiya, N. (2018). A combine approach of preprocessing in integrated signature verification (ISV). International Journal of Engineering & Technology, 7(1.2), 155-159.

[64] Nagpal, S., Dahiya, N., & Dalal, S. (2018). Comparison of Task Scheduling in Cloud Computing Using various Optimization Algorithms. Journal of Computational Information Systems ISSN, 1553-9105.

[65] Jindal, U., Dalal, S., & Dahiya, N. (2018). A combine approach of preprocessing in integrated signature verification (ISV). International Journal of Engineering & Technology, 7(1.2), 155-159

[66] Shakti Arora, S. (2018). Resolving problem of Trust context in Cloud Computing. International Journal of Engineering Research in Computer Science and Engineering (IJERCSE), 5(1), 138-142.

[67] Dalal, S., Dahiya, N., & Jaglan, V. (2018). Efficient tuning of COCOMO model cost drivers through generalized reduced gradient (GRG) nonlinear optimization with best-fit analysis. In Progress in Advanced Computing and Intelligent Engineering (pp. 347-354). Springer, Singapore

[68] Seth, B., & Dalal, S. (2018). Analytical assessment of security mechanisms of cloud environment. In Progress in Advanced Computing and Intelligent Engineering (pp. 211-220). Springer, Singapore.

[69] Kukreja, S., & Dalal, S. (2018). Performance analysis of cloud resource provisioning algorithms. In Progress in Advanced Computing and Intelligent Engineering (pp. 593-602). Springer, Singapore.

[70] Rani, U., Dalal, S., & Kumar, J. (2018). Optimizing performance of fuzzy decision support system with multiple parameter dependency for cloud provider evaluation. Int. J. Eng. Technol, 7(1.2), 61-65.

[71] Dahiya, N., Dalal, S., & Khatri, S. (2017). An Enhanced Bat Algorithm for Data Clustering Problems. International Journal of Advanced Research in Computer Science, 8(3).

[72] Dahiya, N., Dalal, S., & Khatri, S. (2017). Data clustering and its Application to numerical function optimization algorithm. International Journal of Advanced Research in Computer Science, 8(1).

[73] Arora, S., & Dalal, S. (2017). Adaptive Model For Integrity Verification In Cloud Computing System. International Journal of Advanced Research in Computer Science, 8(1), 233-236.

[74] Neeraj Dahiya, S. (2017). Numerical Function Optimization: Model, Procedure And Uses. International Journal of Engineering Science and Technology (IJEST), 9(4), 266-270.

[75] Dahiya, N., Dalal, S., & Khatri, S. (2016). Refinement with Image clustering using Self-Organizing Map and Numerical Function Optimization. International Journal of Computer Science and Information Security, 14(11), 909.

[76] Neeraj Dahiya, S. (2016). A Review on Numerical function optimization Algorithm and its Applications to Data Clustering & Classification. International Journal of Recent Research Aspects, 3(3), 115-121.

[77] Arora, S., & Dalal, S. (2016). Novel Approach of Integrity Verification in Dynamic Cloud Environment. International Journal of Computer Science and Information Security, 14(8), 207.

[78] D. Benikova, M. Wojatzki, and T. Zesch, "What does this imply? examining the impact of implicitness on the perception of hate speech," in International Conference of the German Society for Computational Linguistics and Language Technology. Springer, 2017, pp. 171–179.

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

[80] R. Gomez, J. Gibert, L. Gomez, and D. Karatzas, "Exploring hate speech detection in multimodal publications," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1470–1478.

[81] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," Physica D: Nonlinear Phenomena, vol. 404, p. 132306, 2020.

[82] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," IEEE transactions on Signal Processing, vol. 45, no. 11, pp. 2673–2681, 1997.

