Annotated Corpus of Tweets in English from Various Domains for Emotion Detection

Soumitra Ghosh[1], **Asif Ekbal**[1], **Pushpak Bhattacharyya**[1], **Sriparna Saha**[1],
Vipin Tyagi[2], **Alka Kumar**[2], **Shikha Srivastava**[2], **Nitish Kumar**[2]
[1]Indian Institute of Technology Patna, India
{1821cs05, asif, sriparna, pb}@iitp.ac.in
[2]Centre for Development of Telematics (C-DOT, India)
{vipin, alkakm, shikha, nitish}@cdot.in

Abstract

Emotion recognition is a very well-attended problem in Natural Language Processing (NLP). Most of the existing works on emotion recognition focus on the general domain and in some cases to specific domains like fairy tales, blogs, weather, Twitter etc. But emotion analysis systems in the domains of security, social issues, technology, politics, sports, etc. are very rare. In this paper, we create a benchmark setup for emotion recognition in these specialised domains. First, we construct a corpus of 18,921 tweets in English annotated with Paul Ekman's six basic emotions (*Anger, Disgust, Fear, Happiness, Sadness, Surprise*) and a non-emotive class *Others*. Thereafter, we propose a deep neural framework to perform emotion recognition in an end-to-end setting. We build various models based on Convolutional Neural Network (CNN), Bi-directional Long Short Term Memory (Bi-LSTM), Bi-directional Gated Recurrent Unit (Bi-GRU). We propose a Hierarchical Attention-based deep neural network for Emotion Detection (HAtED). We also develop multiple systems by considering different sets of emotion classes for each system and report the detailed comparative analysis of the results. Experiments show the hierarchical attention-based model achieves best results among the considered baselines with accuracy of 69%.

1 Introduction

Online social media provides a platform for people to share their perspectives on various issues with their close ones or in the public forum. Twitter is a very popular and heavily used platform among the most social media users. The USA and India are the 1st and 3rd leading countries based on number of twitter users as of July 2020[1]. Twitter data serves as a rich source for text analysis tasks as this type of data is not very long, yet very rich in emotion content. This type of communication is free from barriers of age, race, culture, gender, etc. In recent times, understanding people's opinion and sentiment have been the need of the hour to meet various purposes, such as real-time trending, better customer service, winning elections, etc. Particularly in Indian context, we often hear stories how a single social-media post has changed the life of an individual or an organization in a positive[2] or negative way.

Sentiment analysis or opinion mining deals with the automatic identification and extraction of the underlying subjective information from text. It is often synonymously described as 'polarity detection' which is concerned with classifying an instance of data as 'positive', 'negative' or 'neutral'. Contrary to sentiment analysis, the emotional analysis relies on a more fine-grained analysis of the subjectivity information. It deals with the deeper analysis of human emotions and sensitivities. Emotion analysis goes a step further into a person's motives and impulses. It gives valuable and exact insights that are easily transformed into actions. It is usually based on a wide spectrum of moods rather than a couple of static categories. Inside positive sentiment, it detects specific emotions like happiness, pride, satisfaction, thankfulness or excitement, depending on how it is configured. Similarly, negative sentiment may span a variety of emotions like anger, sadness, fear, hopelessness, blame, etc. Usually, sentiment or emotion analysis works better on subjective texts (texts having emotions or feelings) than objective ones (statements or facts).

In this work, we focus on building a corpus of emotion-annotated tweets from the relevant topics

[1] https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

[2] How social-media revived a helpless old couple running an eatery 'baba ka dhaba': https://www.instagram.com/p/CGDAHGxlGTv/

Proceedings of the 17th International Conference on Natural Language Processing, pages 460–469
Patna, India, December 18 - 21, 2020. ©2020 NLP Association of India (NLPAI)

of interest of recent times (i.e. social issues, technology, cyber-security etc.) that are quite a buzz among today's online social platform users. Most of the available corpora for emotion are from the domains limited to blogs, weather, elections, fairy tales and some more. There are some available corpora in the general domain too, but they mostly cover tweets from politics, world news, sports, etc. At the present, hardly any corpora cover diverse domains. We consider Paul Ekman's (Ekman, 1992) basic emotions *(Anger, Disgust, Fear, Happiness, Sadness, Surprise)* for the emotion labelling task. A non-emotive label *Others* is also introduced for tweets which do not fall within the scope of Ekman's basic emotions (Ekman, 1992).

