Towards Sentiment and Emotion aided Multi-modal Speech Act Classification in Twitter

Tulika Saha, Apoorva Upadhyaya, Sriparna Saha and Pushpak Bhattacharyya
Department of Computer Science and Engineering, Indian Institute of Technology Patna, India
(sahatulika15,sriparna.saha,pushpakbh)@gmail.com

Abstract

Speech Act Classification determining the communicative intent of an utterance has been investigated widely over the years as a standalone task. This holds true for discussion in any fora including social media platform such as Twitter. But the emotional state of the tweeter which has a considerable effect on the communication has not received the attention it deserves. Closely related to emotion is sentiment, and understanding of one helps understand the other. In this work, we firstly create a new multi-modal, emotion-TA ('TA' means tweet act, i.e., speech act in Twitter) dataset called *EmoTA* collected from open-source Twitter dataset. We propose a Dyadic Attention Mechanism (DAM) based multi-modal, adversarial multi-tasking framework. DAM incorporates intra-modal and inter-modal attention to fuse multiple modalities and learns generalized features across all the tasks. Experimental results indicate that the proposed framework boosts the performance of the primary task, i.e., TA classification (TAC) by benefitting from the two secondary tasks, i.e., Sentiment and Emotion Analysis compared to its uni-modal and single task TAC (tweet act classification) variants.

1 Introduction

Identification of speech acts is one of the preliminary means of determining the communicative intent or pragmatics of a speaker (for example, statement, request, question etc.). This is true for dialogue system, speech transcription, social media such as Twitter, MySpace etc. Twitter is one of the leading micro-blogging services. By 2019, 330 million users were active monthly and 500 million tweets were sent per day[1]. Identification of tweet acts (TAs- speech acts in Twitter) is highly beneficial for Twitter as well as tweeters. For Twitter, it helps decipher a particular subject in terms of speech acts and discrepancy identification. It also helps in social media monitoring by analysing topic alteration or spamming. It assists the followers in monitoring and scanning the subject with the most advantageous speech acts based on their needs. This helps reduce their search space and encourages them to obtain useful information from out of millions of tweets. It gives the tweeter a greater sense of the content, mood and trend.

A person's emotional state and sentiment greatly impacts its intended content (Barrett et al., 1993). Often sentiment and emotion are treated as two different problems (Do et al., 2019), (Soleymani et al., 2017), (Albanie et al., 2018), (Hossain and Muhammad, 2019), (Majumder et al., 2019). However, sentiment and emotion are are closely related. For example, emotions such as *happy* and *joy* are inherently related to a *positive* sentiment. But emotion is much more nuanced and fine-grained compared to sentiment (Kumar et al., 2019). Emotion along with sentiment provides better understanding of the state of mind of the tweeter. For example, a *question* or *statement* is associated with *anticipation*. An *opinion* is many times associated with *anger* or *disgust*. The close association between emotion and sentiment motivates considering tweeter's sentiment along with emotion while deciphering the tweet acts. For expressive TAs such as "expression", "request", "threat" etc., the tweeter's sentiment and emotion can aid in classifying true communicative intent and vice-versa.

Additionally, multi-modal inputs, i.e., the combination of text and other nonverbal cues (emojis in tweets) (Felbo et al., 2017) help create reliable classification models aiding the identification of emotional state and sentiment of the tweeter which in turn help in determining correct TAs.

In this paper, we leverage the relationships as delineated above to predict TAs of tweets in a multi-modal framework. In this multi-task framework, TAC is treated as the primary task and Sentiment

[1] https://www.omnicoreagency.com/twitter-statistics/

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5727–5737
June 6–11, 2021.

Analysis (SA) and Emotion Recognition (ER) as auxiliary (i.e., secondary) tasks.

Contributions of this paper are as follows : **i.** We create a new dataset called *EmoTA* consisting of tweets with high-quality annotations of TAs, including emotionally aided and multi-modal cues; **ii.** We establish the need for considering the sentiment and emotional state of the tweeter while identifying TAs. **iii.** We propose a *Dyadic Attention Mechanism* (DAM) based multi-task adversarial learning framework for multi-modal TAC, SA and ER. In DAM, we incorporate *intra-modal* and *inter-modal* attention to integrate information across multiple modalities and learn generalized features across multiple tasks; **iv.** We illustrate performance gains by jointly optimizing TAC, SA and ER. Multi-modal and multi-task TAC performs significantly better than its uni-modal and single task TAC variants.

2 Related Works

There exist plenty of works which address the task of TAC as a standalone problem. In (Zhang et al., 2011), (Vosoughi and Roy, 2016), authors proposed Machine Learning based approaches for TAC namely Support Vector Machines (SVM), Logistic Regression etc. In (Saha et al., 2020a), authors proposed a first ever public dataset for the identification of speech acts in Twitter followed by a capsule based network built on top of BERT for TAC. In (Vosoughi, 2015), authors highlighted the importance of identification of tweet acts and established it to be one of the elementary steps for detection of rumours in Twitter. In (Saha et al., 2020c), authors proposed an attention based model built on top of the Transformer for predicting TAs. In (Saha et al., 2020a), authors proposed a capsule based network built on top of BERT for TAC. All these works utilized only the textual modality to identify TAs without any sentiment or emotional correlation of the tweeter. In (Cerisara et al., 2018), authors proposed a LSTM based study for jointly optimizing SA and TAC in a decentralized social media platform called Mastodon. However, they modelled their task as a multi-party conversation pretty different in essence to that of Twitter analysis. In (Jeong et al., 2009), authors presented a semi-supervised approach to identify speech acts in emails and different forums. These works, however, use datasets that comprise of face-to-face or telephone data that can not directly aid in advancing work on endless data in electronic mode such as micro-blogging networks, instant-messaging, etc.

