TRANSFORMER BASED ENSEMBLE LEARNING TO HATE SPEECH DETECTION LEVERAGING SENTIMENT AND EMOTION KNOWLEDGE SHARING

Prashant Kapil and Asif Ekbal

Department of Computer Science and Engineering, IIT Patna, India

ABSTRACT

In recent years, the increasing propagation of hate speech on social media has encouraged researchers to address the problem of hateful content identification. To build an efficient hate speech detection model, a large number of annotated data is needed to train the model. To solve this approach we utilized eleven datasets from the hate speech domain and compared different transformer encoder-based approaches such as BERT, and ALBERT in single-task learning and multi-task learning (MTL) framework. We also leveraged the eight sentiment and emotion analysis datasets in the training to enrich the features in the MTL setting. The stacking based ensemble of BERT-MTL and ALBERT-MTL is utilized to combine the features from best two models. The experiments demonstrate the efficacy of the approach by attaining state-of-the-art results in all the datasets. The qualitative and quantitative error analysis was done to figure out the misclassified tweets and the effect of models on the different data sets.

KEYWORDS

BERT, Multi-task learning, Hate speech, Transformer, Ensemble.

1. INTRODUCTION

The majority of the post on the social media platform are harmless but some express hatred towards a targeted individual or any group based on some attributes such as religion, nationality, colour, gender, nationality, ethnicity, etc. These posts have detrimental effects on their victims, e.g., victims are more likely to have lower self-esteem and a tendency of suicidal thoughts **[1].** The violence due to hate speech has increased worldwide. The USA has seen an increase in hate speech and related violence following the Presidential election. Therefore Governments and social media platforms must build an efficient tool to combat this issue. To detect online hate speech a large number of scientific studies have been done leveraging Machine learning and Deep learning methods. The trend has been shifted to deep learning architectures for feature extraction and training of the classifiers to enhance the performance but they still lack a sufficient number of labelled data. Recently pre-trained language model BERT has shown substantial and consistent improvement in solving the task. Therefore in this paper, we investigated the effects of transferring knowledge from BERT, and ALBERT to distinguish different hate posts trained in single-task learning , multi-task learning paradigm, and stacking of MTL. The semantics of hate speech often contains negative sentiment that is correlated to hate. The effective features from other sources can be used to enhance the performance **[2][3]**. We are also providing the different definitions of hate used in the existing literatures to collect the data in Table 1. The laws by

David C. Wyld et al. (Eds): CONEDU, CSITA, MLCL, ISPR, NATAP, ARIN - 2022
pp. 179-194, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121014

different countries regarding the hate speech is in Table 2. The significant contributions of this work are as follows:

Dataset: We utilized eleven bench mark datasets related to hate domain, harassment, aggressiveness, offensiveness, abusive, spam, racist, sexist, etc. Due to high correlation with the sentiment and emotion data we also utilized three sentiment analysis and five emotion data sets that are publicly available.

Model: We investigated the various state of the art models such as BERT,ALBERT in single task learning and multi task learning framework. The sentiment data is also leveraged in multi task training framework. The stack based ensemble of MTL with BERT and ALBERT as the shared encoder trained on the hate, sentiment, and emotion data is utilized.

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 is described. Section 4 discusses in detail the proposed methodology and Experimental setup. Section 5 reports the evaluation results and comparisons to the state-of-the-art. Error analysis containing qualitative and quantitative analysis of the obtained results is presented in Section 6. Finally, the conclusion and directions for future research are presented in Section 7.

Table1. Definitions of hate to collect data

Authors	Definition
[4]	a language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate or insult the members of the group
[5]	a speech that denigrates any person or any group based on characteristics like race, color, gender, religion, ethnicity, nationality, sexual preferences etc.
[6]	A tweet is offensive if it contains racist or sexist slur, intention to attack, promote violent crimes, threatening minorities, and stereotyping genders.
[7]	It is a bias-motivated hostile speech aimed at a person or group of people with intentions to injure, dehumanize, harass, degrade and victimizing 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 country on hate speech.

Country	Law
USA	Hate speech is legally protected free speech under the First Amendment. However, speech that include obscenity, speech integral to illegal conduct, speech that incites lawless action or 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 Germany 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 population or using insulting malicious slur 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 restriction 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 contents in any form, words that are likely to create hostility against a group of people on the basis of race, color, ethnicity is unlawful.

2. RELATED WORK

The state-of-the-art approaches try to solve this problem by supervised learning. These methods can be divided into two parts.