For effective usage of the dataset, we develop multiple single-task models for emotion classification. We develop CNN, Bi-GRU and Bi-LSTM based deep learning models for the emotion classification task. We propose a Bi-GRU based framework with hierarchical attention (Yang et al., 2016) mechanism to extract important information from each sentence in a tweet effectively. We also develop a couple of models on a sub-set of emotion classes (4 classes and 6 classes) and perform a comparative analysis with the developed systems with 7 classes. The proposed hierarchical attention system for 7 classes attains superior results than the considered baselines with an overall test accuracy of 69% for the emotion classification task.

The rest of the paper is organized as follows. In section 2, we discuss some of the existing work and corpora concerning emotion analysis. Various aspects of resource creation, challenges and analysis are discussed in Section 3. Next, we discuss the methodologies we implement in Section 4. In Section 5, we discuss the implementation details, results and qualitative error analysis. Finally, we conclude in Section 6 and acknowledge the funding agency for this work in Section 7.

2 Background

In the past decades, several annotated corpora have been created for emotion recognition from texts. Various annotation schemes were introduced to serve the specific purpose for which the corpus is created. (Scherer and Wallbott, 1994) collected questionnaires answered by people with different cultural backgrounds to form *The International Survey on Emotion Antecedents and Reactions (ISEAR)* dataset. People reported on their emotional events. The dataset contains a total of 7,665 sentences from reports by approximately 3,000 respondents. Sentences are annotated with single labels, chosen from the set of following labels: *joy, fear, anger, sadness, disgust, shame, and guilt*. The *Affective Text* task (Strapparava and Mihalcea, 2007) in SemEval 2007 was proposed to focus on the emotion classification of news headlines extracted from news web sites. Given a set of predefined six emotion labels (Paul Ekman's basic emotions (Ekman, 1992)), classify the titles with the appropriate emotion label and/or with a valence indication *(positive/negative)*. (Aman and Szpakowicz, 2007) published a dataset of blog content consisting of 5,205 sentences from 173 blogs. Each instance is annotated with an emotion label from Ekman's basic emotions (Ekman, 1992) and also with an intensity score for that emotion. (Alm, 2008) researched the text-based emotion prediction problem in the literature domain. The author provided an annotated corpus of 15,302 sentences from 176 stories annotated from among the following seven emotion classes *(angry, disgusted, fearful, happy, sad, positive surprise, and negative surprise)*.

Crowdflower's dataset, *The Emotion in Text*[3], is a noisy single-labelled crowd-sourced annotated corpus of tweets. It primarily follows Plutchik's 8 basic emotions (Plutchik, 2001) in addition to another 3 emotions *(love, confusion and no emotion)*. The *Electoral-Tweets* dataset, published by (Mohammad et al., 2015), targets the domain of elections (2012 US Presidential election). It consists of over 100,000 crowdsourced responses to two detailed online questionnaires (the questions targeted emotions, purpose, and style in electoral tweets). (Ghazi et al., 2015) published the Emotion-Stimulus dataset to predict the cause of emotion in the text. The dataset consists of 820 sentences which are annotated both with emotions (one label per sentence) and their causes, and 1,549 sentences which are marked only with their emotion. Ekman's basic emotions (Ekman, 1992) with an added class *Shame* have been used for the annotation. *The Hashtag Emotion Corpus*, also known as *Twitter Emotion Corpus (TEC)*, was published by (Mohammad and Kiritchenko, 2015), and consists of 21,051 tweets. This resource was created to understand if emotion-word hashtags can successfully be used as emotion labels. Ekman's basic emotions

[3] https://data.world/crowdflower/sentiment-analysis-in-text

(Ekman, 1992) have been considered for the annotation process. Tweets were scraped that contained hashtags in the form *#emotion* corresponding to Ekman's (Ekman, 1992) 6 basic emotions (like #anger, #disgust).