Apart from these, identification of speech acts has been studied extensively for dialogue conversations starting from early 2000's with (Stolcke et al., 2000) being one of the benchmark works where the authors presented varieties of approaches such as Hidden Markov Models, Neural Networks and Decision Trees to identify dialogue acts on a benchmark dialogue data known as the Switchboard (SWBD) (Godfrey et al., 1992) dataset. In (Saha et al., 2021), authors studied the role of emotion in identifying dialogue acts for a dyadic conversation by considering thee textual and the audio modality of the utterances in the conversation. In (Saha et al., 2020b), authors proposed studying the role of emotion in determining dialogue acts on a dyadic and multi-party conversational dataset in a multi-modal framework (incorporating text, audio and video). However, tweets are unstructured and noisy communications with spelling mistakes, random coinages with limitations in expression because of character constraint per tweet. This makes it very different from face-to-face or other conversations.

3 Dataset

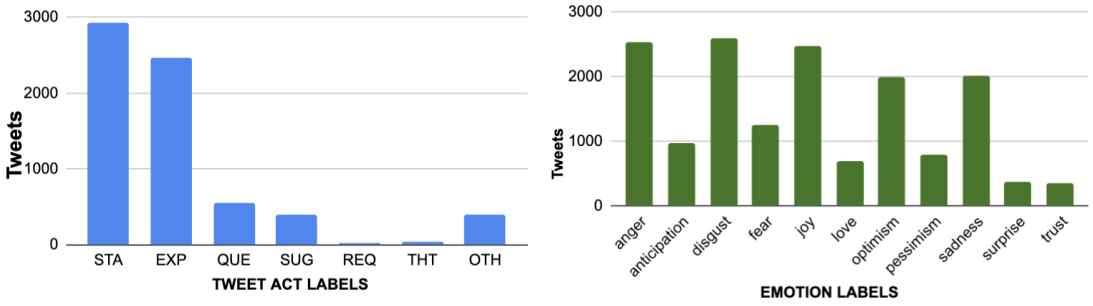
Here, we discuss the details of the newly created dataset, *EmoTA*.

3.1 Data Collection

To begin with, we scanned the literature for the latest SA and ER dataset for Twitter in order to gather potentially emotionally rich tweets to explore its impact on TAC. Initially, we came across several SA and ER datasets for Twitter such as (Oleri and Karagoz, 2016), (Mohammad and Kiritchenko, 2018), *SemEval-2018* (Mohammad et al., 2018), *BTD* (Wang et al., 2012), *TEC* (Mohammad, 2012), *CBET* (Shahraki and Zaiane, 2017), *STS-Gold* (Mohammad and Turney, 2013), *STS* (Go et al., 2009), *SS-Twitter* (Thelwall et al., 2012) etc. However, we chose to use SemEval-2018 dataset for further investigation of our task at hand. The reason behind this choice was that most of the ER datasets were annotated with only six Eckman's (Ekman, 1999) or eight Plutchik's (Plutchik, 1980) emotion categories. Whereas SemEval-2018 dataset contains tweets annotated with multi-label 11 emotion categories which aids the diversity of the problem statement. Intuitively, it was indeed possible to go the other way round and search for Twitter dataset annotated with TAs such as (Zhang et al., 2011),

Tweet	TA	Emotion	Sentiment
And it pisses me off more they killed people who surrendered. Hands up and all. If hands visible you shouldn't be fearing for your life	exp	anger, disgust, fear	negative
We're going to get City in the next round for a revenge	tht	anger	negative
Ryan Gosling and Eva Mendes finally; B joyful an funny/dont boss/dont argue/do everything with kids/go on mini car trips/ focus on love	sug	joy, love	positive
@MendipHillsAONB do we think the swallows and swifts have gone? Photo'd 3 nights ago, not seen since. #sad #Autumn	que	pessimism, sadness	negative

Table 1: Sample tweets from the EmoTA dataset with its corresponding true TA, Emotion and Sentiment labels



(a) (b)

Figure 1: (a) Distribution of tweet act labels, (b) Distribution of emotion labels.

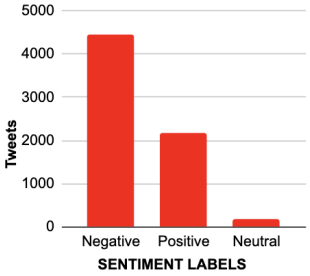


Figure 2: Distribution of sentiment labels

(Vosoughi and Roy, 2016), (Saha et al., 2020a) etc. However, the tweets in these datasets were devoid of nonverbal cues such as emojis which are quite excessively used in Twitter.

