A Transformer based Approach for Identification of Tweet Acts

Tulika Saha*[1], Aditya Prakash Patra[1], Sriparna Saha[1], Pushpak Bhattacharyya[1]
[1]Department of Computer Science and Engineering
Indian Institute of Technology Patna
Bihar, India 801103
*Email: sahatulika15@gmail.com

***Abstract*—Speech acts help in uncovering and understanding the communicative intention and behavior of a speaker utterance. This is pertinent for communication on any platform that embodies social media platforms like Twitter. This paper presents a supervised speech act (tweet act in our case) classifier for tweets for assessing the content and intent of tweets, thereby, exploring the valuable communication amongst the tweeters. With the recent success of Bi-directional Encoder Representations from Transformers (BERT), a newly introduced language representation model that provides pre-trained deep bi-directional representations of vast unlabeled data, we introduce *BERT-extended* that is built on top of BERT. Our model is based on calculating attention weights over the representations of tokens of a sequence to identify tweet acts. The proposed model attained a benchmark accuracy of 75.97% and outperformed several strong baselines and state-of-the-art approaches on an open-source, tweet act annotated Twitter dataset.**

***Index Terms*—Twitter, Tweet Acts, BERT, Attention**

I. Introduction

Various social media and micro-blogging platforms such as Twitter, MySpace, Facebook and Tumblr etc. have become immensely prevalent as communication tools among Internet users. With hundreds and millions of messages exchanged on a daily basis, these social networks form a dynamic and valuable source for studying the interests of individuals and for analyzing user generated contents. Among numerous social media platforms, Twitter is one of the leading micro-blogging services with people discussing current issues, sharing opinions, suggestions and views about various topics, sharing facts, information and news, writing about their life, asking for solutions to queries and so on. Twitter is being used in such a magnitude that in 2016, there were approximately 325 million monthly active users and more than 500 million tweets (status messages) were created on a daily basis [1]. Twitter's enormous volume and its public nature has made it a prudent source of data to study user behavior and nature. Therefore, the crux of this paper is to analyze Twitter content, i.e., what people are tweeting about.

A reasonable amount of research has been carried out for analyzing the linguistic essence of tweets [2], [3], [4] but very little research has been carried out on understanding the pragmatics or semantics of tweets. Pragmatics capture the intended meaning of an utterance and look beyond the literal meaning. Thus, it captures and identifies the behavior and communicative intention of an user utterance. A very well known and accepted formalization for understanding pragmatics is called "Speech Act Theory", which was proposed by [5] and advanced by [6]. The theory introduced amongst various other features, a precise and a definite taxonomy of communicative acts known as speech acts [7]. Thus, the current paper primarily addresses the task of recognizing such speech acts in Twitter for studying the pragmatics and thereby understanding the tweet contents automatically.

The recognition of speech acts in an automated framework has compelling influences on Twitter and tweeters. For the Twitter itself, it helps to determine what constitutes a particular topic or subject with respect to speech acts and if there is a divergence in a specific topic. It helps in social media monitoring by diagnosing topic alteration or spamming. It allows the readers of the platform to follow and scan for a particular topic with the most beneficial speech act based on their needs. Consequently, it assists them to reduce their search space and allow them to become frequent readers amongst millions of tweets. It also provides a better sense of understanding for the tweeters of the user's behaviour, attitude and general notion etc.

A wide-range of works have been carried out for analyzing speech acts in the context of dialogues known as Dialogue Act Classification (DAC) in computational linguistics which includes notable works such as [8], [9], etc. However, given the limited tweet length (now 280 characters, initially it was 140), noisy, unusual and peculiar nature of tweets makes it ineligible for employing typical data mining and information retrieval methods.

Recently, the introduction of BERT, a language representation model has established state-of-the-art results for a vast spectrum of Natural Language Processing tasks. However, there has been very little to no effort in exploring BERT with respect to data in electronic mode for eg., tweets. In this paper, we present a BERT based model named *BERT-extended* for the identification of tweet acts. Our proposed approach is based on calculating attention over the word representation obtained from the BERT model. This notion stems from the fact that not all tweet words or tokens contribute equally to the identification of a particular tag. Empirically, we show that our proposed approach outperforms several strong baselines and state-of-the-art models significantly.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

The key contributions of this paper are the following :

- *This paper presents a novel speech act classifier for tweets named as **BERT-extended** which is built on top of BERT*;
- *The model leverages from the calculation of attention over the word representation obtained from the BERT model for specific tweet acts to learn cumulative features pertaining to speech acts and Twitter*;
- *The proposed model attained state-of-the-art results for the task of tweet act classification*.