DailyDialogs is a dataset of dialogs published by (Li et al., 2017) spanning over a variety of topics and better structured than any social media data. The SSEC corpus (Schuff et al., 2017) is an annotation of the SemEval 2016 Twitter stance and sentiment corpus (Mohammad et al., 2017) with Plutchik's emotion labels (Plutchik, 2001). The authors studied the relation between emotion annotation and the other annotation layers like stance and sentiment. The *EmoInt* dataset published by (Mohammad and Bravo-Marquez, 2017) for evaluation of the *WASAA-2017 Shared Task of Emotion Intensity (EmoInt)* contains 7,097 tweets annotated with a pair of emotion tag and intensity score of the corresponding emotion. The annotation was done via crowdsourcing with primarily (but not limited to) one among the following 4 emotions *anger, joy, sadness, and fear* and their respective intensity score ranging between 0 to 1. The *Affect in Tweets Dataset* of the SemEval 2018 Task 1 (Mohammad et al., 2018) was introduced to determine the intensity of both emotion and sentiment as well as multi-label emotion classification of tweets.

Recent works have shown the effectiveness of multi-task systems by learning several correlated tasks simultaneously (Akhtar et al., 2019; Majumder et al., 2019; Akhtar et al., 2020).

3 Corpus Creation

In this section we briefly describe the data creation process starting from the collection, pre-processing, annotation and inter-annotator reliability.

3.1 Data Collection

We use the Twython [4] python library (wrapper) to extract tweets from Twitter's Standard search API [5]. Tweets were extracted using certain domain-specific keywords *(terror, cyber-security, technology)* and their combinations between the time interval January 2018 and August 2019. Several hundred thousand tweets were collected initially. This was filtered using various lexicons to increase the coverage of the affect oriented tweets. Irrelevant (questions, requests, poems) and code-mixed instances were removed. Before using the corpus for our experiments, we perform some basic pre-processing to remove noise from the data (removing URLs [6], mentions, non-ASCII characters, punctuations except '.', '!' and '?', conversion to lowercase, etc). Smileys are replaced by their meanings (for example: *:-(* as *sad*, *<3* as *Love*). Frequently used contractions are replaced by their full versions (for example: *can't* as *cannot*, *he's* as *he is*). Each tweet is sentence tokenized considering '.', '!' and '?' as the sentence delimiters.

3.2 Data Annotation

The entire dataset is manually annotated by three annotators with the single-labelling scheme at the document level, that is, a tweet can have at most one emotion label. Some pre-annotated instances from each category of emotion were prepared to be shared with the annotators to facilitate the annotation process. Here, we use Ekman's (Ekman, 1992) basic emotions that have been widely used among several emotion classification tasks so far. In addition to these 6 basic emotions *(Anger, Disgust, Fear, Joy, Sadness, Surprise)*, we introduce a non-emotive class *Others* to mark those instances that do not fall in the above emotion categories. A few annotation samples are shown in Table 1.

3.3 Inter-Annotator Agreement

We measure the agreement among different annotators using Cohen's Kappa Statistic (Cohen, 1960) for the emotion task. The average agreement attained over the entire dataset was 0.74 which indicates that the annotations are of fair quality. Highest individual agreement attained was for the class *Joy* (0.87) followed by *Others* (0.83). Lowest attained agreement score (0.62) was for the *Surprise* class. This may be attributed to two reasons: the very low number of instances in the surprise class and the presence of both positive and negative types of surprise instances.