3.2 Data Annotation

To the best of our knowledge, we were unaware of any sizable and open sourced Twitter dataset annotated for its TA and emotion labels. Hence, the SemEval-2018 dataset has been manually annotated for its TA categories. Unlike dialogic conversations, there isn't a standard TA tag-set available for annotating tweets. However, we made use of 7 TA categories of (Saha et al., 2020a) for annotating SemEval-2018 dataset as opposed to 5 and 6 TA categories of (Zhang et al., 2011) and (Vosoughi and Roy, 2016), respectively. The 7 TA tags are "Statement" (sta), "Expression" (exp), "Question" (que), "Request" (req), "Suggestion" (sug), "Threat" (tht) and "Others" (oth). For the current work, we selected a subset of SemEval-2018 dataset amounting to 6810 tweets to create *EmoTA* dataset. Three annotators who were graduate in English linguistics were accredited to annotate the tweets with the appropriate TA tags. They were asked to annotate these tweets individually by only viewing the tweet available without the information of the pre-annotated emotion tags. This was done so as to assure that the dataset does not get biased by specific TA-emotion pairs. The conflicting annotations were resolved through discussions and mutual agreements. The inter-annotator score over 80% was considered as reliable agreement. It was determined based on the count that for a given tweet more than two annotators agreed on a particular tag.

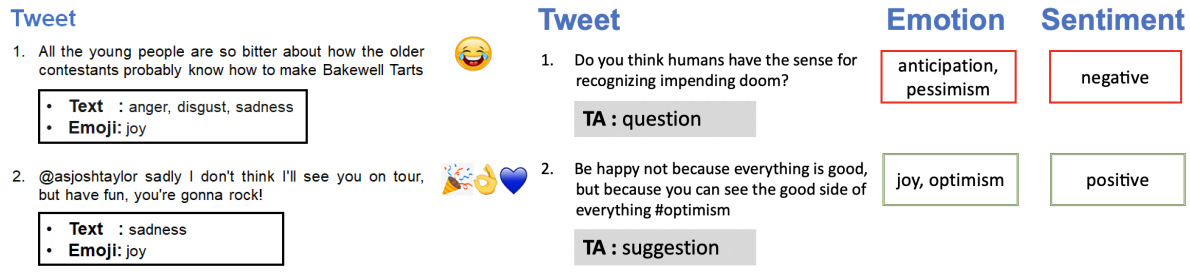
For annotating the dataset with sentiment labels, we followed a semi-supervised approach instead of manual annotation which is cost intensive. We used the IBM Watson Sentiment Classifier[2], an open-sourced API readily available for obtaining silver standard sentiment label of the tweets categorized into 3 tags namely "Positive", "Negative" and "Neutral".

3.3 Emotion-Tweet Act Dataset : *EmoTA*

The *EmoTA* dataset[3] now comprises of 6810 tweets with the corresponding gold standard TA and multi-label emotion tags. Each of the tweet contains its Tweet ID and two modalities: text and emoji. Few sample tweets along with the corresponding TA, sentiment and emotion labels from the proposed dataset are shown in Table 1. Distributions of TA,

[2]https://cloud.ibm.com/apidocs/natural-language-understanding#sentiment

[3]The dataset with its TA and emotion tags will be made publicly available to the research community.



(a) (b)

Figure 3: (a) Importance of emoji in analysis of tweets, (b) Importance of emotion and sentiment in TAC.

emotion and sentiment labels across the dataset are shown in Figure 1a, 1b and 2, respectively.

3.4 Qualitative Aspects

Below, we analyze using some samples from the dataset that require sentiment-emotion aided and multi-modal reasoning.

Role of Sentiment and Emotion. In Figure 3b, we demonstrate using two examples from the dataset to establish our hypothesis that sentiment and emotional states of the tweeter can aid the identification of TAs. In the first instance, the tweeter questions about the impending doom supposedly because of a pessimistic expectation arising due to the negative sentiment. Similarly, in the second instance, because of a joyous emotion emerging due to positive sentiment, the tweeter shares an optimistic suggestion with the readers. The above examples highlight the need for incorporating these additional user behavior, i.e., sentiment and emotion while reasoning about TAs. Thus, stressing the requirement of addressing such synergy amongst TAC, SA and ER.

Role of Multi-modality. In Figure 3a, we present two examples from the dataset to highlight the importance of including other nonverbal features such as emoji present in the tweet along with the text for several tweet analysis tasks. In the first example tweet, the text represents an overall negative sentiment with emotion such as anger and disgust. However, the presence of an emoji face with tears of joy gives it an emotion of joy along with the other emotions. Similarly, in the second example tweet, the text represents the emotional state of the tweeter as sad, whereas the ok, celebration and heart emojis depict the feeling of joy. These instances show that the presence of complementary information in the form of emojis aids the process of any twitter analysis task including TAC.

4 Proposed Methodology

The proposed multi-tasking, multi-modal approach and implementation details are outlined in this section.

4.1 Feature Extraction

The procedure for feature extraction across multiple modalities is discussed below.

Textual Features. To extract textual features of a tweet U having n_u number of words, the representation of each of the words, $w_1, ..., w_u$, where $w_i \in \mathbb{R}^{d_u}$ and w_i's are obtained from BERT (Devlin et al., 2019) which is a multi-layered attention aided bidirectional Transformer Encoder model based on the original Transformer model (Vaswani et al., 2017) where $d_u = 768$.

Emoji Features. To extract emoji features from a tweet, we use *emoji*, a python based library for eliciting the pictorial image of an emoji (primarily that of a face, object or symbols). A total of 1816 kind of emojis are available along with its different types. We then use *emoji2vec* (Eisner et al., 2016), which provides $d_v = 300$ dimensional vector representation for each of the emojis present in the tweet. Let's say a tweet contains n_v number of emoji. Thus, we obtain the final emoji representation V for a tweet as $V \in \mathbb{R}^{n_v \times d_v}$.

