BERT-Caps: A Transformer-Based Capsule Network for Tweet Act Classification

Tulika Saha, *Graduate Student Member, IEEE*, Srivatsa Ramesh Jayashree, Sriparna Saha, *Senior Member, IEEE*, and Pushpak Bhattacharyya

***Abstract*— Identification of speech acts provides essential cues in understanding the pragmatics of a user utterance. It typically helps in comprehending the communicative intention of a speaker. This holds true for conversations or discussions on any fora, including social media platforms, such as Twitter. This article presents a novel tweet act classifier (speech act for Twitter) for assessing the content and intent of tweets, thereby exploring the valuable communication among the tweeters. With the recent success of Bidirectional Encoder Representations from Transformers (BERT), a newly introduced language representation model that provides pretrained deep bidirectional representations of vast unlabeled data, we introduce BERT-Caps that is built on top of BERT. The proposed model tends to learn traits and attributes by leveraging from the joint optimization of features from the BERT and capsule layer to develop a robust classifier for the task. Some Twitter-specific symbols are also included in the model to observe its influence and importance. The proposed model attained a benchmark accuracy of 77.52% and outperformed several strong baselines and state-of-the-art approaches.**

***Index Terms*— Bidirectional Encoder Representations from Transformers (BERT), capsule networks, speech acts, Twitter.**

I. Introduction

IN RECENT times, there has been an upsurge of various social media and microblogging platforms, e.g., Twitter[1], MySpace,[2] Facebook,[3] and Tumblr.[4] This is specifically because these platforms provide an efficient means of communication among Internet users all across the globe. These social networks that allow the exchange of millions of messages and information each day are diverse and valuable resources for exploring and researching the interests of people and for analyzing the contents created by users. Among an array of several such social media platforms, Twitter forms one of the most dominant microblogging service interwoven by interesting people, alluring events, topics, and so on. It provides users or tweeters with an effortless medium to share opinions and views about various trending topics or discuss topics in general, provide suggestions, ask for solutions to their queries, express happenings about their life, share facts, information and news update, and so on. Twitter is being used to such a large extent that as of 2018, there were 330 million monthly active users and over 500 million tweets (status messages) were generated each day [1]. Twitter's social nature with its massive volume has collectively turned it into a frugal data source for observing and studying actions, behavior, and user's personality. Hence, comprehending Twitter content, i.e., what people are tweeting about, forms the core of this article.

There has been a fair amount of research to examine the linguistic element of tweets, such as those of [2]–[4], but there has been very little work on understanding the pragmatics of tweets. Pragmatics, in general, focus on capturing the intended meaning of an utterance and look beyond the literal meaning, thereby emphasizing the context and intention of the speaker. Thus, it captures and identifies the behavior and communicative intention of a user utterance. Pragmatics lies at the core of any linguistic system, providing a base for all language communications and contacts. It is a crucial feature to the understanding of language, what it is used for, and the reaction that comes after. 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, among various other features, a precise and definite taxonomy of communicative acts known as speech acts [7]. Thus, the current article primarily addresses the task of recognizing such speech acts in tweets for studying pragmatics and, thereby, understanding the tweet contents automatically.

The recognition of speech acts in an automated framework has a compelling influence on Twitter as well as tweeters. For the Twitter platform itself, identification of tweet acts helps in figuring out what a specific topic or theme is composed of in terms of speech acts and whether there is an inconsistency in a particular topic. For example, topics, such as news or message broadcast, typically encompass statements, assertions, and facts. Thus, a visible change or alteration in general inferences the trends in speech acts, say some recommendation or threat and so on, can be inferred to be a case of topic variation or spamming. Thus, it fundamentally assists in monitoring and supervising social media content. It benefits the readers and followers of the platform, by allowing the users to follow or search for a particular topic by utilizing the most convenient and preferred speech act based on their needs. For example, suggestions and advices can be useful under topics, such as traveling, health-care, or cooking, and so on or statements, and

Manuscript received January 18, 2020; revised May 6, 2020 and June 19, 2020; accepted July 20, 2020. Date of publication September 16, 2020; date of current version November 10, 2020. *(Corresponding author: Tulika Saha.)*

The authors are with the Department of Computer Science and Engineering, IIT Patna, Patna 801106, India (e-mail: sahatulika15@gmail.com).