A. **Classical methods: [4]** created unigram, bigram, and trigram weighted by TF-IDF. The syntactic structure is captured by the Part of speech (POS) tag. The sentiment score and readability of all the tweets along with surface-level features were merged to fed into a logistic regression, naive bayes, support vector machine, decision trees, and random forests. In **[5]** the features included were unigram, bigram, trigram, and four grams for each tweet, and user-based features such as location, and gender is fed into a logistic regression to solve the task. **[6]** leveraged characters 3-5 grams, unigram, and bigrams as the n-gram features. The linguistic features along with syntactic features such as POS and dependency relations to detect hate speech.

B. **Deep neural networks**: In the last 5 years the neural network-based approaches outperformed the traditional classical methods as the former can capture more abstract features helpful in the classification. **[7]** proposed a CNN_GRU where the first layer is a word embedding layer. The features from the embedding layer followed by the dropout are fed into 1D Convolutions with 100 filters and a window size of 4. The extracted features from the CNN are fed into the GRU for the final classification. **[8]** proposed a deep learning architecture that utilizes the user features and network features combined with automatically extracted hidden patterns within the text of tweets. **[9]** investigated deep neural networks, namely Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) by initializing word embeddings with random embedding, FastText word embeddings **[10]** and GloVe word embeddings **[11**] using data by **[5]** . **[12]** constrained their work to binary classification between *abusive* and *not abusive*. Their character-based approach outperformed token-based and distributional-based features on the dataset by **[6]. [13]** trained four CNN models, based on character 4-grams, word vectors based on semantic information built using word2vec, randomly generated word vectors, and word vectors combined with character n-grams utilizing the dataset by **[5].**

[14] considered *SentiWordNet, Affinn, Bing Liu, General Inquirer, Subjectivity clues* and *NRC* to explore the relationship between sentiment and toxicity in social media messages from 3 domains, namely Reddit, Wikipedia talk labels and Toxic comment classification. The toxicity detector is a Bi-GRU layer with words represented by 300d FastText pre-trained word embeddings **[10]** characters represented by 60 dimensions one-hot vector and 3 sentiment values obtained from 3 best lexicons based on their study. These input values are then concatenated together into a vector of 363 dimensions. **[15]** proposed a transfer learning approach advantaging the pre-trained language model BERT to enhance the performance of the hate speech detection system and generalize it to new datasets. **[16]** proposed a hybrid methodology to infuse external knowledge into a supervised model for abusive language detection. The external knowledge is lexical features with BERT at the sentence or term level. Transformer-based BERT outperforms traditional deep neural networks in all the tasks. In the semeval task6, 7 out of the top 10 teams used BERT with variations in the parameters and the pre-processing steps. **[17]** described the architecture of BERT-CNN which utilizes the merged output of the last four layers of BERT to pass into the convolution layer.

Table 3. Statistics of hate data used in the experiment

Datasets	Training	Testing	Inter Annotator Agreement score
D1	Hate: 1430, Offensive:19190 Neutral: 4163	Cross Validation	0.92
D2	Racism:1923, Sexism:2871 Neutral:10682	Cross Validation	0.84
D3	OAG: 3419, CAG: 5297 NAG: 6285	Cross Validation	0.72
D4	Offensive:4400, Non-Offensive:8840	Cross Validation	0.83
D5	Harassment: 5285, Neutral: 15075	Cross Validation	0.84
D6	HOF: 2261, NOT: 3591	HOF: 288, NOT: 865	0.61
D7	Hate: 4210, Neutral: 5790	Hate: 1260 Neutral: 1740	0.62
D8	Hate: 1097, Neutral: 8571	Cross Validation	0.36
D9	OAG: 548, CAG: 570 NAG: 4211	OAG: 286, CAG: 224 NAG: 690	0.69
D10	HOF: 2501, NOT: 1342	HOF: 483, NOT: 798	--
D11	Hate: 4965, Normal: 53851 Spam : 14030, Abusive: 27150	Cross Validation	0.70

3. DATA SETS

In this section we will be briefly describe all the 11 datasets related to the hate domain used in this paper. The statistics of all the hate related data is in Table 3.

3.1. Hate Domain Data

Data 1(D1) [4]: They begin with a hate speech lexicon containing words and phrases identified by internet users as hate speech, compiled by hatebase.org. Using Twitter API, 33458 user data were crawled. A random sample of 25K tweets was manually coded by Crowdsource workers. The tweet was categorized into hate, offensive, and neutral. The intercoder-agreement score provided by the CF is 92%. Only 5% of tweets were tagged as hate by the majority of coders and only 1.3% were coded unanimously, demonstrating the imprecision of the hate lexicon. While f*g, b***h, n**ga are used in both offensive and hate speech the terms f**got and n**ger is generally associated with hate speech. Many of the tweets considered most hateful contain multiple racial and homophobic slurs.