Remaining of the paper is arranged as follows : Section II introduces a short formal description of the problem statement. Section III, gives a concise summary of the related works and the motivation behind solving this problem. The details of the proposed methodology has been demonstrated in Section IV. Section V presents the implementation details. The empirical results and the detailed analysis of the same are presented in Section VI. Lastly, the conclusion and the directions for future work are explained in Section VII.

II. Problem Statement

The main objective of this particular task is to identify speech acts in Twitter, i.e., tweet act classification. Given a tweet X, the task is to assign the most appropriate tweet act (say y) among a set of tags ($Y = \{y_1, y_2, ...y_i\}$ where i is the number of tweet acts). Thus, it is a multi-class classification problem. Formally, it can be represented as

$$y = argmax_{y' \in Y} P(y'|X) \tag{1}$$

where $P(y'|X)$ is the probability of assigning tweet act, y', to the given tweet, X. We acknowledge the fact that in real-life situations this presumption may not always hold and that one tweet may demonstrate multiple speech acts. But because of the limited and short length of tweets, tweets with multiple speech acts are infrequent and exceptionally rare to obtain and thus, we consider this simplifying assumption competent in curtailing the complexity of the given problem statement.

III. Related works

A. Background

An extensive literature survey was carried out to explore various work done on Tweet Act Classification. In [10], authors proposed Support Vector Machine (SVM) based approach to model the task of tweet act classification. They employed their approach on manually annotated data-set consisting of 8613 tweets with the proposed tag-set of five tweet acts namely "Comment", "Statement", "Suggestion", "Question" and "Miscellaneous". Their model was based on manually extracted hand-crafted features from the tweets such as cue words and phrases which included maximum occurrences of unigrams, bigrams and trigrams, some non-cue words which included opinion words, emoticons, abbreviations etc. followed by some character based features such as twitter-specific characters and punctuations. Their proposed method achieved F1-score of 0.7. Later in [11], authors proposed Logistic Regression (LR) based speech act classifier for tweets. Their work also employed manually extracted features such as semantic level features which included speech act verbs, N-grams (typically unigrams, bigrams and trigrams), emoticons etc. followed by syntactic level features such as abbreviations, twitter-specific symbols and so on along with dependency sub-trees and part of speech (typically interjections and adjectives). Authors of these work also manually annotated a data-set of nearly 7500 tweets with six tweet acts which are "Expression", "Assertion", "Request", "Recommendation", "Question" and "Miscellaneous". They reported F1-score of 0.7 based on their approach.

In [12], authors presented first-ever and the only deep learning based tweet act classifier. Their approach was a Convolutional Neural Network (CNN) based model with SVM loss function to address the multi-class classification problem more efficiently. They also incorporated few hand-crafted features to boost the robustness of their proposed model. Along with it they released an open-source manually annotated data-set with seven tweet acts. More details of the data-set is discussed in Section V, since this data-set has been utilized to demonstrate the current work. In [13], 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 [14], authors proposed a multi-task approach for joint sentiment and tweet act classification. They aimed to demonstrate that transfer learning can be efficiently achieved between these tasks and analyzed some specific correlation between these two tasks in social media platform. Whereas, our work is purely based on learning a classifier for tweet act classification without any transfer learning or multi-task based scenarios.

Apart from these, identification of speech acts has been studied extensively for dialogue conversations starting from early 2000's. In [8], 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) [15] data-set. In [16] authors presented a Naive Bayes based dialogue act classifier. Later with the advancement of deep learning, several neural networks based approaches were widely proposed. In [9], authors presented a stacked Long Short Term Memory (LSTM) based approach on the SWBD data-set. In [17], authors presented a hierarchical encoder based approach to model the task of dialogue act classification. Other prominent works in this context includes those of [18], [19], [20], [21], [22], [23] etc. In [24], authors presented a semi-supervised approach to identify speech acts in emails and different forums. These works, however, use datasets that comprises 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.