4.2 Network Architecture

The proposed network consists of four main components : (i) *Modality Encoders* (ME) produces respective modality encodings by taking as input the uni-modal features extracted above, (ii) *Dyadic Attention Mechanism* (DAM) that comprises dual attention mechanisms such as *intra-modal* and *inter-modal* attentions, (iii) *Adversarial Loss* to make the feature spaces of task-specific and shared layers of each task mutually exclusive, (iv) *Classification Layer* that contains output channels for the three tasks at hand (TAC, SA and ER) to learn generalized representations across all the tasks.

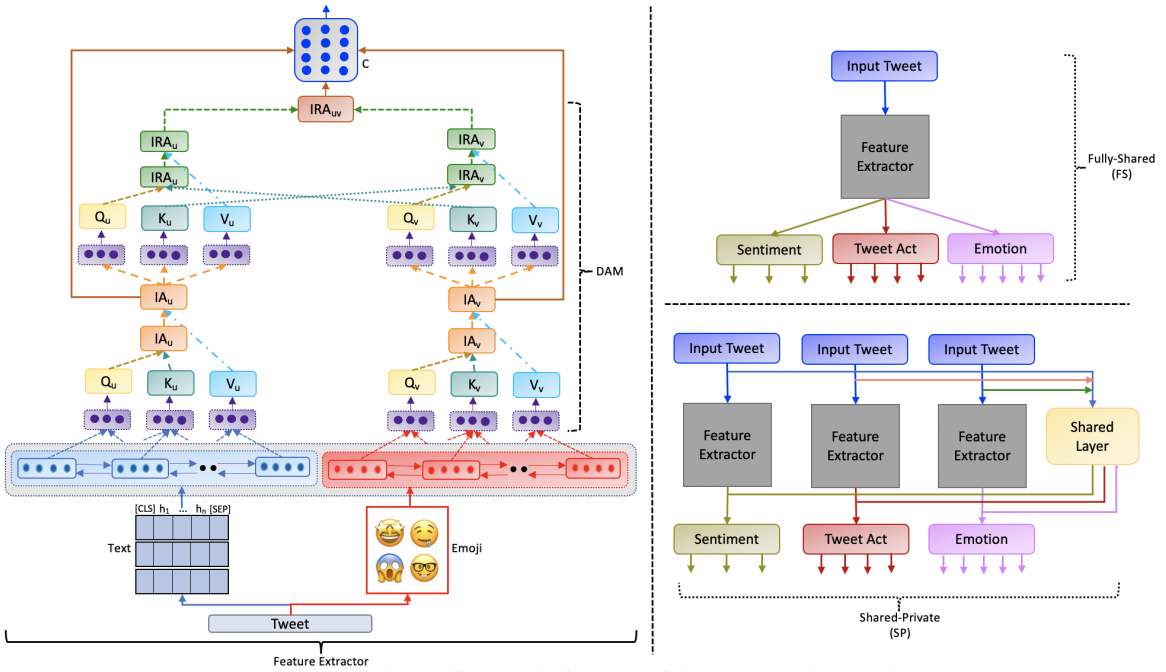


Figure 4: The architectural diagram of the proposed network

4.2.1 Modality Encoders

In this section we discuss how the architectural framework encodes different modalities.

Text and Emoji Modalities. The features U and V obtained from each of the modalities corresponding to a tweet (discussed above) are then passed through two discrete Bi-directional LSTMs (Bi-LSTMs) (Hochreiter and Schmidhuber, 1997) to sequentially encode these representations and learn complementary semantic dependency based features into hidden states from these modalities. In case of textual modality (say), the final hidden state matrix of a tweet is obtained as $H_u \in \mathbb{R}^{n_u \times 2d_l}$. d_l represents the number of hidden units in each LSTM and n_u is the sequence length. In the similar way, a representation of corresponding emoji modality encoding as $H_v \in \mathbb{R}^{n_v \times 2d_l}$ is obtained. The number of representations from modality encoders vary depending on the variant of the multi-task learning framework used (e.g., fully shared (FS) or shared-private model (SP)). In a FS variant, two representations are obtained one for text and another for emoji cumulatively for optimizing all the three tasks. However, for a SP model, six encoding representations are obtained. Three for text and the remaining for emoji forming a pair of text-emoji representations for each of the three tasks.

4.2.2 Dyadic Attention Mechanism

We use a similar concept as in (Vaswani et al., 2017), where the authors proposed to compute attention as mapping a query and a set of key-value pairs to an output. So, the representations obtained from the modality encoders above are passed through three fully-connected layers each termed as queries and keys of dimension $d_k = d_f$ and values of dimension $d_v = d_f$. For a FS model, we have two triplets of (Q, K, V) as : (Q_u, K_u, V_u) and (Q_v, K_v, V_v). Similarly for a SP model, we have six such triplets as : (Q_{u1}, K_{u1}, V_{u1}), (Q_{v1}, K_{v1}, V_{v1}), (Q_{u2}, K_{u2}, V_{u2}), (Q_{v2}, K_{v2}, V_{v2}), (Q_{u3}, K_{u3}, V_{u3}), (Q_{v3}, K_{v3}, V_{v3}) where pair of two triplets are from the textual and emoji modality encoders for each of the tasks[4]. These triplets are then used to compute attention values for different purposes in various combinations which include intra attention and inter-modal attention.