Data 2(D2) [5] : The data consists of tweets collected over 2 months. In total 16914 tweets were annotated into racism, sexism, and neutral out of 136052 tweets. The corpus is collected by performing an initial manual search of slurs and terms used about religious, sexual, ethnic minorities, and gender. They presented a list of criteria based on critical race theory to identify racist and sexist slurs. The inter-annotator agreement score is 0.84. 85% of all disagreements occur in the annotation of sexism.

Data 3(D3) [18]: 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 4(D4) [19] : They compiled the Offensive Language Dataset(OLID), where the tweets were annotated using a fine-grained three-layer annotation schema. They retrieved the examples in OLID from Twitter using API and searched for the keywords and constructions often included in 'she is or 'to"Breitbart news. Some of the keywords leveraged are ANTIFA, MAGA, liberals, conservatives, etc. The full datasets consist of 50\% tweets from political and 50\% tweets from non-political keywords. The Fleiss Kappa score is 83\% for the first layer.

Data 5(D5) [20]: It introduces a hand-coded corpus of online harassment data of 35K tweets. It has 15% harassment and 85% non-harassment tweets. They collected a sample of tweets from the blocked user in the Block together. The list terms such as #white genocides, #fuckniggers, #whitepower, #whitelivesmatter, #fucking faggot, #the Jews, etc were searched. Each tweet was labeled by 2 annotators, where the third coder is to break the tie of 2711 tweets. The cohen kappa score is 0.84.

Data 6(D6) [21] : The content was scrapped from Storm front using web-scraping techniques. The extracted forum content was published between 2002 and 2017.A subset of 22 sub-forums covering diverse topics and nationalities is randomly sampled to gather individual posts uniformly distributed among sub-forums and users. The average percentage agreement, Cohen's kappa coefficient, and Fleiss kappa coefficient are 91.03%, 61.4%, and 60.7% respectively. The most occurring hateful words were ape, scum, savages, filthy, mud, homosexuals, etc.

Data 7(D7) [22] : The data have been collected using different gathering strategies in the period from July to September 2018. The different approaches to collecting the tweets are (1) monitoring potential victims of hate accounts, (2) downloading the history of identified haters, and (3)filtering Twitter streams with keywords, i.e. words, hashtags, and stems. The frequent occurring keywords were migrants, refugee, #buildthatwall, bitch, hoe, and women.

Data 8(D8) [23]: the authors searched with heuristics 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 0.36.

Data 9}(D9) [24] : The sampling of the datasets was planned 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. The trending hashtags used to sample the tweets were #resignmodi, #TMCTerror, #chinesevirus, #islamophobia, #covidvaccine, #IndiaCovidcrisis, etc. The inter-annotator agreement score is 69%.

Data 10(D10) [25]: 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 11(D11) **[26]**: The first step is to collect random tweets by utilizing Twitter API. They collected all the tweets provided by the API over 10 days, consisting of 32 million in total. They store the data in elastic search and basic filtering techniques. They also applied simple text analysis and machine learning to create a boosted set of tweets that will be used to improve the coverage of the minority classes. Finally, they randomly sampled a small data D1 for the exploratory analysis and the remaining D2 for the large scale annotation.

Table 4. Statistics of emotion and sentiment data used in the experiment

Authors	Labels	Total
Kaggle Airline data	Positive, Negative, Neutral	14640
[27]	Positive, Negative, Neutral	20632
[28]	Ekman's Emotion	21051
[29]	Ekman's Emotion	7665
[30]	Ekman's Emotion	13118
[31]	Ekman's Emotion	7303
[32]	Sadness, joy	2585
[33]	Positive, Negative, Neutral	63192

3.2. Sentiment and Emotion Data

Table 4 consists of the sentiment and emotion datasets used for our experiment. We have utilized three sentiment data tagged into positive, negative, and neutral whereas five emotion data is being used tagged based on Ekman's emotion fear, anger, joy, sadness, disgust, and surprise.

4. METHODOLOGY

4.1. Pre-processing

Social media posts contain a lot of noisy texts which are not considered as useful features for the classification. We perform the following steps to remove the noise, and make it ready for machine learning experiments:

1. All the characters like |:,? were removed along with the numbers and URLs.
2. 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.
3. Word segmentation is being done using the Python based word segment to preserve the important features present in hashtag mentions.
4. All the emoticons were categorized into 5 categories, namely *love*, *sad*, *happy*, *shocking* and *anger*. The unicode character of emoticon in text is substituted with one category.
5. All the @ (ex.@abc) mentions were replaced with the common token, i.e *user*.
6. 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.
7. 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 7 transformer based approaches which are discussed in this section.

4.2. Models

1.**Model 1(M1):BERT [34]** : It stands for Bidirectional Encoder representations from Transformer is designed to pre-train deep bidirectioanl representations from unlabeled text by jointly conditioning on on both left and right context in all layers. There are two steps in this framework: pre-training and fine-tuning. During pre-training, the model is trained on unlabeled data over pre-training tasks. For the fine-tuning, the BERT model is initialized with the pre-trained parameters and all of the parameters are fine-tuned using labeled data from the downstream tasks. The pre-training of BERT is done by two unsupervised tasks.