B. Motivation

Below are some observations that were made after an exhaustive literature analysis and these factors motivated us

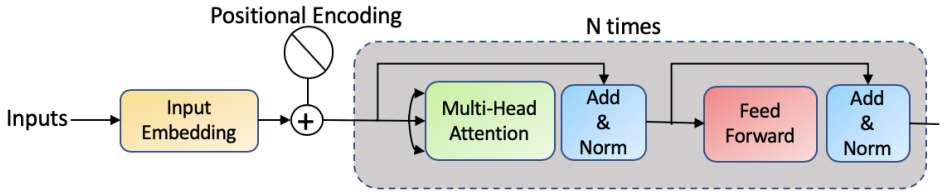

Fig. 1. The basic architecture of the Transformer Encoder model

to study the task of tweet act classification :

- Few works that are done for tweet act classification employ considerable amount of manual feature engineering to formulate fine grained hand-crafted and Twitter-specific features.
- The existing approaches make the entire process extremely tedious, time-consuming and inept.
- An even higher amount of robustness is required for the evaluation of tweets with regards to accuracy and precision.
- Tweets are replete with random coinages, spelling mistakes, collective usage of letters and symbols etc. Thus, existing approaches for Dialogue Act Classification (DAC) cannot be directly applied to the micro-blogging platform because of the noisy, informal and limited length of the tweets as opposed to the telephonic or face to face data-set used for DAC.
- Thus, there is clearly a dearth of works that address the task of tweet act classification as it forms a significant means for social content monitoring.

IV. Proposed Methodology

In this section, we first introduce the BERT model concisely followed by the extended BERT-extended model proposed for our task.

a) ***BERT****:* BERT [25] is a multi-layered attention aided bidirectional Transformer Encoder model based on the original Transformer model [26]. The BERT model is pre-trained based on typically two strategies on large-scale unlabeled corpora, which are the Masked Language Model and the Next Sentence Prediction tasks. The input representation to the model is a concatenation of WordPiece embeddings [27], positional embeddings, and the segment embedding. Its processing can be outlined as a sequence of Multi-Head Attention, Add & Normalization, and Feed-Forward layers repeated N times (with residual connections [28]). For an input token sequence $x = (x_1, ..., x_n)$, it outputs a sequence of representations as $h = (h_0, h_1, ..., h_n, h_{n+1})$ to capture pertinent contextual information for each token. A special classification embedding ([CLS]), denoted as h_0 is included as the first token and a special token ([SEP]), denoted as h_{n+1} is inserted as the last token. The basic architecture of the BERT model is shown in Figure 1. Several works such as [29], [30], extends the h_0 or the [CLS] token obtained from the pre-trained BERT model for sentence-level classification as

$$P_y = softmax(h_0 W^T + b), \tag{2}$$

where P_y represents class label probabilities and the predicted category can be obtained as

$$y = argmax(P_y) \tag{3}$$

where y represents the final classified category among set of labels.

b) **Proposed BERT-extended***:* We extend the BERT model for the task of tweet act classification. Our model is based on the BERT outputted representations pre-trained over a corpus. To utilize these representations for the given classification task, we fine-tune it over our task-specific dataset. Let $W_B \in \mathbb{R}^{k \times n}$ be the weight matrix obtained from the BERT model for a sequence of tokens, i.e., we ignore the [CLS] and the [SEP] token to obtain representation for each token, t_j. Next, we perform a series of operations with these word representations to obtain an optimal sentence/tweet representation to classify the tweets. Our proposed approach is based on calculating the attention over the word representation obtained from the BERT model. The idea stems from the fact that not all words or tokens in a tweet are equally contributing towards identifying a particular tag. Thus, we eliminate less importance bearing words and put emphasis in determining those word representations that cumulatively aid towards classification of a tweet. The architecture of the proposed model is shown in Figure 2. The process is described as follows :

- The weight matrix W_B is normalized along every column to obtain values in the range of mod 1. Let the normalized weight matrix be $W_N \in \mathbb{R}^{k \times n}$.
- Next, we compute the row-wise average of the normalized weight matrix W_N to obtain a weight vector say $V_a \in \mathbb{R}^{k \times 1}$.
- We then calculate, cosine similarity or the dot product of the obtained vector V_a with the normalized word representations obtained from the BERT model. The notion is to eliminate representations below a particular threshold (average in our case) as the learnt representations do contain noises with respect to a particular tag. Let the obtained vector be $C \in \mathbb{R}^{1 \times n}$.