3.4 Corpus Analysis

Looking into the content of the corpus, we observe that certain words frequently occur in instances across all the emotion classes. For understanding the role of a certain entity (or event) as a contributing factor towards generating a particular kind of emotion, we observe some frequently

[4] https://pypi.org/project/twython/

[5] https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets

[6] We use Beautifulsoup to remove URLs: https://pypi.org/project/beautifulsoup4/

Actual tweet	Emotion
Tropical Cyclone Mona to Hit Fiji Islands on Saturday January 5, 2019 https://t.co/nwknedBFba	Others
@congressdotgov Louis Farrakhan promotes terrorism and racism. Why doesn't anyone talk about him calling white people Satan?	Anger
@WillieHarveyJr @CycloneFB Thanks for being a CYCLONE!!! Honored to have you!	Joy
Crime will not go WAY DOWN because of a border wall, sir. There is crime from US CITIZENS EVERYDAY. https://t.co/HOg4EfEfnP	Sadness

Table 1: Samples of annotated tweets from our dataset

occurring words (like terrorism, technology, crime, etc.) and their frequencies of occurrences among all the emotion classes. It is found that although some words tend to occur across all the emotion classes, their frequencies of distribution differ considerably based on the type of word and the nature of the emotion. Words like *terrorism, weapons* occur more frequently in classes like *anger, disgust, fear* than in *joy, sadness, others* classes. The annotated dataset has a very highly skewed distribution of instances over the 7 classes that we consider. There are four severely under-represented classes, namely *(disgust, fear, sadness and surprise)* and one over-represented class *(others)*.

4 Methodologies

We develop various deep learning-based multi-task models for automatic detection of emotion and its intensity. As base learning techniques, we use Convolution Neural Network (CNN) (Kim, 2014), Long Short Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Unit (GRU) network (Cho et al., 2014). We build three separate multi-task models (CNN based, Bi-GRU based and Bi-LSTM based) on top of pre-trained word embedding (GloVe [7] (Pennington et al., 2014)). The embedding layer is initialized with the pre-trained weights and is learned during the training in accordance with our dataset. We employ word attention (Bahdanau et al., 2014) mechanism to focus on the informative words in a document (tweet) and obtain an aggregated representation(document vector) which is passed through two fully-connected layers (100 neurons in each layer) and an output layer (with 7 neurons, one for each class) with Softmax activation. We use the categorical cross-entropy as the loss function.

4.1 Convolution Neural Network (CNN)

In the past few years, CNN has produced some break-through results in various NLP tasks. CNN relies heavily upon two operations for extracting features: convolution (produce feature maps) and pooling (dimension reduction). (Kim, 2014); (Akhtar et al., 2016); (Singhal and Bhattacharyya, 2016) have used CNN in different sentiment analysis tasks. Our CNN based classification system employs 3 convolution layers in parallel with 100 filters of sizes 2, 3 and 4 respectively. The output of the layers is added (merged) to produce a single output of the same shape as the individual layer's output. Max pooling operation (poolsize = 2) is performed on the convoluted output which is further passed through an attention layer (Bahdanau et al., 2014) to get an aggregated representation (document vector) of the informative words in a document (tweet). Lastly, the output from the attention layer is passed through two fully connected layers (with 100 neurons in each layer) with ReLu activation (Glorot et al., 2011) and an output layer (with 15 neurons, one for each class) with Softmax activation.

4.2 Bidirectional Long Short Term Memory Network (Bi-LSTM)

LSTMs (Hochreiter and Schmidhuber, 1997) are well known for their ability to preserve long term dependencies in the text, thus eliminating the vanishing gradient problem. LSTM employs 3 gates (forget, input and output) for regulating the amount

[7] http://nlp.stanford.edu/data/wordvecs/glove.840B.300d.zip

of information it wants to retain in its cell state (memory). Bi-LSTMs have shown promising results for several applications. Bi-LSTMs run the inputs in both forward and backward passes (positive time direction and negative time direction) generating two hidden states which, when combined, preserve information from both past and future. We use a Bi-LSTM layer having 256 neurons with Tanh recurrent-activation and 25% dropout and recurrent-dropout. The encoded representation from the Bi-LSTM layer is passed through the word attention (Bahdanau et al., 2014) layer which is further passed through 2 fully connected layers (with 100 neurons in each layer) with ReLu activation (Glorot et al., 2011) and an output layer (with 7 neurons, one for each class) with Softmax activation.