Digital Object Identifier 10.1109/TCSS.2020.3014128

[1]http://twitter.com
[2]https://myspace.com
[3]http://facebook.com
[4]http://tumblr.com

expressions under affairs related to politics or economics, and so on. Thus, by simply limiting the search space, it primarily guides and helps the users to become diligent readers of the platform among millions of tweets. Also, recognizing speech acts provides an enhanced perception of the behavior, attitude, and general mindset of the tweeters. For example, the users are baffled if they are asking questions about a topic or are pleased or maybe dissatisfied if they are expressing their views and so on.

Extensive amounts of works have already been done for identifying speech acts in the context of dialogs known as dialog act classification (DAC) in computational linguistics, which includes notable works, such as in [8] and [9]. However, because of the limited tweet length (now 280 characters, initially, it was 140), noisy, unusual, and eccentric nature of tweets makes it unsuitable for applying typical data mining and information retrieval approaches. Recently, the introduction of Bidirectional Encoder Representations from Transformers (BERT) [10], a language representation model, has established state-of-the-art results for a vast spectrum of natural language processing tasks [11], [12]. However, it is not fully known how BERT works in the context of tweets because it came out very recently. Primarily, the usage of BERT allows the training of the language model bidirectionally that has a deeper sense of language context and flow compared with single-direction language models where a text sequence is viewed as either from left to right or combined left-to-right and right-to-left training. The bidirectional or rather nondirectional characteristic allows the model to learn the context of a word based on all of its surroundings (left and right of the word). In this article, we present a BERT-based model named BERT-Caps for the identification of tweet acts. Our proposed approach tends to learn traits and attributes by leveraging from the joint optimization of features from the BERT and capsule layer [13] to develop a robust classifier for the task. Some Twitter-specific symbols are also included in the model to observe its influence and importance. Empirically, we show that our proposed approach outperforms several strong baselines and state-of-the-art models significantly.

The key contributions of this article are the following.

1) This article presents a novel speech act classifier for tweets named BERT-Caps that is built on top of BERT.
2) The model leverages from the joint optimization of features of the BERT embeddings and capsule layers to learn cumulative features pertaining to speech acts and Twitter.
3) The proposed model attained state-of-the-art results for the task of tweet act classification.

The remainder of this article is organized as follows. Section II presents a short formal description of the problem statement. Section III presents a brief description of the related works and the motivation behind solving this problem. The proposed methodology has been discussed in Section IV. Section V presents the implementation details. The experimental results and the corresponding analysis are given in Section VI. Finally, the concluding remarks and the directions for future work are discussed in Section VII.

II. Problem Statement

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

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

where F is the developed tweet act classifier. We acknowledge the fact that, in real-life situations, this presumption may not always hold and that one tweet may demonstrate multiple speech acts. However, because of the limited and short length of tweets, multiple speech acts tweets 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

This particular section discusses work done so far on tweet act classification followed by the motivation behind solving this problem.

A. Background

An extensive literature survey was carried out to explore various works done on tweet act classification. Zhang *et al.* [14] 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: "Comment," "Statement," "Suggestion," "Question," and "Miscellaneous." Their model was based on manually extracted handcrafted features from the tweets, such as cue words and phrases that included maximum occurrences of unigrams, bigrams, and trigrams, some noncue words that included opinion words, emoticons, abbreviations, and so on, followed by some character-based features, such as Twitter-specific characters and punctuations. Their proposed method achieved an F1-score of 0.7. Later, Vosoughi and Roy [15] proposed logistic regression (LR)-based speech act classifier for tweets. Their work also employed manually extracted features, such as semantic level features that included speech act verbs, N-grams (typically unigrams, bigrams, and trigrams), and emoticons, followed by syntactic level features, such as abbreviations, Twitter-specific symbols, and so on along with dependence subtrees 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: "Expression," "Assertion," "Request," "Recommendation," "Question," and "Miscellaneous." They reported an F1-score of 0.7 based on their approach.

Saha *et al.* [16] presented first-ever and the only deep learning (DL)-based tweet act classifier. Their approach was a convolutional neural network (CNN)-based model with SVM loss function to address the multiclass classification problem more efficiently. They also incorporated a few handcrafted 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 are discussed in Section V since this data set has been utilized to demonstrate this work. Vosoughi [17] highlighted the importance of identification of tweet acts and established it to be one of the elementary steps for the detection of rumors on Twitter. Cerisara *et al.* [18] proposed a multitask 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 platforms, whereas our work is purely based on learning a classifier for tweet act classification without any transfer learning or multitask-based scenarios.