Masked LM: This method masks 15% of wordpiece tokens in each sequence at random. The final hidden vectors corresponding to the masked token are fed into an output softmax over the vocabulary. The objective is to predict the masked words rather than reconstructing the entire sentence.

Next Sentence Prediction: In this, the model is trained to understand sentence relationship by pretraining for a binarized next sentence prediction task. When choosing the sentences A and B for each training example 50% of the time B is the actual sentence that follows A and 50% of the time it is the random sentence from the corpus.

In the fine-tuning the task specific inputs and outputs are fed into BERT and all the parameters are fine-tuned end to end. At the output, the CLS representation is fed into an output layer for classification.

2**. Model 2(M2) ALBERT [35] :.** The design choice of ALBERT uses three new techniques over BERT.

(i) Factorized embedding parameterization: In BERT and RoBERTa the word piece embedding size E is equal to hidden layer size E=H. The word piece embeddings learn context independent representations whereas the hidden layers capture context dependent representations. ALBERT in order to make more efficient usage of the total model parameters dictate the H>>E.

(ii) Cross layer parameter sharing: There are multiple ways to share the parameters (i) sharing the feed forward network across layers or only sharing attention parameters. The ALBERT share all parameters across the layers to improve parameter efficiency.

(iii) Inter sentence coherence loss: They propose a loss based on coherence which is sentence order prediction loss that avoids topic prediction and focuses on modelling inter sentence coherence.

The general architecture of transformer encoder block is in Figure 1.

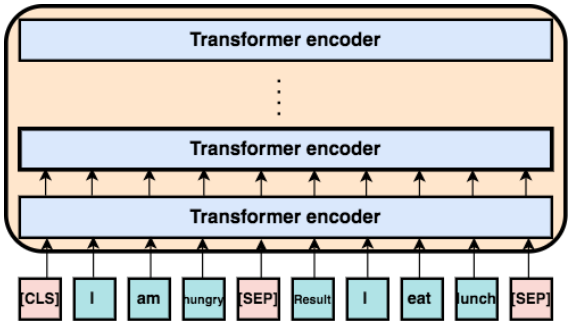

Figure 1. Transformer Encoder

4.3. Multi-Task Learning

Multi-tasking learning aims at solving more than one problem simultaneously. The 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 generalize better by achieving better performance for the individual tasks.

[36] 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 for the purpose of improving the new task's learning performance.

[37] 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 own specific layers with some sharable part.
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

The architecture of the MTL-DNN is shown in Figure 2. The lower layers are shared across all the tasks, while the top layers represent task-specific outputs. In our experiment all the tasks are classification. 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.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 Devlin et al 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.

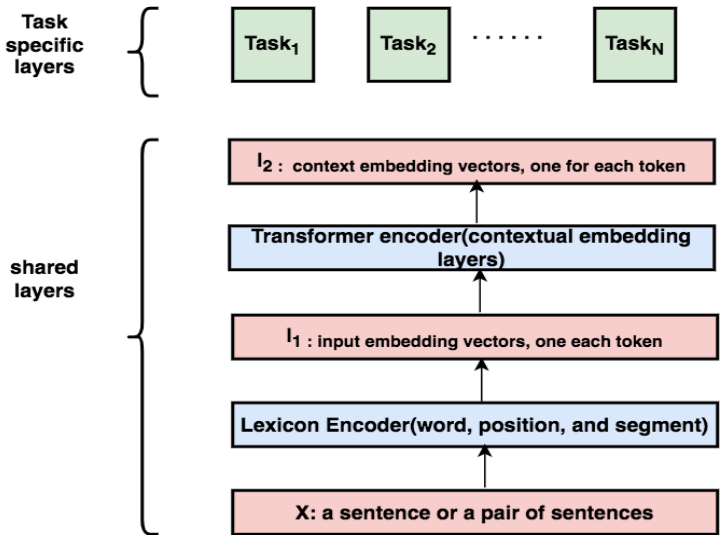

Figure 2. Multi task learning architecture with BERT/ALBERT as shared encoder

Transformer Encoder (l2): It consists of multi-layer bidirectional Transformer encoder (Vaswani et al) 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 labelled as class c is predicted with softmax:

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

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.

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

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

We experimented with BERT, and ALBERT as shared encoder in MTL which we termed as **(iii)Model 3(M3):MTL with BERT** and **(iv)Model 4(M4):MTL with ALBERT** as the shared encoder.