$$C = (V_a^T . W_N) \tag{4}$$

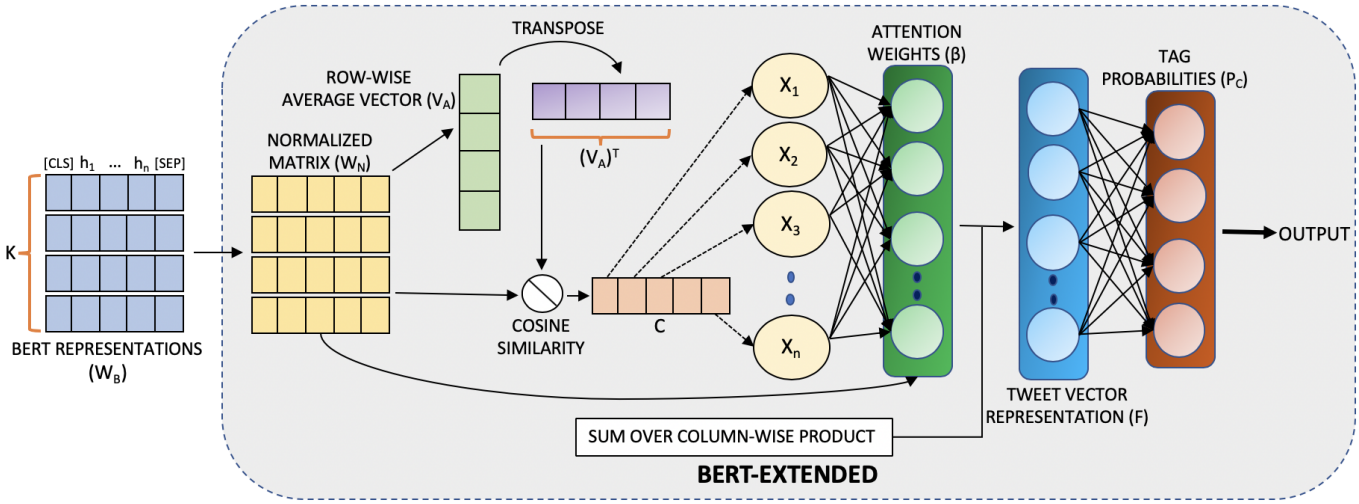

Fig. 2. The architecture of the BERT-extended model

- We obtain the softmax of the weight vector C to squash the range of values between 0 - 1 as tokens which are more contributing towards a specific tag will get higher values and vice-versa. Let the obtained weight vector be $\beta \in \mathbb{R}^{1\times n}$. This operation in a way computes the attention over the sequence of tokens in relation to a tweet act.

$$\beta = softmax(C) \tag{5}$$

- The final tweet vector representation $F \in \mathbb{R}^{1\times n}$ is obtained as

$$F = \sum_{i=1}^{n}(W_N.\beta^T), \tag{6}$$

  i.e., sum over column-wise product of W_N and attention weight vector β.
- The tweet-level category probability, P_c, is obtained as

$$P_c = softmax(F.W_c + b_c), \tag{7}$$

  and the classified tag is obtained by,

$$t = \text{argmax}(P_c), \tag{8}$$

  where $W_c \in \mathbb{R}^{|C|}$ and b_c are the tweet-level classifier weight matrix and biases, respectively, and $|C|$ is the number of tweet acts.

V. Implementation Details

In this section, we first discuss the details of the dataset used followed by the experimental set-up.