Intra-modal Attention. We compute intra-modal attention (IA) for all these individual modalities in order to learn the interdependence between the current words and the preceding part of the tweet. In a way, we aim to relate different positions of a single sequence to estimate a final representation of the same sequence for individual modalities (Vaswani et al., 2017). Thus, the IA scores for individual modalities are calculated as :

$$IA_j = softmax(Q_j K_j^T) V_j \quad (1)$$

[4]Subscript 1, 2 and 3 represent TAC, ER and SA task, respectively.

where $IA \in \mathbb{R}^{n_u \times d_f}$ for IA_u, $IA \in \mathbb{R}^{n_v \times d_f}$ for IA_v for FS model and six such IA scores for SP model.

Inter-modal Attention. The IA scores obtained above are then used to compute inter-modal attention (IRA). We re-iterate the same process (explained above) to now form triplets of (Q, K, V) for these IA scores and then compute IRA scores amongst triplets of all IA scores by computing the matrix multiplication of combination of queries and keys of different IA modality scores using Equation 1. In this manner, we obtain one IRA score as $IRA_{uv} \in \mathbb{R}^{n_u \times d_f}$ for FS variant and three IRA scores for SP model as IRA_{uv1}, IRA_{uv2} and IRA_{uv3}. This is done to distinguish important contributions between various modalities to achieve optimal representation of a tweet.

Attention Fusion. Next, we concatenate each of these computed IA and IRA vectors as :

$$C = concat(IRA_{uv}, IA_u, IA_v), \text{for FS} \quad (2)$$

$$C_1 = concat(IRA_{uv1}, IA_{u1}, IA_{v1}), \text{for SP} \quad (3)$$

$$C_2 = concat(IRA_{uv2}, IA_{u2}, IA_{v2}), \text{for SP} \quad (4)$$

$$C_3 = concat(IRA_{uv3}, IA_{u3}, IA_{v3}), \text{for SP} \quad (5)$$

Next, we obtain mean of these three different concatenated attention vectors for the SP variant or directly use the obtained C attention vector for the FS variant to obtain the final representation of a tweet.

$$M = mean(C_1, C_2, C_3) \quad (6)$$

Shared Layer. Additionally, for the SP model, other than having task-specific layers, we allow a shared layer to learn task invariant features. Here, the shared layer is in the form of a fully-connected layer of dimension d_f. The inputs to the shared layer are the hidden representations of three IRA vectors : IRA_{uv1}, IRA_{uv2} and IRA_{uv3}. Thus for a given tweet, the loss of the shared layer is minimized if the model correctly classifies the tasks of each of the tweets in the input. This helps learn domain invariant feature space for different tasks.

Adversarial Loss. The goal of this adversarial loss function is to tune the weights of the shared layer so that it learns a representation that misleads the task discriminator. The adversarial loss l_{adv}, aims to make the feature space of shared and task-specific layers to be mutually exclusive (Liu et al., 2017). We follow the similar strategy as that of (Liu et al., 2017), where a task discriminator D (say) maps the shared feature to its original task. Thus, on a correct prediction when the loss at the shared layer decreases, the adversarial loss increases and vice-versa. Alternatively, the shared layer is tuned to work in an adversarial way, thereby prohibiting the discriminator to predict one of the three tasks. The adversarial loss is computed as :

$$l_{adv} = \min_{F}(\max_{D}(\sum_{n=1}^{N}\sum_{k=1}^{K} d_k^n \log[D(F(x_k^n))])) \quad (7)$$

where d_k^n represents the true label amongst the type of the tasks, N, and x_k^n is the k^{th} example for task n. The min-max optimization problem is addressed by the gradient reversal layer (Ganin and Lempitsky, 2015).

4.2.3 Classification Layer

The final representation of the tweet obtained from the DAM module is shared across three channels pertaining to the three tasks, i.e., TAC, SA and ER (for FS model) and three DAM representations for three individual tasks are subjected to individual output layer (for SP model). The task-specific loss (l_t), shared loss (l_s) and adversarial loss (l_{adv}) are used as

$$l_f = l_t + \alpha l_s + \gamma l_{adv}, \text{for SP model} \quad (8)$$

$$l_f = l_s + \gamma l_{adv}, \text{for FS model} \quad (9)$$

where α and γ are hyper-parameters.

4.3 Experimentation Details

Hyper-parameters. 80% of the tweets of the *EmoTA* dataset were used for training and the remaining 20% were used for testing the models. The same training and testing data were used for all the experiments in order to ensure fair comparison of models. To encode different modalities, a Bi-LSTM layer with 100 memory cells was used. Dense layers of dimensions 100 were used for d_f. The three channels contain 7, 3 and 11 output neurons, for TA, sentiment and emotion tags, respectively. *Categorical crossentropy* loss is used for TA and sentiment channels and *Binary crossentropy* loss function is used for emotion channel. A learning rate of 0.01 and *Adam* optimizer were used in the final experimental setting. *All these values of the parameters were selected after a careful sensitivity analysis.*

Pre-processing. We employ NLTK based *TweetTokenizer* to tokenize tweets. Urls were removed. User mentions were replaced by <user> token. Numbers occurring in the tweet were replaced by