Apart from these, identification of speech acts has been studied extensively for dialog conversations starting from the early 2000's. Stolcke *et al.* [8] presented varieties of approaches, such as hidden Markov models, neural networks, and decision trees to identify dialog acts (DAs) on a benchmark dialog data known as the Switchboard (SWBD) [19] data set. Grau *et al.* [20] presented a naive Bayes-based DA classifier. Later, with the advancement of DL, several neural network-based approaches were widely proposed. Khanpour *et al.* [9] presented a stacked long short-term memory (LSTM)-based approach on the SWBD data set. Kumar *et al.* [21] presented a hierarchical encoder-based approach to model the task of DAC. Verbree *et al.* [22] presented a J48 classifier-based model with several handcrafted features, such as n-grams and part of speech n-grams, to predict DAs. Kalchbrenner and Blunsom [23] proposed a recurrent convolutional network-based model to capture sentence as well as discourse-level compositionality to model context across a sentence as well across a dialog. Liu *et al.* [24] developed a hierarchical convolutional and recurrent network to capture dialog history in order to predict DAs, thus stressing on the role of context across a dialog. Saha *et al.* [25] presented several machine learning (ML)-based models involving clustering and word embeddings along with conditional random fields (CRFs), DL-based models with CNNs, and Bi-LSTMs concatenated with several clustering-based features for dialogs. Sankar and Ravi [26] presented a Seq2seq-based approach to generate texts conditioned on several attributes, such as DA. For this, they trained a DA classifier on the Reddit data set to assist their model to generate responses accordingly. Jeong *et al.* [27] presented a semisupervised approach to identify speech acts in emails and different forums. These works, however, use data sets that comprise of face-to-face or telephone data that cannot directly aid in advancing work on endless data in electronic mode, such as microblogging networks and instant-messaging.

B. Motivation

The followings are some observations that were made after an exhaustive literature analysis, and these factors motivated us to study the task of tweet act classification.

1) Few works that are done for tweet act classification employ a considerable amount of manual feature engineering to formulate fine-grained handcrafted and Twitter-specific features [14], [15].
2) The existing approaches make the entire process extremely tedious, time-consuming, and inept [14], [15].
3) An even higher amount of robustness is required for the evaluation of tweets with regards to accuracy and precision.
4) Tweets are replete with random coinages, spelling mistakes, collective usage of letters and symbols, and so on. Thus, existing approaches for DAC cannot be directly applied to the microblogging 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 [9], [21], [24], and so on.
5) 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.

Motivated by these factors, this article aims to propose a DL-based model to reduce human intervention and the need for fine-grained feature engineering, thereby making the entire process less time-consuming and more robust. The proposed model outperforms several strong baselines and state-of-the-art approaches to predict tweet acts with increased precision.

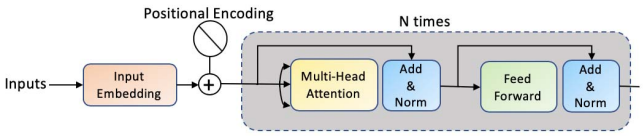

Fig. 1. Basic architecture of the transformer encoder model.

IV. Proposed Methodology

In this section, we first introduce the BERT model and capsule networks concisely followed by the developed BERT-Caps model for our task.

A. BERT

BERT [10] is a multilayered attention-aided bidirectional transformer encoder model based on the original transformer model [28]. It is pretrained on BooksCorpus (800M words) [29] and English Wikipedia (2500M words). The input representation to the model is a concatenation of WordPiece embeddings [30], positional embeddings, and the segment embedding. Its processing can be outlined as a sequence of multihead attention, add, & Normalization and feedforward layers repeated N times (with residual connections [31]). For an input token sequence $x = (x_1, \ldots, x_l)$, it outputs a sequence of representations as $h = (h_0, h_1, \ldots, h_l, h_{l+1})$ 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_{l+1} is inserted as the last token. The basic architecture of the BERT model is shown in Fig. 1.