4.4. Sentiment and Emotion knowledge

High-quality annotation data is scarce in hate speech detection, which makes the task stereotype words and hence suffer from inherently biased training. Sentiment analysis research has been carried out for many years, and there are abundant high-quality labelled datasets. There is a high degree of correlation between two tasks,

Negative sentiment can be an indicator of hate as reported in the previous research. Therefore, we adopt a multi-task learning method for sentiment knowledge sharing, so as to better extract sentiment features and apply them to hate speech detection.

Model 5(M5): The BERT is used as shared encoder with eleven hate and eight sentiment task trained jointly.

Model 6(M6): The ALBERT is used as shared encoder with eleven hate and eight sentiment task trained jointly.

The same architecture as in Figure 2 is used for M5 and M6.

4.5. Ensemble learning

The Ensemble learning strategy have been proposed to effectively generalize machine learning techniques in several domains including text classification. The existing approaches on ensemble learning have outperformed the baseline classifiers by reducing the variance of predictions. In our experiments we utilized stacking based approach that combines multiple machine learning

algorithm via meta learning. In this, the base level algorithms are trained on complete datasets, the meta-model is trained on the final outcomes of all the base model as the feature. This model is termed as **Model 7(M7): Stack based Ensemble** and figure 3 explains the idea.

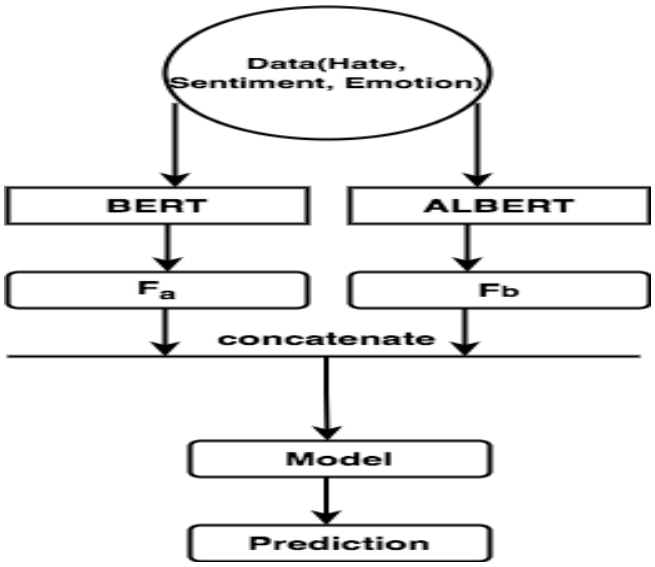

Figure 3. Stack based Ensemble

4.6. Experimental Setup

All the deep learning models were implemented using Keras, a neural network package **[38]** with Tensorflow **[39]** as 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. For some data with the seperate test set, model is trained on train data and performance is evaluated using test data. Categorical cross-entropy is used as a loss function, and Adam **[40]** optimizer is used for optimizing the network.

We use Adam optimizer and 2e-5 for the transformer models. The batch size of 30 is used to train the shared encoder and epoch of 2 is found to be optimal. The value for bias is randomly initialized to all zeros, 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 a various NLP tasks.

5. RESULTS, COMPARISON AND ANALYSIS

We report the accuracy and weighted-F1 score of all the eleven datasets in Table 5 and Table 6. Table 7 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 Ensemble of BERT and ALBERT trained in MTL with sentiment and emotion features outperformed the other methods. The ensemble based approach obtained the best results for all the eleven hate domain data. We are also presenting the qualitative and quantitative analysis on the obtained results to highlight some of the errors that need to be rectified to improve the system performance.

Table 5. Weighted-F1 of eleven datasets.

Datasets	M1	M2	M3	M4	M5	M6	M7
D1	91.10	89.50	92.73	90.81	93.13	90.93	**93.63**
D2	85.50	84.40	89.66	87.03	89.98	87.36	**90.10**
D3	78.80	77.80	83.32	82.52	83.51	82.78	**83.78**
D4	79.70	80.10	83.18	84.57	83.48	84.71	**83.92**
D5	75.90	77.40	82.54	83.37	82.92	83.63	**83.89**
D6	80.50	79.10	82.48	82.58	82.89	82.96	**83.54**
D7	56.40	55.80	59.80	56.62	59.91	56.74	**60.14**
D8	83.80	83.80	86.58	86.56	87.13	86.82	**87.52**
D9	79.20	79.98	80.82	81.22	81.46	81.69	**81.96**
D10	72.30	71.50	76.32	74.23	77.18	74.80	**77.78**
D11	80.70	80.90	81.93	82.32	82.28	82.46	**82.89**