Model	TAC + SA								TAC + ER								TAC + SA + ER							
	Five-Class				Seven-Class				Five-Class				Seven-Class				Five-Class				Seven-Class			
	Text		Text+ Emoji		Text		Text+ Emoji		Text		Text+ Emoji		Text		Text+ Emoji		Text		Text+ Emoji		Text		Text+ Emoji	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
FS	72.06	69.87	74.73	72.02	62.25	59.66	66.85	64.35	73.72	71.05	76.60	74.32	63.58	61.00	68.73	66.20	78.01	75.85	78.16	76.01	71.29	68.85	75.62	73.20
FS + Adv	73.92	71.05	75.61	73.32	63.67	61.27	69.54	67.03	75.57	73.05	77.35	75.00	65.11	62.80	71.24	69.02	80.01	77.59	81.34	79.08	72.90	70.51	76.21	73.95
SP	73.41	70.91	76.81	74.52	62.71	60.25	67.62	65.28	75.05	72.85	77.12	74.93	64.63	62.35	69.30	67.02	78.41	76.00	80.68	78.28	72.02	69.90	76.50	74.33
SP + Adv (without DAM)	74.73	72.06	75.86	73.33	64.13	61.75	70.32	68.04	76.11	73.80	77.57	75.20	65.80	63.16	71.86	69.60	80.32	78.00	81.49	79.14	73.24	70.90	77.60	75.28
SP + Adv (Glove)	73.82	71.22	77.27	75.00	66.71	64.46	69.94	67.61	75.61	73.28	78.42	76.05	68.81	66.36	72.26	69.83	79.35	77.15	81.79	79.46	73.31	70.90	78.17	76.00
SP + Adv (only IA)	76.21	73.85	78.62	76.35	69.73	67.30	71.75	69.50	77.64	75.21	80.68	78.37	71.07	68.95	73.05	71.00	81.64	79.27	83.04	81.16	75.62	73.35	79.95	77.62
SP + Adv (only IRA)	-	-	78.75	76.30	-	-	72.17	70.05	-	-	80.82	78.55	-	-	73.59	71.29	-	-	83.49	81.15	-	-	80.10	78.02
SP + Adv (with DAM)	76.21	73.85	**79.37**†	**77.01**	69.73	67.30	**72.90**†	**70.63**	77.64	75.21	**80.97**†	**78.70**	71.07	68.95	**74.08**†	**72.00**	81.64	79.27	**84.08**†	**81.85**	75.62	73.35	**80.32**†	**78.16**

Table 2: Results of all the baselines and the proposed multi-task models in terms of accuracy and weighted F1-score. † indicates that the reported results are statistically significant

Model	Single Task TAC							
	Five-Class				Seven-Class			
	Text		Text+Emoji		Text		Text+Emoji	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
FS (without DAM)	70.31	68.60	72.60	70.25	59.86	57.20	64.21	62.10
FS (only IA)	72.61	70.30	73.35	71.17	64.02	62.90	67.16	65.00
FS (only IRA)	-	-	73.74	71.26	-	-	67.70	65.28
FS (with DAM)	72.61	70.30	**74.73**	**72.15**	64.02	62.90	**68.57**	**66.15**
FS (Glove)	71.05	68.74	72.23	69.90	60.68	58.62	66.17	64.00
FS (emoji as text)	74.09	71.91	-	-	66.66	64.12	-	-
FS (with DAM) (sentiment as feature)	74.16	71.75	76.35	74.20	66.24	64.15	69.73	67.25
FS (with DAM) (emotion as feature)	75.35	73.34	78.50	75.82	68.03	65.83	71.46	69.19
FS (with DAM) (sentiment & emotion as features)	77.42	74.81	79.05	76.74	68.83	66.12	72.21	70.03

Table 3: Results of the single task TAC models in varying combinations

<number> token. *Ekphrasis* (Baziotis et al., 2017) was used to extract hashtags by segmenting long string into its constituent words. All the characters of the tweet were lower-cased. Since the dataset is under-represented for most of the TA tags, we over-sample 80% of the tweets used for training as : the mediocrely represented tags (e.g., sug, que and oth) are over-sampled to be equally represented as the most represented tags (e.g., sta and exp). Similarly, the highly under-represented classes (e.g., req and tht) are over-sampled to be equally represented as the mediocrely represented tags in the *EmoTA* dataset. All the results reported below are on the 20% test data without any over-sampling.

5 Results and Analysis

A series of experiments were conducted for evaluating the proposed approach. Experiments were conducted for single task and several combinations of multi-task framework with TAC being the pivotal task along with varying modalities. A thorough ablation study is performed to analyze the importance of each of the attention mechanisms of the proposed architectural framework along with several variations of multi-task learning (e.g., FS, SP etc.). *Note that we aim to enhance the performance of TAC with the help of other two auxiliary tasks. Following this, we report results and analysis with TAC strictly being the pivotal task in all the task combinations.* Since, the dataset is unbalanced for all the task categories, we report results for different dimensions of TAC in the following set-up:

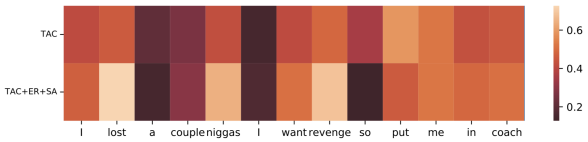


Figure 5: The visualization of the learned weights for a tweet from IA_u layer- u_1: "I lost a couple niggas I want revenge so put me in coach." for single task TAC (baseline), multi-task TAC+SA+ER (proposed) models

Model	SA & ER							
	Five-Class				Seven-Class			
	Text		Text+Emoji		Text		Text+Emoji	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Single Task SA	87.26	86.05	88.52	87.20	88.85	87.30	90.10	89.00
Single Task ER	81.57	60.77	84.63	64.32	80.07	73.51	81.58	76.63
SA + TAC	89.31	87.85	90.74	89.09	89.60	88.75	91.55	90.35
ER + TAC	83.52	65.86	86.09	67.02	81.37	75.21	84.21	78.30
SA + ER (for SA)	92.30	91.06	93.02	92.00	90.33	88.65	92.73	90.37
SA + ER (for ER)	84.61	68.77	87.37	70.19	82.72	70.00	85.30	72.04
SA + ER + TAC (for SA)	92.06	91.13	93.19	92.38	92.49	90.53	93.68	91.82
SA + ER + TAC (for ER)	85.39	70.04	88.31	72.77	83.26	79.66	86.01	81.05

Table 4: Results of the proposed model for the single and multi-task SA and ER

- **Five-class** Classification : This includes the top 5 highly occurring TA tags namely sta, exp, que, sug and oth.
- **Seven-class** Classification : This includes all the 7 categories of TAs used in the annotation process.