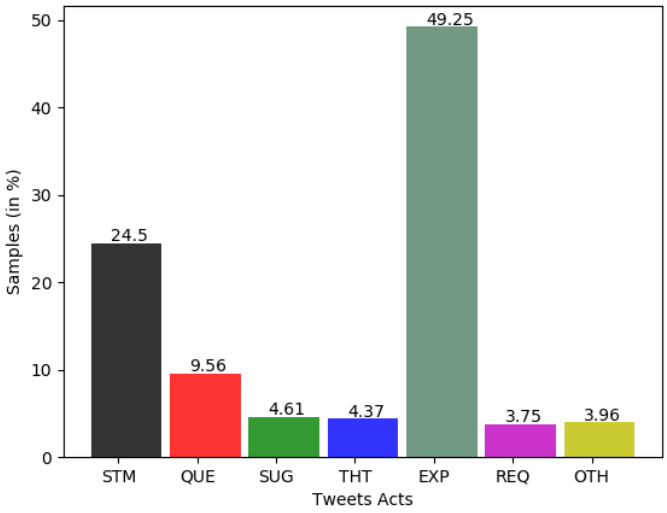

Fig. 3. Distribution of Tweet Acts in the Dataset

a) ***Dataset:*** We conducted experiments over an open-access dataset released by the authors of the paper [12]. In this particular work, authors released tweets of three different types which are *Long-standing*, *Event-oriented* and *Entity-oriented*, and picked up two topics for each of these three types for a total of six topics which are *Ursula K. Le Guin*, *Immigration and Travel Ban*, *J. K. Rowling*, *Gaza Attack*, *GifHistory* and *YesAllWomen*. These tweets were collected over a period of time and was archived in *https://www.docnow.io/catalog/* to be used publicly. Authors utilized *twarc*[1], a command line tool and Python library for archiving Twitter JSON data and gathered numerous tweets corresponding to the tweet IDs for each of the topics mentioned above. Along with it, they also introduced seven different categories of tweet acts

[1]https://github.com/DocNow/twarc

TABLE I
AN EXAMPLE TWEET FOR EACH OF THE TWEET ACTS

Tweet Act	Example
EXPRESSION (EXP)	Last night, I dreamed that the service component for my course was painting Ursula Le Guin's house for her.
STATEMENT (STM)	RT @jk_rowling: A 9-year-old Nigerian girl has written a book about the effects of terrorism on children.
SUGGESTION (SUG)	RT @JoshBoltzIsDead: Instead of being mad at #YesAllWomen because you think it paints men as rapey misogynists you should be mad at men who do these things to women.
THREAT (THT)	The Maiming Fields of #Gaza Even the BBC has shown films of the deliberate shooting of people who were standing harmlessly or running away, including children and journalists. The sniper-fire is mostly not to the head, with most of the wounds to the lower torso and legs.
REQUEST (REQ)	RT @Duh_Mar_Rah: @MatthewACherry #GifHistory please I must know this story.
QUESTION (QUE)	Will Trump rewrite the immigration order?
OTHERS (OTH)	RT @narrynicotine: #yesallwomen

TABLE II
RESULTS OF THE BASELINES, STATE-OF-THE-ART AND THE PROPOSED MODELS IN TERMS OF ACCURACY AND F1-SCORE. † REPRESENTS THAT THE RESULTS ARE STATISTICALLY SIGNIFICANT

Model	Accuracy †	F1-score †
LSTM (Baseline-1)	62.41%	0.61
Bi-LSTM (Baseline-2)	63.75%	0.62
CNN-LSTM (Baseline-3)	64.84%	0.63
CNN-GRU (Baseline-4)	64.51%	0.63
CNN-Bi-LSTM (Baseline-5)	65.02%	0.64
CNN-Bi-GRU (Baseline-6)	64.96%	0.63
BERT (Baseline-7)	73.17%	0.72
Bi-LSTM-extended (Baseline-8)	66.27%	0.66
BERT-extended (Our model)	**75.97%**	**0.74**

influenced from the works of [10], [11] which are "*Statement*", "*Expression*", "*Suggestion*", "*Request*", "*Question*", "*Threat*" and "*Others*". Table I shows an example tweet for each of the tweet acts. The distribution of the tags across the dataset is shown in Figure 3. A total of 7735 tweets annotated with these tweet acts were released. They performed a dataset split of 80% - 20% with the train and test set comprising of 6188 and 1547 tweets, respectively. Thus, we use the same train and test set for our experimentation so as to make a direct comparison with this work. For more details of the dataset and tag-set, refer to paper [12][2].

*b) **Experimental Setup**:* In the subsequent experiments, we use the pre-trained BERT model available from the paper [25]. This model comprises of 12 layers and the size of the hidden state is 768. The multi-head self-attention is composed of 12 heads for a total of 112M parameters. The following hyper-parameters were tuned to obtain the best results reported below :

1) a learning rate of 0.0001 was optimal,
2) a dropout value of 0.1 was ideal for our setting,
3) Adam optimizer was used for all the experiments

For the baseline models, kernel size of 3 with 64 feature maps has been used for the CNN layer whereas 90 hidden units have been used for the corresponding recurrent layer.

[2]The dataset can be downloaded from https://github.com/sahatulika15/Tweet-Act-classification

VI. RESULTS AND ANALYSIS

To analyze the performance of the proposed model, we compare it with several strong baselines and state of the art models in terms of overall accuracy and F1 measure.