Table 6. Accuracy of eleven datasets

Datasets	M1	M2	M3	M4	M5	M6	M7
D1	91.90	90.80	93.60	92.70	94.17	93.10	**94.72**
D2	86.40	85.41	90.69	87.34	91.93	87.72	**91.99**
D3	78.90	78.10	85.14	84.43	92.18	90.98	**92.63**
D4	79.80	80.10	83.93	84.84	84.63	85.34	**85.78**
D5	76.90	78.50	84.51	84.38	84.98	84.65	**85.67**
D6	80.93	78.12	87.23	86.43	87.76	83.32	**87.99**
D7	57.82	59.23	62.64	59.98	63.12	60.78	**63.93**
D8	85.10	85.10	87.38	90.51	89.10	91.78	**92.13**
D9	79.84	80.62	81.32	82.54	81.98	84.34	**84.96**
D10	76.70	74.10	79.83	77.23	81.34	79.32	**81.54**
D11	80.90	81.50	81.91	82.12	82.67	83.32	**83.72**

Table 7. Comparison to the state-of-the-art systems and the proposed approach

Best Model (Weighted-F1)	Comparison (Weighted-F1)
D1 (**93.63**)	**[4]**: (90), **[41]:** (91.10)
D2 (**90.10**)	**[42]**: (83), **[43]**: (86), **[8]:** (87)
D3 (**83.78**)	**[44]:** (58.72)
D4 (**83.92**)	**[45]**: (72.85), **[46]**: (78.3)
D5 (**83.89**)	**[41]:** (72.75), **[47]:** (73.6)
D6 (**83.54**)	**[48]**:(74.65), **[49]:**(74.31),
D7 (**60.14**)	**[50]:**(54.60), **[51]:**(51.90),
D8 (**87.52**)	**[52]:** (82.01), **[53]:** (78.40)
D9 (**81.96**)	**[54]:**(80.20), **[55]:**(75.90)
D10 (**77.78**)	**[56]**: (80.89), **[57]:**(81.99)
D11 (**82.89**)	**[58]:**(78.40), **[58]:**(80.10)

5.1. Quantitative Analysis

The sentences in Neutral class play a very crucial role in determining the annotators' global knowledge about any specific topic and how much they can distinguish between free speech or any subtypes of harmful speech. We analyzed the misclassification rate of 5 datasets from one class into other over the baseline BERT model and the best performing stacked MTL using BERT and ALBERT in Table 8. In the M1 for D1 in Table 8 ,9.2% of hate tweets were misclassified to neutral showing the model's ability to distinguish the hateful text. The addition of

emotion and sentiment features in the MTL setting with ALBERT for D1 improved with only 27.42% and 6.6% misclassification to offensive and neutral. The error rate for all the other 4 datasets were improved with MTL based approach. The most notable improvement is 61% improvement in case of harassment class.

5.2. Qualitative Analysis

Table 9 and Table 10 consists of True positive of hate sub class and false positive of hate sub class. We show seven different types of hate speech that were correctly classfied by all the models. Some of the misclassified non-hate into hate is shown in Table 9. The lack of adequate contextual information is one of the factor involved due to which model is not able to distinguish non-hate from hate.

Table 8. Misclassification comparison between Model 1(M1) and Model 7(M7)

Model	**Class**	**Misclassification**
M1	Hate (D1)	Offensive (63.76%), Neutral (9.2)
M7	Hate (D1)	Offensive (27.42%), Neutral (6.6%)
M1	Offensive (D1)	Hate (3.9%) , Neutral (3.5%)
M7	Offensive (D1)	Hate (2.1%), Neutral (1.2%)
M1	Racism (D2)	Sexism (0.8%), Neutral (23%)
M7	Racism (D2)	Sexism (0.4%), Neutral (8.78%)
M1	Sexism (D2)	Racism (1.34%), Neutral (35.66%)
M7	Sexism (D2)	Racism (0.76%), Neutral (14.03%)
M1	OAG (D3)	CAG (58.07%), NAG (16.61%)
M7	OAG (D3)	CAG (14.73%), NAG (6.8%)
M1	Offensive (D4)	NOT (40.56%)
M7	Offensive (D4)	NOT (15.79%)
M1	Harassment (D5)	Non-Harassment (79.18%)
M7	Harassment (D5)	Non-Harassment (18.65%)

6. CONCLUSION AND FUTURE WORK

Our study is based on the assumption that discourse of hate speech detection involves other affective components such as sentiment and emotion. We have leveraged the labeled corpora for each tasks and experimented on single task learning and multi-task learning paradigm. Our results demonstrates that stack based multi-task architectures are the best performing model and emotion and sentiment knowledge sharing improves system performance and advances hate speech detection. The plausible extensions include the inclusion of more affective phenomenon correlated to hate speech such as sarcasm/irony **[59]**, "big five" personality traits **[60]**, and emotion roe labeling **[61].**