B. Capsule Networks

In neural network-based methods, spatial patterns acquired at lower levels contribute to the representation of higher level

concepts. For example, CNN constructs convolutionary feature detectors to extract local patterns from a vector sequence and uses max-pooling to select the most prominent ones. Being a spatially sensitive model, CNN pays a price for the inefficiency of replicating feature detectors on a grid. On the other hand, methods that are spatially insensitive are perfectly efficient at the inference time regardless of any order of words or local patterns. However, they are unavoidably more restricted to encode rich structures presented in a sequence. Improving the efficiency to encode spatial patterns while keeping the flexibility of their representation capability is, therefore, a bottleneck. Capsule networks are, thus, known to address this issue. They introduce an iterative routing process to decide the credit attribution between nodes from lower to higher layers. As an outcome, their model could encode the intrinsic spatial relationship between a part and a whole constituting viewpoint invariant knowledge that automatically generalizes to novel viewpoints.

C. Proposed BERT-Caps

We leverage the representations from the BERT model for the task of tweet act classification. Our model is based on the BERT outputted representations pretrained over a corpus specified earlier. To utilize these representations for the given classification task, we fine-tune it over our task-specific data set. Let $W_B \in \mathbb{R}^{k \times l}$ 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 of a given tweet. Consecutively, the obtained representation is passed through a series of operations and, in turn, hierarchical layers to obtain an optimal sentence/tweet representation to classify the tweets that are described as follows.

1) **N-gram Convolutional Layer:** The obtained representation from the BERT model with a sequence length of l and each token representation of k-dimension is first passed through a convolution layer to extract and learn abstract features from N-grams. This layer outputs feature maps by sheafing the inner layers of convolution operation followed by pooling

$$f_i = W_B * k_i + b \tag{2}$$
$$\hat{f}_i = p(f_i) \tag{3}$$

where k_i is the kernel with bias b that outputs feature map f_i by convolution and $p(.)$ is the pooling operation. Next, we congregate t such feature maps to form a t-channel layer as

$$F = [\hat{f}_1, \hat{f}_2, \ldots.., \hat{f}_t]. \tag{4}$$

Thus, the n-gram convolutional layer captures an abstract representation of the phrases, i.e., n-grams, reflecting its semantic meaning at different positions that finally span over the entire sentence.

2) **Bidirectional Recurrent Layer:** The obtained feature maps or the t-channel layer F are then passed through a bidirectional LSTM to sequentially encode these feature maps into hidden states and learn semantic dependence-based features as

$$\overrightarrow{h_t} = LSTM_{fd}(f_t, \overrightarrow{h}_{t-1}) \tag{5}$$
$$\overleftarrow{h_t} = LSTM_{bd}(f_t, \overleftarrow{h}_{t+1}). \tag{6}$$

For each of these feature maps, its corresponding forward and backward hidden states $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$, respectively, from the forward $LSTM_{fd}$ and the backward $LSTM_{bd}$ are concatenated to obtain a single hidden state h_t. The complete hidden state matrix is obtained as

$$H = [h_1, h_2, \ldots., h_t] \tag{7}$$

where $H \in \mathbb{R}^{t \times 2d}$ and d represents the number of hidden units in each LSTM.
This layer, thus, captures the context within particular phrases to learn long term dependencies in the tweet.

3) **Primary Capsule Layer:** The generated semantic and context-dependent features are passed through the primary capsule layer, thus fragmenting the instantiated parts collectively with another convolution operation. Primary capsules typically utilize vectors as opposed to scales in order to retain the instantiated parameters belonging to each feature. This is important because, in addition to depicting the intensity of the activation, it also maintains some details of the instantiated parts in the input of this layer. Thus, in this particular manner, the capsule can be considered as a short depiction of the instantiated parts that are captured by the kernel of the convolution. Thus, sliding over the hidden states H, each kernel k_i outputs a sequence of capsule, $c_i \in \mathbb{R}^d$ of dimension d. These outputted capsules consist of a channel C_i belonging to the primary capsule layer

$$C_i = s(k_i * H + b) \tag{8}$$

where s represents the nonlinear squash function and b is the bias weight parameter of the capsule. Thus, now, all the I such channels can be compiled as

$$C = [C_1, C_2, \ldots, C_I]. \tag{9}$$

Analogous to spatial orientation in images, this layer basically captures local ordering of words in tweets as well as the corresponding semantic representations. As explained earlier, this layer primarily takes a step toward countering the flaws of conventional CNN by replacing its scalar-output features with the vector-output capsules in order to retain the instantiated parameters.