*a) **Comparison with the Baselines**:* We compare our BERT-extended approach with the following baselines :

1) **LSTM model (Baseline - 1)** : In this baseline, we only utilize the Long Short Term Memory layer with categorical crossentropy loss to report the results. We utilize the Glove embeddings to represent words;
2) **Bi-LSTM model (Baseline - 2)** : In this baseline, we only utilize the Bi-directional LSTM layer with categorical crossentropy loss to report the results;
3) **Convolutional and LSTM Model (Baseline - 3)** : This baseline has the N-gram convolutional followed by a LSTM layer with categorical crossentropy loss;
4) **Convolutional and GRU Model (Baseline - 4)** : This baseline has a convolutional followed by a GRU layer with categorical crossentropy loss;
5) **Convolutional and Bi-LSTM Model (Baseline - 5)** : This baseline has a convolutional layer followed by a Bi-directional LSTM layer with categorical crossentropy loss;
6) **Convolutional and Bi-GRU Model (Baseline - 6)** : This baseline has a convolutional layer followed by a Bi-directional GRU layer with categorical crossentropy loss;
7) **Original BERT model (Baseline - 7)** : Authors of the BERT paper recommend using the [CLS] token, i.e., the final hidden state h_0 for classification task as it represents a fixed dimensional pooled representation of the sequence/tweet. Thus, we simply use this pre-trained model fine-tuned over this dataset with a layer of hidden units without any other extension to report results for this baseline;
8) **Bi-LSTM-extended (Baseline - 8)** : We use a single Bidirectional LSTM (Bi-LSTM) in place of BERT with our extended version to curate this baseline. This is done as besides providing the advantage, the BERT models are extremely computation intensive because of the presence of multi-layered Transformer encoders. This

TABLE III
SAMPLE UTTERANCES WITH ITS PREDICTED LABEL FOR THE PROPOSED BERT-EXTENDED AND BEST TWO PERFORMING (BASELINE-7 AND BASELINE-8) MODELS

Tweet	True Label	BERT-extended Predicted Label	Baseline-7 Predicted Label	Baseline-8 Predicted Label
RT <user>: Israel has turned Gaza's sea into a battlefield. IOF routinely fires on boats, injuring, killing & arresting Palestinian	THT	THT	EXP	STM
RT <user>: On Revolution Day, Iranians turn out in huge numbers to defy threats & insults by US govt ; praise American people for rejecting travel ban	STM	STM	STM	EXP
RT <user>: <user>This needs to be a full on Netflix documentary show!! #GifHistory	SUG	EXP	EXP	EXP
RT <user>: # <hashtag>When my lover abused me I adjusted my behaviour to avoid provoking him (no, it didn't work).	EXP	EXP	STM	STM

TABLE IV
F1-SCORES OF TWEET ACTS FOR THE TOP TWO BASELINES AND THE PROPOSED APPROACH

Model	Class						
	REQ	QUE	SUG	STM	THT	EXP	OTH
BERT (Baseline-7)	0.57	0.67	0.50	0.65	0.51	0.82	0.60
Bi-LSTM-extended (Baseline-8)	0.61	0.58	0.40	0.54	0.47	0.76	0.49
BERT-extended (Our model)	0.71	0.77	0.65	0.74	0.60	0.84	0.63

baseline also highlights the importance and advantage of the proposed approach.

The results of all these baselines along with the proposed approach in terms of accuracy and F1 measures are shown in Table II. As is evident, the proposed BERT-extended model produced better results compared to all other baselines. This gain is rather intuitive as our model computes attention weights over the token representations obtained from the BERT model for identification of a particular tweet act. Our proposed model showed an improvement of almost 3% and 9% in terms of accuracy compared to the top two baselines 7 & 8, respectively. The contribution of the extended part of the proposed approach can be realized by comparing the baselines 2 & 8 as the extended part shows an improvement of almost 3% over the baseline 2. The F1-scores of individual tags for the top two baselines and the proposed model are shown in Table IV. Also, the confusion matrix of the proposed approach during testing is shown in Table V for a detailed analysis on the misinterpretation of tags. We also present some sample predictions during testing for the top two baselines and the proposed model in Table III. All the reported results are statistically significant as we have employed the Welch's t-test [31] at 5% significance level.