4.3 Bidirectional Gated Recurrent Unit (Bi-GRU)

Unlike LSTMs where 3 gates are involved, GRUs has 2 gates (update and reset gate) to control the amount of information it wants to retain, making it simpler and faster internally than LSTMs. Bidirectional GRUs takes into account the use of information from both the past time steps and future time steps to make decisions about the present state. Unlike LSTMs, GRUs (Cho et al., 2014) do not have any explicit cell state (memory) but still handle the vanishing gradient problem and learn long-term dependencies by the help of its gates mechanism. We use a Bidirectional GRU layer having 256 neurons. Word attention (Bahdanau et al., 2014) is applied on the encoded output (from the GRU layer) which is further passed through 2 fully connected layers (with 100 neurons in each layer) with ReLu activation (Glorot et al., 2011) and an output layer (with 7 neurons, one for each class) with Softmax activation.

4.4 Hierarchical Attention Based Deep Neural Framework for Emotion Detection (HAtED)

In recent works, Hierarchical attention (Bahdanau et al., 2014) based deep learning systems have gained popularity because of their good and consistent performance in various classification tasks when compared to the existing state-of-the-art techniques. In this work, we try to exploit the advantages of such an approach to improve upon our attention mechanism by focusing on words (at the sentence level) as well as sentences (at document level).

HAtED focuses on each sentence in a tweet individually resulting in sentence vectors which are further attended upon to produce a document vector. The intuition is to focus upon important words in a sentence as well as important sentences in a document (tweet) for a particular emotion. For encoding of the sentences, we leverage Bi-GRU (256 neurons) based word encoder. Without making major changes to the basic architecture of the hierarchical attention framework as in the original work (Yang et al., 2016), we tweaked the last few layers to solve our objective. We pass the document vector through a dense layer (100 neurons with ReLU activation) followed by an output layer (7 neurons with Softmax activation). We use categorical cross-entropy loss function for the classification task.

Besides HAtED, we also develop two separate Hierarchical Attention-based models considering various sets of emotion classes. They are as follows:

- **HAtED$^{\text{4-C}}$**: We develop HAtED$^{\text{4-C}}$ following the same architecture as HAtED but considering a sub-set of the 7 classes as considered in HAtED. We take motivation from the WASSA-2017 shared task (Mohammad and Bravo-Marquez, 2017) on Emotion Intensity and consider instances from the following 4 emotion classes in our dataset: *Anger, Fear, Joy, Sadness*. Leaving out two of the severely under-represented classes to build HAtED$^{\text{4-C}}$ helps us to get a better approximation of the effectiveness of our proposed approach as the negative impact of having unbalanced dataset is considerably reduced.

- **HAtED$^{\text{6-C}}$**: In this setting, we do not consider the Others class and train our model on Ekman's 6 Basic Emotion Classes (Ekman, 1992) which are as follows *Anger, Disgust, Fear, Joy, Sadness, Surprise*. Overall architecture is similar to that of HEtED with only change being the number of neurons in the output layer (6 neurons).

All the models described in section 4 are trained and tuned independently. Training of models is done through backpropagation using the Adam optimizer (Kingma and Ba, 2014). We employ 25% Dropout (Srivastava et al., 2014) in all the fully connected layers to prevent over-fitting.