Table 9. True Postives

Sentence No.	Type	Tweets
1.	Toxic	bitch please whatever
2.	Non-Toxic	@user your sadness is exactly what the terrorist want
3.	Direct Attack	@user the jew are te mastermind idiot
4.	Indirect Attack	@user he is a dumb and dumber
5.	Doubtful	shame on icc now please stop it
6.	References	What an idiot. \#buildthatwall
7.	Annotators Bias	ball is in our court not yours

Table 10. False Positives

1.	Toxic	@user a woman you should not complain about cleaning up your house a man should always take the trash out
2.	External Knowledge	user eebo who want to get there nose in these bad bois then scally chav sock fetish game of basketball hoe
3.	Hate is subjective	user @user love frat boy with soft long

REFERENCES

[1] Vazsonyi, A. T., Machackova, H., Sevcikova, A., Smahel, D., & Cerna, A. (2012). Cyberbullying in context: Direct and indirect effects by low self-control across 25 European countries. *European Journal of Developmental Psychology*, *9*(2), 210-227.

[2] Zhou, X., Yong, Y., Fan, X., Ren, G., Song, Y., Diao, Y., ... & Lin, H. (2021, August). Hate speech detection based on sentiment knowledge sharing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 7158-7166).

[3]. del Arco, F. M. P., Halat, S., Padó, S., & Klinger, R. (2021). Multi-task learning with sentiment, emotion, and target detection to recognize hate speech and offensive language.

[4] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, May). 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).

[5] Waseem, Z., & Hovy, D. (2016, June). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop* (pp. 88-93).

[6] Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016, April). Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web* (pp. 145-153).

[7] Zhang, Z., Robinson, D., & Tepper, J. (2018, June). Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference* (pp. 745-760). Springer, Cham.

[8] Founta, A. M., Chatzakou, D., Kourtellis, N., Blackburn, J., Vakali, A., & Leontiadis, I. (2019, June). A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM conference on web science* (pp. 105-114).

[9] Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017, April). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion* (pp. 759-760).

[10] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, *5*, 135-146.

[11] Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

[12] Mehdad, Y., & Tetreault, J. (2016, September). Do characters abuse more than words?. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 299-303).

[13] Gambäck, B., & Sikdar, U. K. (2017, August). Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online* (pp. 85-90).

[14] Brassard-Gourdeau, E., & Khoury, R. (2019, August). Subversive toxicity detection using sentiment information. In *Proceedings of the Third Workshop on Abusive Language Online* (pp. 1-10).

[15] Mozafari, M., Farahbakhsh, R., & Crespi, N. (2019, December). A BERT-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications* (pp. 928-940). Springer, Cham.

[16] Koufakou, A., Pamungkas, E. W., Basile, V., & Patti, V. (2020). HurtBERT: incorporating lexical features with BERT for the detection of abusive language. In *Fourth Workshop on Online Abuse and Harms* (pp. 34-43). Association for Computational Linguistics.

[17] Safaya, A., Abdullatif, M., & Yuret, D. (2020, December). Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (pp. 2054-2059).

[18] Kumar, R., Reganti, A. N., Bhatia, A., & Maheshwari, T. (2018). Aggression-annotated corpus of hindi-english code-mixed data. *arXiv preprint arXiv:1803.09402*.

[19] Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*.

[20] Golbeck, J., Ashktorab, Z., Banjo, R. O., Berlinger, A., Bhagwan, S., Buntain, C., ... & Wu, D. M. (2017, June). A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on web science conference* (pp. 229-233).

[21] De Gibert, O., Perez, N., García-Pablos, A., & Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*.

[22] Basile, V., Bosco, C., Fersini, E., Debora, N., Patti, V., Pardo, F. M. R., ... & Sanguinetti, M. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation* (pp. 54-63). Association for Computational Linguistics.

[23] Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., & Patel, A. (2019, December). 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* (pp. 14-17).

[24] Mandl, T., Modha, S., Shahi, G. K., Madhu, H., Satapara, S., Majumder, P., ... & Jaiswal, A. K. (2021). 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*.

[25] Bhattacharya, S., Singh, S., Kumar, R., Bansal, A., Bhagat, A., Dawer, Y., ... & Ojha, A. K. (2020). Developing a multilingual annotated corpus of misogyny and aggression. *arXiv preprint arXiv:2003.07428*.

[26] Founta, A. M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., ... & Kourtellis, N. (2018, June). Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

[27] Rosenthal, S., Farra, N., & Nakov, P. (2017, August). SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)* (pp. 502-518).

[28] Mohammad, S. (2012). # Emotional tweets. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)* (pp. 246-255).

[29] Scherer, K., & Wallbott, H. (1997). The ISEAR questionnaire and codebook. *Geneva Emotion Research Group*.