4) **Connecting the Capsule Layers via Dynamic Routing:** Normally, the capsule network creates capsules in the next subsequent layer by following the "routing-by-agreement" algorithm. This algorithm replaces the pooling operation of the conventional convolution layer that typically eliminates the location information. However, this kind of information is extremely essential as it enhances the robustness of the network and aids the clustering of features for prediction. Thus, in between

two consecutive layers say l and $l+1$, a predictor vector $\hat{v}_{b|a}$ is estimated from the capsule v_a as

$$\hat{v}_{b|a} = W_{ab} v_a \tag{10}$$

where v_a is the capsule in layer l. In the layer $l+1$, a capsule z_b is computed as

$$z_b = \sum_a u_{ab} v_{b|a} \tag{11}$$

where u_{ab} represents coupling coefficients determined by the dynamic routing algorithm. This algorithm determines the intensity of connection or link between capsules, i.e., from capsule a of the lth layer to capsule b of the $l+1$th layer that justifies the coupling coefficient u_{ab}. The primary value of coupling coefficient q_{ab} is revised with routing by agreement p_{ab}, which is obtained as

$$p_{ab} = v_{\hat{b}|a} \cdot m_b \tag{12}$$

where m is the length of a capsule. The length of the capsule represents the probability that input sample has, the object capsule describes which is the activation of the capsule. Thus, the length of the capsule is restricted in the range of [0, 1] with a nonlinear squashing function

$$m_b = \frac{\|z_b\|^2}{1 + \|z_b\|} \frac{z_b}{\|z_b\|^2}. \tag{13}$$

Thus, the value of agreement p_{ab} is added to the value to compute the value of the capsules in the next layer as

$$q_{ab} \leftarrow = q_{ab} + p_{ab}. \tag{14}$$

This whole process from (11) $\rightarrow$ (13) $\rightarrow$ (12) $\rightarrow$ (14) is run iteratively to optimize the value of the coupling coefficients and the capsules in the subsequent layers. With the help of this algorithm, it decides the importance and agreement of words for a particular prediction task. Primarily, it learns which words in the tweet are important or overlooks those which are unrelated for a given tag. For example, a word in the tweet, such as "please," is more likely to be crucial and informational for a tag, such as "REQ." However, it is less important if it appears in relation to a tag, such as "STM." Such agreement of words in a tweet is learned via the routing algorithm from the primary capsule to class capsule layer.

5) **Class Capsule Layer With Loss:** This final capsule layer consists of Y class capsules where each one corresponds to a class label or category. The length of instantiated parameters in each capsule denotes the probability of the input sample belonging to this class label, whereas the direction of each set of instantiated parameters maintains the traits and aspects of the feature attributes that can be thought of as an encoded vector for that input sample. In order to enhance the difference between the length of different class capsules and to aid the process of generalization, we use a separate margin loss [32] as

$$L_b = R_b \max(0, e^+ - \|c_b\|^2) + \lambda(1 - R_b) \max(0, e^- - \|c_b\|^2)) \tag{15}$$

where c_b represents the capsule for category b. e^+ and e^- denote the top and bottom margins, respectively, which forces the length to be between these two margins. λ represents the weight of the classes that are not present. It helps to scale down the weight of the absent classes, thus preventing the length of the capsules to reduce too much.

6) **Additional Features:** To this, we also incorporate some shallow syntactic as well as semantic features that are included as binary features in the primary capsule layer through a fully connected (dense) layer. The used features are as follows.
   a) *Twitter-Specific:* Twitter-specific symbols, such as @, RT, and #, provide valuable information for specific tweet acts, and their explicit positions in the tweet are beneficial indicators of tweet acts. The presence of these symbols in the beginning of the tweet is marked as features.
   b) *Abbreviations and Punctuations:* For conversation on social media platform, users often use abbreviations for relevant phrases. For example, *ttyl* for "talk to you later," *bff* for "best friend forever," and so on. These abbreviations that are indicative of informal speech are marked as features; 2000 such abbreviations were accumulated from an online dictionary.[5],[6] Along with it, the presence of punctuations was also used as features as they also contributes toward distinction of certain tweet acts, e.g., ? for "QUE" and ! for "EXP."
   c) *Emoticons:* Emoticons are being used ubiquitously in variety of online fora and indicate presence of emotions. Thus, 1174 emoticons were gathered using a text-based emoticon,[7] and their presence were marked as features.
   d) *Opinion and Vulgar Words:* 5203 opinion words that include strong positive and negative words, e.g., *hate* and *passion*, were obtained using the "Harvard General Inquirer" lexicon [33] as they indicate presence of certain cues that aids the classification of tweet acts. Similarly, 349 vulgar words were also collected from an online portal[8] and were used as features.