Emotion	Train	Test	Val	Total
Anger	2475	688	275	3438
Disgust	865	241	97	1207
Fear	445	123	49	617
Joy	2911	809	323	4043
Sadness	647	180	72	899
Surprise	242	67	27	336
Others	6033	1677	671	8381

Table 3: Distribution of instances over respective emotion classes.

5 Experiments

5.1 Experimental Settings

Dataset: For our experiments, we split our curated dataset into three parts: train, validation and test sets in a 70:20:10 ratio, respectively. As mentioned earlier, the dataset is highly skewed with several under-represented classes (disgust, fear, sadness, surprise). The data distribution over the various emotion classes is shown in Table 3.

Implementation: We use the Python-based libraries Keras and Scikit-learn for the implementation. We use 300-dimension GloVe (Pennington et al., 2014) pre-trained embeddings to initialize the embedding layer in our models which are further learned during training on our data to obtain emotion-enriched word representations. First, we develop three basic deep-learning models (CNN, Bi-GRU and Bi-LSTM) which have been extensively used in various classification tasks on textual data. Considering these models as baselines, we build three hierarchical attention based Bi-GRU systems for the emotion classification task. Two of these three systems are built for 4 (HAtED$^{\text{4-C}}$) and 6 (HAtED$^{\text{6-C}}$) emotion classes, respectively. The third one (HAtED) is a hierarchical attention based emotion detection system for 7 classes.

Evaluation metrics: As our dataset is unbalanced, we consider the macro-average measure of precision (P), recall (R) and F1-scores as our evaluating metrics for the emotion detection task. In this case, we also compute the overall test accuracy.

5.2 Results and discussion

Table 2 and table 4 show the per-class precision, recall, F1-score and accuracy values for all the implemented models. Scores for classes with fewer instances (*Disgust, Fear, Sadness, Surprise*) are not at par with that of the better-represented classes with (*Anger, Joy, Others*). The 4-class model (HAtED$^{\text{4-C}}$) achieves better scores compared to its 6-class (HAtED$^{\text{6-C}}$) and 7-class (HAtED) variants for those classes. This may be attributed to the lower degree variance in data because of the smaller number of classes. The hierarchical attention-based systems outperformed the base learning systems (i.e. CNN, Bi-GRU and Bi-LSTM) in terms of the various metrics considered. Overall performance of the models on the test set is shown in Table 5. We compare and evaluate the performance of those systems which are built on **7-classes** *(i.e. CNN, Bi-GRU, Bi-LSTM and HAtED)*. HAtED outperforms the other models (for 7 classes) with a test accuracy of 69% and macro-average F1-score of 0.46. Performance of CNN is better than the HAtED model in terms of precision. Results of the HAtED$^{\text{4-C}}$ and

	CNN				Bi-LSTM				Bi-GRU			
Emotion	P	R	F1	A	P	R	F1	A	P	R	F1	A
Anger	0.55	0.69	0.61	0.69	0.45	0.75	0.59	0.75	0.56	0.66	0.61	0.66
Disgust	0.29	0.01	0.02	0.01	0.60	0.01	0.03	0.01	0.26	0.16	0.20	0.16
Fear	0.75	0.03	0.05	0.03	0.64	0.06	0.11	0.06	0.41	0.17	0.24	0.17
Joy	0.80	0.75	0.77	0.75	0.79	0.79	0.79	0.79	0.79	0.78	0.79	0.78
Sadness	0.27	0.02	0.03	0.02	0.00	0.00	0.00	0.00	0.31	0.11	0.16	0.11
Surprise	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Others	0.67	0.84	0.75	0.84	0.71	0.80	0.75	0.80	0.71	0.82	0.76	0.82

Table 2: Per-class Precision (P), Recall (R), F1-score (F1) and Accuracy (A) values for the CNN, Bi-LSTM and Bi-GRU models