Table 3 and 2 illustrate the results of single task TAC and varying combinations of multi-task proposed models for different set-up (as mentioned

Tweet	True	TAC	TAC+SA	TAC+ER	TAC+SA+ER
@BreezyWeekes hey breezy, you wanna give me some of that coffee you posted on your snap?? please	req	exp	req	req	req
We're going to get City in the next round for a revenge	tht	sta	exp	tht	tht
@voguemagazine, did you not learn from @FreePeople 's viral insult to ballet? Stop trying to wrongfully stick models into pointe shoes	sug	que	exp	exp	sug
I wonder if Corey will vote for Nicole?? #snacole #bb18 #paulsgonnawin #finale #halfamill	que	que	que	que	que

Table 5: Sample tweets from the EmoTA dataset with its corresponding ground truth and predicted labels for different single and multi-task models

Model	Acc.	F1
JointDAS (TAC + SA) (Cerisara et al., 2018)	59.05	57.60
CNN-SVM (TAC) (Saha et al., 2019)	61.32	59.75
Transformer (TAC) (Saha et al., 2020c)	65.46	63.65
Bert-Caps (TAC) (Saha et al., 2020a)	67.10	65.00
Proposed (TAC)	**68.57**	**66.15**

Table 6: Comparative Analysis with the state of the art models

above). As evident, the addition of non-verbal cues in the form of emojis improves the uni-modal textual baseline consistently. This improvement implies that the proposed architecture utilizes the interaction among the input modalities very effectively. This highlights the importance of incorporating multi-modal features for different Twitter analysis tasks. We also report result for utilizing emoji as textual feature instead of treating it as a different modality in the single task TAC framework. Also, the five-class set-up gave better results than the seven-class set up. This is pretty obvious, as with 5-class set-up, the model needs to distinguish and identify lesser fine-grained features compared to the 7-class set-up. Additionally, the under-representation of two tags in the *EmoTA* dataset for the 7-class set-up also effects its performance.

As seen in Table 2, the multi-task framework with all the three tasks (i.e., TAC + SA + ER) consistently gave better results as compared to single task TAC. In the bi-task variant, TAC+SA, shows little improvement in different metrics as opposed to TAC+ER over and above the single task TAC. This gain is rather intuitive as sentiment alone is sometimes unable to convey complete information of the tweeter's state of mind. E.g., a *negative* sentiment can occur because of various emotions such as *disgust*, *fear*, *sadness* etc. Similarly, a *positive* sentiment can take place because of emotions such as *happiness*, *surprise* etc. Thus, with sentiment alone, sometimes this discreteness or fine differences in the state of mind cannot be completely determined and conveyed. To illustrate this, in Figure 5, we provide a visualization of the learned weights of a tweet from the IA_u layer (as this layer contains word-wise attention scores). For this particular tweet, its true TA label is *tht*. With the multi-task framework, the importance of warning bearing words are learnt well such as *lost, revenge* compared to the single-task TAC where attention is laid on expression bearing word such as *put me*. Additionally, we also report results for cases where sentiment and emotion were directly used as features in the single task TAC models to leverage from instead of deploying a multi-task based approach in Table 3.

As stated above, we treat SA and ER as auxiliary tasks aiding the primary task, i.e., TAC. However, we report the performance of SA and ER tasks on the proposed model for single as well as multi-task frameworks in Table 4 for further investigations. However, we do not make any explicit effort to enhance their performance.

Comparison amongst Different Multi-task Architecture. In terms of varying ways of multi-tasking such as FS, SP along with adversarial loss (adv), it was observed that SP model gave better results compared to FS model. Additionally, incorporating adversarial loss further boosted the performance of different multi-task models. Intuitively, as TAC shares lesser amount of correlation with SA and ER compared to SA and ER themselves, FS model was not sufficient enough to learn diverse features across different tasks. This observation is in conformity with the existing literature. We also demonstrate the importance of different attentions used for the best performing multi-task model, i.e., SP+Adv. Furthermore, we also report results by replacing BERT model to extract textual representation with Glove embeddings (Pennington et al., 2014). Results indicate that each of these aspects contributed significantly to aid the performance of the proposed multi-tasking framework. *All the reported results here are statistically significant* (Welch, 1947).

Comparison with the State of the Art Models. We also compare our proposed approach with the recent state of the art models for single task TAC as we are unaware of any other work which jointly optimized tweet act, emotion and sentiment in Twitter. In Table 6, we report the results for the same by re-implementing those on the *EmoTA* dataset. As

evident, the proposed model outperformed these SOTA approaches.