*b) **Comparison with the State-of-the-Art**:* We have also performed a comparative study of the state-of-the-art models [10], [11], [12] with the results reported in the paper [12]. This is because we solely use this dataset to perform all the experiments in our work as we were unaware of any other open-access and sizable Twitter data annotated with its corresponding speech act at the time of writing. Table VI shows the results of all the state of the art approaches. As is evident from the table, our proposed BERT-extended model outperformed all other state-of-the-art approaches by significant margin.

TABLE V
CONFUSION MATRIX FOR THE PROPOSED BERT-EXTENDED

	STM	EXP	QUE	OTH	SUG	REQ	THT
STM	**174**	98	7	2	6	4	6
EXP	45	**721**	22	13	5	0	2
QUE	6	15	**119**	0	0	0	2
OTH	15	18	0	**56**	2	1	1
SUG	5	28	0	5	**35**	1	1
REQ	1	6	9	5	2	**24**	0
THT	6	4	0	1	1	0	**46**

TABLE VI
RESULTS OF THE STATE-OF-THE-ART AND THE PROPOSED MODELS IN TERMS OF ACCURACY AND F1-SCORE VALUES

Model	Accuracy	F1-score
SVM (Zhang et al., 2011)	66.45%	0.65
LR (Vosoughi et al., 2016)	68.70%	0.67
CNN-SVM (Saha et al., 2019)	73.75%	0.71
BERT-extended (Our model)	**75.97%**	**0.74**

*c) **Error Analysis**:* A thorough analysis was conducted to understand where our proposed model faltered. The possible reasons are as follows :

- **Imbalanced dataset** : One of the fundamental reasons being the skewed dataset, i.e., contribution of most the tweet acts in the dataset is very less with much of the representation from "EXP" and "STM" tags. Thus, its F1-score is relatively improved than rest of the tags as shown in Table IV;
- **Subset tags** : Tags such as "THT" suffer massively with respect to F1-score as majority of the tweets of

this category are subset of "STM" tags with the former being a fine-grained tag of the latter. For example, "*RT @AviMayer: A new form of terror: farmland belonging to Kibbutz Nir Am, near Israel's border with Gaza, goes up in flames*" has been misinterpreted as "STM" instead of "THT". Also, the number of instances of "THT" tag is relatively less in the released dataset. Similar confusions were noticed between classes where tweet such as "*@jk_rowling would be amazing if @MerrynPippin could get her college fundme link RT'd by you http://t.co/AVSC7NnliV http://t.co/FRqIeFHP3m*" is mis-classified as "EXP" rather than "SUG";

- **Miscellaneous** : Tag like "OTH" is also affected owing to the absence of any predefined structure of the tweet belonging to this category.

The performance gain of the proposed BERT-extended can be attributed to the following : (i) *The presence of BERT embeddings which provides deep bi-directional contextual representations by multiple attention heads attending to different sections of input in parallel;* **(ii)** *The proposed approach is based on calculating the attention over the word representation obtained from the BERT model. It leverages from the elimination of less importance bearing words and learns word representations that cumulatively aid towards classification of a particular tag;* **(iii)** *The proposed BERT-extended seeks advantage from the joint optimization of these two significant layers to learn features pertaining to speech acts and Twitter.*

VII. Conclusions and Future Work

In this paper, we present a novel model for the identification of speech acts in Twitter on top of BERT. We treat this problem as a multi-class classification problem and introduce *BERT-extended* classifier for the task. Our proposed model is based on calculating the attention weights over the token representations of a sequence obtained from the pre-trained BERT model. We compare our proposed approach with several strong baselines and state-of-the-art approaches. Our model attained a benchmark overall accuracy and F1 measure of 75.97% and 0.74 respectively.

In future, attempts can be made to boost the system's efficiency in classifying the tweets with more precision and accuracy. An even fine-grained taxonomy can be curated to capture the communicative intention of the user in detail. Different other aspects of communication such as emotion and sentiment analysis can be integrated in the model in order to leverage from the state of mind of the tweeter.

Acknowledgement

Sriparna Saha would like to acknowledge the support of SERB WOMEN IN EXCELLENCE AWARD 2018 for conducting this research.

References

[1] G. Blogger, "Twitter users statistics 2016 infographics," Dec 2016. [Online]. Available: https://www.globalmediainsight.com/blog/twitter-users-statistics/

[2] A. C. Pandey, D. S. Rajpoot, and M. Saraswat, "Twitter sentiment analysis using hybrid cuckoo search method," *Information Processing & Management*, vol. 53, no. 4, pp. 764–779, 2017.