	HAtED$^{4\text{-}C}$				HAtED$^{6\text{-}C}$				HAtED			
Emotion	**P**	**R**	**F1**	**A**	**P**	**R**	**F1**	**A**	**P**	**R**	**F1**	**A**
Anger	0.81	0.85	0.83	0.85	0.64	0.87	0.74	0.87	0.66	0.78	0.71	0.78
Disgust	-	-	-	-	0.42	0.15	0.22	0.15	0.35	0.28	0.31	0.28
Fear	0.44	0.38	0.41	0.38	0.57	0.39	0.46	0.39	0.38	0.15	0.22	0.15
Joy	0.88	0.91	0.90	0.91	0.89	0.89	0.89	0.89	0.78	0.81	0.79	0.81
Sadness	0.62	0.45	0.52	0.45	0.44	0.49	0.46	0.49	0.32	0.11	0.11	0.11
Surprise	-	-	-	-	0.25	0.01	0.03	0.01	0.00	0.00	0.00	0.00
Others	-	-	-	-	-	-	-	-	0.73	0.79	0.76	0.79

Table 4: Per-class Precision (P), Recall (R), F1-score (F1) and Accuracy (A) values for the Hierarchical attention based Bi-GRU systems. '-' indicates no score corresponding to a particular metric for a specific emotion class as that particular system do not consider that emotion class during training.

HAtED$^{6\text{-}C}$ models show that less variance in data (fewer number of classes) enhances the decisive power of the models for the classes it is built upon.

We observe from 5 that when we consider a fewer number of classes during training our system (HAtED$^{4\text{-}C}$), reducing the negative impact of severely under-represented classes (disgust, surprise), the scores for all the metrics improves remarkably when compared to HAtED. The scores from the HAtED$^{6\text{-}C}$ model shows that leaving out the *Others* class has also resulted in improving the scores of HAtED system. In other words, the introduction of *Others* class leads to an increase in misclassifications by the HAtED system. This effect is quite similar to the situation of having *Neutral* class in training for a Sentiment classification task which eventually leads to performance degradation of the sentiment classifier.

Models	P	R	F1	Acc.
(Considering 7 classes)				
Baselines				
CNN	**0.47**	0.33	0.32	0.66
Bi-GRU	0.43	0.38	0.39	0.67
Bi-LSTM	0.46	0.34	0.32	0.67
Proposed				
HAtED	0.46	**0.46**	**0.46**	**0.69**
Considering 4 classes				
HAtED$^{4\text{-}C}$	0.68	0.64	0.66	0.81
Considering 6 classes)				
HAtED$^{6\text{-}C}$	0.53	0.46	0.46	0.71

Table 5: macro-average Precision, Recall, F1-score values and Accuracy scores on the test set are shown in the table. Values in bold signify the best attained scores for the respective metrics among the 7-class models.

5.3 Qualitative Analysis

We perform a detailed qualitative analysis of the results from the models that we developed. Table 6 (correct classifications) and Table 7 (incorrect classifications) show some samples of predictions by the HAtED model. Indeed, implicit tweets (i.e. instances not having any explicit mention of affect information) are the major sources of errors. For example:

- **Tweet**: @Ru_NRD Truueee. They said they got involved to fight terrorism and ISIS but we all know why they are really involved. **Actual**: *Disgust*; **Predicted**: *Joy*.

Tweets having multiple emotions (say, anger as well as disgust) seems to hinder the models' overall performance.

- **Tweet**: @nimish4fk @RatanSharda55 @kushal_mehra @vivekagnihotri Why are you surprised? Congress is the power that created naxalism in india and used it to grab and retain power. **Actual**: *Disgust*; **Predicted**: *Anger*.

It is observed that certain instances from *joy* and *others* class are misclassified as belonging to some negative emotion class. This is primarily due to the presence of word(s), in such instances, which have mostly occurred in negative contexts in the overall dataset.