[30] Li, Y., Su, H., Shen, X., Li, W., Cao, Z., & Niu, S. (2017). Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

[31] Plaza-del-Arco, F. M., Strapparava, C., Lopez, L. A. U., & Martín-Valdivia, M. T. (2020, May). EmoEvent: A multilingual emotion corpus based on different events. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 1492-1498).

[32] Zhao, J., Liu, K., & Xu, L. (2016). Sentiment analysis: mining opinions, sentiments, and emotions.

[33] Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016, June). Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)* (pp. 31-41).

[34] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[35] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

[36] Xue, Y., Liao, X., Carin, L., & Krishnapuram, B. (2007). Multi-Task Learning for Classification with Dirichlet Process Priors. *Journal of Machine Learning Research*, *8*(1).

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

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

[39] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... & Zheng, X. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

[40] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[41] Chakrabarty, T., Gupta, K., & Muresan, S. (2019, August). Pay "attention" to your context when classifying abusive language. In *Proceedings of the Third Workshop on Abusive Language Online* (pp. 70-79).

[42] Park, J. H., & Fung, P. (2017). One-step and two-step classification for abusive language detection on twitter. *arXiv preprint arXiv:1706.01206*.

[43] Kshirsagar, R., Cukuvac, T., McKeown, K., & McGregor, S. (2018). Predictive embeddings for hate speech detection on twitter. *arXiv preprint arXiv:1809.10644*.

[44] Kapil, P., Ekbal, A., & Das, D. (2020). Investigating deep learning approaches for hate speech detection in social media. *arXiv preprint arXiv:2005.14690*.

[45] Cambray, A., & Podsadowski, N. (2019). Bidirectional recurrent models for offensive tweet classification. *arXiv preprint arXiv:1903.08808*.

[46] Liu, P., Li, W., & Zou, L. (2019, June). NULI at SemEval-2019 Task 6: Transfer Learning for Offensive Language Detection using Bidirectional Transformers. In *SemEval@ NAACL-HLT* (pp. 87-91).

[47] Naseem, U., Razzak, I., & Hameed, I. A. (2019). Deep Context-Aware Embedding for Abusive and Hate Speech detection on Twitter. *Aust. J. Intell. Inf. Process. Syst.*, *15*(3), 69-76.

[48] Mishra, S., & Mishra, S. (2019). 3Idiots at HASOC 2019: Fine-tuning Transformer Neural Networks for Hate Speech Identification in Indo-European Languages. In *FIRE (Working Notes)* (pp. 208-213).

[49] Baruah, A., Barbhuiya, F., & Dey, K. (2019, June). Abaruah at semeval-2019 task 5: Bi-directional lstm for hate speech detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 371-376).

[50] Ding, Y., Zhou, X., & Zhang, X. (2019, June). Ynu_dyx at semeval-2019 task 5: A stacked bigru model based on capsule network in detection of hate. In *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 535-539).

[51] Montejo-Ráez, A., Jiménez-Zafra, S. M., Garcia-Cumbreras, M. A., & Díaz-Galiano, M. C. (2019, June). SINAI-DL at SemEval-2019 Task 5: Recurrent networks and data augmentation by paraphrasing. In *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 480-483).

[52] MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PloS one*, *14*(8), e0221152.

[53] Berglind, T., Pelzer, B., & Kaati, L. (2019, August). Levels of hate in online environments. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 842-847).

[54] Risch, J., & Krestel, R. (2020, May). Bagging BERT models for robust aggression identification. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying* (pp. 55-61).

[55] Mishra, S., Prasad, S., & Mishra, S. (2020, May). Multilingual Joint Fine-tuning of Transformer models for identifying Trolling, Aggression and Cyberbullying at TRAC 2020. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying* (pp. 120-125).

[56] Mitra, A., & Sankhala, P. (2022). Multilingual Hate Speech and Offensive Content Detection using Modified Cross-entropy Loss. *arXiv preprint arXiv:2202.02635*.

[57] Glazkova, A., Kadantsev, M., & Glazkov, M. (2021). Fine-tuning of Pre-trained Transformers for Hate, Offensive, and Profane Content Detection in English and Marathi. *arXiv preprint arXiv:2110.12687*.

[58] Lee, Y., Yoon, S., & Jung, K. (2018). Comparative studies of detecting abusive language on twitter. *arXiv preprint arXiv:1808.10245*.

[59] Reyes, A., Rosso, P., & Buscaldi, D. (2012). From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, *74*, 1-12.

[60] Flek, L. (2020, July). 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).

[61] Mohammad, S., Zhu, X., & Martin, J. (2014, June). 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).

AUTHORS

1. **Prashant Kapil** is a PhD scholar at 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 India, for providing financial support in the form of UGC NET-JRF/SRF. Research interests: AI, NLP, and ML

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