[3] F. Laylavi, A. Rajabifard, and M. Kalantari, "Event relatedness assessment of twitter messages for emergency response," *Information Processing & Management*, vol. 53, no. 1, pp. 266–280, 2017.

[4] S. M. Mohammad, X. Zhu, S. Kiritchenko, and J. Martin, "Sentiment, emotion, purpose, and style in electoral tweets," *Information Processing & Management*, vol. 51, no. 4, pp. 480–499, 2015.

[5] J. L. Austin, *How to do things with words*. Oxford university press, 1975, vol. 88.

[6] J. R. Searle and J. R. Searle, *Speech acts: An essay in the philosophy of language*. Cambridge university press, 1969, vol. 626.

[7] J. R. Searle, "A taxonomy of illocutionary acts," 1975.

[8] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[9] H. Khanpour, N. Guntakandla, and R. Nielsen, "Dialogue act classification in domain-independent conversations using a deep recurrent neural network," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 2012–2021.

[10] R. Zhang, D. Gao, and W. Li, "What are tweeters doing: Recognizing speech acts in twitter." *Analyzing Microtext*, vol. 11, no. 05, 2011.

[11] S. Vosoughi and D. Roy, "Tweet acts: A speech act classifier for twitter." in *ICWSM*, 2016, pp. 711–715.

[12] T. Saha, S. Saha, and P. Bhattacharyya, "Tweet act classification : A deep learning based classifier for recognizing speech acts in twitter," 07 2019, pp. 1–8.

[13] S. Vosoughi, "Automatic detection and verification of rumors on twitter," Ph.D. dissertation, Massachusetts Institute of Technology, 2015.

[14] C. Cerisara, S. Jafaritazehjani, A. Oluokun, and H. T. Le, "Multi-task dialog act and sentiment recognition on mastodon," in *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, 2018, pp. 745–754.

[15] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1. IEEE, 1992, pp. 517–520.

[16] S. Grau, E. Sanchis, M. J. Castro, and D. Vilar, "Dialogue act classification using a bayesian approach," in *9th Conference Speech and Computer*, 2004.

[17] H. Kumar, A. Agarwal, R. Dasgupta, and S. Joshi, "Dialogue act sequence labeling using hierarchical encoder with crf," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[18] S. Sitter and A. Stein, "Modeling the illocutionary aspects of information-seeking dialogues," *Information Processing & Management*, vol. 28, no. 2, pp. 165–180, 1992.

[19] J. M. Budd and D. Raber, "Discourse analysis: Method and application in the study of information," *Information Processing & Management*, vol. 32, no. 2, pp. 217–226, 1996.

[20] D. Verbree, R. Rienks, and D. Heylen, "Dialogue-act tagging using smart feature selection; results on multiple corpora," in *Spoken Language Technology Workshop, 2006. IEEE*. IEEE, 2006, pp. 70–73.

[21] N. Kalchbrenner and P. Blunsom, "Recurrent convolutional neural networks for discourse compositionality," in *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*. Association for Computational Linguistics, 2013, pp. 119–126. [Online]. Available: http://aclweb.org/anthology/W13-3214

[22] Y. Liu, K. Han, Z. Tan, and Y. Lei, "Using context information for dialog act classification in dnn framework," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2170–2178.

[23] T. Saha, S. Srivastava, M. Firdaus, S. Saha, A. Ekbal, and P. Bhattacharyya, "Exploring machine learning and deep learning frameworks for task-oriented dialogue act classification," 07 2019, pp. 1–8.

[24] M. Jeong, C.-Y. Lin, and G. G. Lee, "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*. Association for Computational Linguistics, 2009, pp. 1250–1259.

[25] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "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)*, 2019, pp. 4171–4186.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "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*, 2017, pp. 5998–6008.

[27] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90

[29] Q. Chen, Z. Zhuo, and W. Wang, "BERT for joint intent classification and slot filling," *CoRR*, vol. abs/1902.10909, 2019.

[30] G. Castellucci, V. Bellomaria, A. Favalli, and R. Romagnoli, "Multi-lingual intent detection and slot filling in a joint bert-based model," *CoRR*, vol. abs/1907.02884, 2019.

[31] B. L. Welch, "The generalization ofstudent's' problem when several different population variances are involved," *Biometrika*, vol. 34, no. 1/2, pp. 28–35, 1947.