Error Analysis. An in-depth analysis revealed several scenarios as to why the proposed model faltered which are as follows : ***i.*** **Imbalanced Dataset** : As visible in Figure 1a, except for "sta" and "exp" tags, all the classes are under-represented in the *EmoTA* dataset. Even though we apply over-sampling to partially counter this issue but still the tags such as "req" and "tht" contain very little tweets for the model to learn fine differences amongst different categories. In accordance with this, we observe that five-class performs exceptionally better than the seven-class classification set-up; ***ii.*** **Fine-grained tags** : It was also observed that the tweets which were mis-classified were subset of each other. For instance, tweet such as "*don't get discouraged! it's early on; it can get overwhelming. keep reading; use cue cards it'll get better!!*" is wrongly predicted as "exp" rather than "sug" which in the superficial way is a subset of the former tag; ***iii.*** **Miscellaneous** : Tweets belonging to "oth" tag was also majorly mis-classified as there was no fixed pattern of tweets belonging to this category. To counter this, even more fine-grained categories of TAs needs to be identified and modelled. *Sample utterances for the error analysis are shown in Table 5.*

6 Conclusion and Future Work

In this paper, we studied the role of sentiment and emotion in speech act classification in Twitter. We curate a novel dataset *EmoTA*, that contains pre-annotated tweets with emotions collected from open-source dataset and annotated with TAs and sentiment categories. We propose a Dyadic Attention Mechanism based multi-modal (emojis and text), adversarial multi-task framework for joint optimization of TAs, sentiment and emotions. The DAM (dyadic attention mechanism) module employs intra-modal and inter-modal attention to fuse multiple modalities and learn generalized features across all the tasks. Results show that multi-modality and multi-tasking boosted the performance of TA identification compared to its uni-modal and single task TAC variants.

In future, attempts will be made to predict TAs with more precision by incorporating fine-grained modality encodings and also identifying which other NLP tasks (e.g., named entity recognition) might assist TAC as a task.

Acknowledgments

Dr. Sriparna Saha gratefully acknowledges the Young Faculty Research Fellowship (YFRF) Award, supported by Visvesvaraya Ph.D. Scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia) for carrying out this research.

References

Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. 2018. Emotion recognition in speech using cross-modal transfer in the wild. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 292–301.

Lisa Feldman. Barrett, Michael Lewis, and Jeannette M. Haviland-Jones. 1993. *Handbook of emotions*. The Guilford Press.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada. Association for Computational Linguistics.

Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, and Hoa Le. 2018. Multi-task dialog act and sentiment recognition on mastodon. *arXiv preprint arXiv:1807.05013*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Hai Ha Do, PWC Prasad, Angelika Maag, and Abeer Alsadoon. 2019. Deep learning for aspect-based sentiment analysis: a comparative review. *Expert Systems with Applications*, 118:272–299.

Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bosnjak, and Sebastian Riedel. 2016. emoji2vec: Learning emoji representations from their description. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media, SocialNLP@EMNLP 2016, Austin, TX, USA, November 1, 2016*, pages 48–54. Association for Computational Linguistics.

Paul Ekman. 1999. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1615–1625. Association for Computational Linguistics.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE.

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

M Shamim Hossain and Ghulam Muhammad. 2019. Emotion recognition using deep learning approach from audio–visual emotional big data. *Information Fusion*, 49:69–78.

Minwoo Jeong, Chin-Yew Lin, and Gary Geunbae Lee. 2009. Semi-supervised speech act recognition in emails and forums. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1250–1259. Association for Computational Linguistics.

Abhishek Kumar, Asif Ekbal, Daisuke Kawahra, and Sadao Kurohashi. 2019. Emotion helps sentiment: A multi-task model for sentiment and emotion analysis. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1–10. Association for Computational Linguistics.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825.

Saif Mohammad. 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)*, pages 246–255.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

Saif Mohammad and Svetlana Kiritchenko. 2018. Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Obrahim Oleri and Pinar Karagoz. 2016. Detecting user emotions in twitter through collective classification. In *KDIR*, pages 205–212.

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. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

Tulika Saha, Dhawal Gupta, Sriparna Saha, and Pushpak Bhattacharyya. 2021. Emotion aided dialogue act classification for task-independent conversations in a multi-modal framework. *Cognitive Computation*, 13(2):277–289.

Tulika Saha, Srivatsa Ramesh Jayashree, Sriparna Saha, and Pushpak Bhattacharyya. 2020a. Bert-caps: A transformer-based capsule network for tweet act classification. *IEEE Transactions on Computational Social Systems*, 7(5):1168–1179.

Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya. 2020b. Towards emotion-aided multi-modal dialogue act classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4361–4372.

Tulika Saha, Aditya Prakash Patra, Sriparna Saha, and Pushpak Bhattacharyya. 2020c. A transformer based approach for identification of tweet acts. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Tulika Saha, Sriparna Saha, and Pushpak Bhattacharyya. 2019. Tweet act classification: A deep learning based classifier for recognizing speech acts in twitter. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Ameneh Gholipour Shahraki and Osmar R Zaiane. 2017. Lexical and learning-based emotion mining from text. In *Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing*, volume 9, pages 24–55.

Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2012. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Soroush Vosoughi. 2015. *Automatic detection and verification of rumors on Twitter*. Ph.D. thesis, Massachusetts Institute of Technology.

Soroush Vosoughi and Deb Roy. 2016. Tweet acts: A speech act classifier for twitter. In *Tenth International AAAI Conference on Web and Social Media*.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2012. Harnessing twitter" big data" for automatic emotion identification. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 587–592. IEEE.

Bernard L Welch. 1947. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35.

Renxian Zhang, Dehong Gao, and Wenjie Li. 2011. What are tweeters doing: Recognizing speech acts in twitter. *Analyzing Microtext*, 11(05).