The architectural representation of the proposed BERT-Caps model is shown in Fig. 2.

V. Implementation Details

In this section, we first discuss the details of the data set used followed by the experimental setup.

[5]https://www.netlingo.com/category/acronyms.php
[6]https://pc.net/emoticons/
[7]https://pc.net/emoticons/
[8]https://www.noswearing.com/dictionary

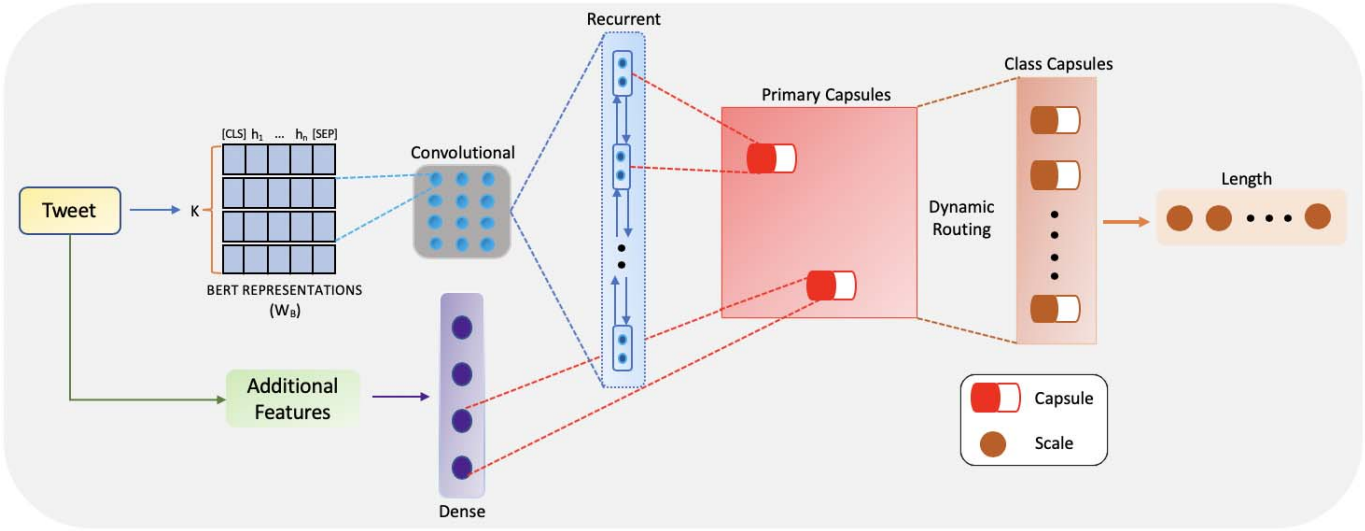

Fig. 2. Architectural diagram of the proposed BERT-Caps model.

A. Data Set

All the experiments were conducted over an open-access data set released by Saha *et al.* [16]. In this particular work, the authors released tweets of three different types, *Long-standing*, *Event-oriented*, and *Entity-oriented*, and picked up two topics for each of these three types, thus accumulating six topics in total: *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 were archived in https://www.docnow.io/catalog/ to be used publicly. The authors utilized *twarc*,[9] 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 abovementioned topics. Along with it, they also introduced seven different categories of tweet acts influenced from the works of [14], [15]: "Statement," "Expression," "Suggestion," "Request," "Question,", "Threat," and "Others." Table I shows the number of tweets for each of these types present in the data set. Table II shows an example tweet for each of the tweet acts. A total of 7735 tweets annotated with these tweet acts were released. They performed a data set 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 data set and tag-set, refer to the article [16].

B. Hyperparameters

In the subsequent experiments, we use the pretrained BERT model available from the article [10]. This model comprises of 12 layers, and the size of the hidden state is 768. The multihead self-attention is composed of 12 heads for a total of 112M parameters. The number of units used in the convolution layer is 64. The size of the kernel used is 5. The number of units used in the recurrent layer is 90. The following hyperparameters were tuned to obtain the best results reported as follows.