6 Conclusion

In this work, we have proposed a benchmark deep learning setup for emotion detection. The

Tweet	Act_Emo	Pred_Emo
’Criminals are evolving their social engineering tactics in an attempt to trick even the most savvy individuals Stay alert to the latest scam strategies to avoid becoming a victim’	Anger	Anger
’Cyclone Penny re-forms and could still about face toward Queensland coast’	Fear	Fear
’DrumFit is a fun way to blend technology and physical education at Bear Bytes Tech Expo’	Joy	Joy
’#Cyberattacks Skyrocketed in 2018. Are You Ready for 2019? Meet the premier #cyber industry at #ISDEF2019! https://t.co/KUghSb2foD’	Others	Others

Table 6: Samples of correct predictions from the HAtED model. Act_Emo and Pred_Emo means Actual Emotion and Predicted Emotion respectively.

Tweet	Act_Emo	Pred_Emo
’For those who are thinking Casteism doesn’t exist and saying don’t divide Hindus Wake up, u were already divided by the Varna system made by Upper caste’	Anger	Disgust
’When your college finally gets a MS cyber security program!!!’	Joy	Others
’Schools Volcano Explosion Experiment Goes Horrifically Wrong In India’	Fear	Disgust
’I still remember when Arsenal lost 4-0 to Chelsea on my birthday. Terrorism.’	Sadness	Anger

Table 7: Samples of incorrect predictions from the HAtED model.

corpus introduced in this work has been built from diverse domains of tweets carrying various emotions. We have built baseline models with CNN, Bi-GRU and Bi-LSTM. The performance of the Bi-GRU variant has improved significantly when we leveraged the effectiveness of hierarchical attention mechanism in HAtED. Comparison of results has demonstrated that the HAtED model outperforms the baselines by a fair margin (69% test accuracy on the emotion classification task) showing the efficacy of our approach.

Scarcity of instances in some emotion classes have resulted in low per-class performances for those classes showing the scope of improvement for our proposed system. We intend to extend the dataset with the goal to get a balanced distribution of instances over all the classes. We would also like to address the present problem in a multi-label multi-class setting since it was observed that a considerable amount of tweets have shown the presence of more than one emotion. With the intuition that sentiment may play a positive role in assisting the emotion classification task, we are eager to build a parallel multi-task system for automatic emotion (primary task) and sentiment detection (secondary task).

References

Md Shad Akhtar, Asif Ekbal, and Erik Cambria. 2020. How intense are you? predicting intensities of emotions and sentiments using stacked ensemble. *IEEE Computational Intelligence Magazine*, 15(1):64–75.

Md Shad Akhtar, Ayush Kumar, Asif Ekbal, and Pushpak Bhattacharyya. 2016. A hybrid deep learning architecture for sentiment analysis. In *Proceedings* *of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 482–493.

Shad Akhtar, Deepanway Ghosal, Asif Ekbal, Pushpak Bhattacharyya, and Sadao Kurohashi. 2019. All-in-one: Emotion, sentiment and intensity prediction using a multi-task ensemble framework. *IEEE Transactions on Affective Computing*.

Ebba Cecilia Ovesdotter Alm. 2008. *Affect in* text and speech*. Citeseer.

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In *International Conference on Text, Speech and Dialogue*, pages 196–205. Springer.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. 2015. Detecting emotion stimuli in emotion-bearing sentences. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 152–165. Springer.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Navonil Majumder, Soujanya Poria, Haiyun Peng, Niyati Chhaya, Erik Cambria, and Alexander Gelbukh. 2019. Sentiment and sarcasm classification with multitask learning. *IEEE Intelligent Systems*, 34(3):38–43.

Saif M Mohammad and Felipe Bravo-Marquez. 2017. Wassa-2017 shared task on emotion intensity. *arXiv preprint arXiv:1708.03700*.

Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA.

Saif M. Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326.

Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):26.

Saif M Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. 2015. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51(4):480–499.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.

Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23.

Prerana Singhal and Pushpak Bhattacharyya. 2016. Borrow a little from your rich cousin: Using embeddings and polarities of english words for multilingual sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3053–3062.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.