1) Learning rate of 0.0001 was optimal.
2) Dropout value of 0.1 was ideal for our setting.
3) Capsule length of 16 was found to be optimum.
4) Dynamic routing algorithm for five iterations gave optimum results.
5) Adam optimizer was used for all the experiments.

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

TABLE I
NUMBER OF TWEETS FOR EACH TYPE AND ITS TOPIC

Type	Topic	# Tweets
Event	Immigration and Travel Ban	1322
	Gaza Attack	1580
Long-Standing	#YesAllWomen	1570
	#GifHistory	550
Entity	J. K. Rowling	1292
	Ursula K. Le Guin	1421

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. An extensive ablation study was conducted to highlight the contribution of the proposed model in several aspects.

A. Comparison With the Baselines

We compare our BERT-Caps approach with the following baselines.

1) *Original BERT Model (Baseline-1):* The authors of the BERT article recommends using the [CLS] token, i.e., the final hidden state h_0 for classification tasks as it represents a fixed dimensional pooled representation of the sequence/tweet. Thus, we simply use this pretrained model without any extension to report results for the first baseline.

TABLE II
EXAMPLE TWEET AND DEFINITION FOR EACH OF THE TWEET ACTS

Tweet Act	Definition	Example
EXPRESSION (EXP)	for any kind of expression of feeling or thought	@BookwormBlues Wow Best writers out there are female and it seems in this world they don't exist Ursula Le Guin NK Jemisin Rachel Aaron Andre Norton Connie Willis JK Rowling Barbara Hambly I can go on and on.....
STATEMENT (STM)	facts, any kind of statement or specifically asserting something	RT @AmyMek: Meanwhile in Florida.... Muslim IMAM at a Florida Mosque says killing gays is the compassionate thing to do #Stonewall #MuslimBan
SUGGESTION (SUG)	giving any kind of suggestion and recommendation	RT @NickNipclose: The MRA movement much like 911 twoofers and tea baggers needs to be considered a domestic terrorist threat. #YesAllWomen
THREAT (THT)	for any kind of social threat, uncertainity, warning	RT @aj_iraqi: According to the latest Peace Index, 83% of Jewish Israelis believe the IDF's open-fire policy at the border with Gaza is justified
REQUEST (REQ)	for any kind of request or appeal made	RT @ChickSpinster: @MatthewACherry #GifHistory I'd be most grateful if you could illuminate my favorite gif.
QUESTION (QUE)	for any question asked	@jk_rowling When writing your books, did you ever get severe writers block?,How did get through it?
OTHERS (OTH)	for any other tweet that doesn't fall in any of the above categories	This! RT @AnamRathor, #yesallwomen

TABLE III
ACCURACY AND F1-SCORE OF ALL THE BASELINES AND THE PROPOSED MODEL. † REPRESENTS THAT THE RESULTS ARE STATISTICALLY SIGNIFICANT

Model	Accuracy †	F1-score † (Weighted Avg.)	Prediction Time (in sec)
BERT (Baseline-1)	73.17%	0.7173	0.979×10^{-3}
BERT-Caps + CE loss (Baseline-2)	71.27%	0.7009	1.680×10^{-3}
Capsule (Baseline-3)	68.68%	0.6595	0.982×10^{-3}
Capsule + CE loss (Baseline-4)	52.10%	0.3582	0.972×10^{-3}
Proposed without BERT (Baseline-5)	75.74%	0.7487	1.562×10^{-3}
Baseline-5 + CE loss (Baseline-6)	70.38%	0.6810	1.549×10^{-3}
n-gram CNN + Bi-Recurrent (Baseline-7)	73.02%	0.7186	1.133×10^{-3}
Bi-Recurrent + Capsule (Baseline-8)	74.66%	0.7349	1.533×10^{-3}
Baseline-8 + CE loss (Baseline-9)	72.21%	0.7066	1.510×10^{-3}
Bi-Recurrent (Baseline-10)	72.23%	0.7102	0.511×10^{-3}
n-gram CNN (Baseline-11)	70.38%	0.6889	0.818×10^{-3}
Proposed without extra features (Baseline-12)	76.39%	0.7535	1.647×10^{-3}
BERT-Caps	**77.52%**	**0.7719**	1.840×10^{-3}

2) *BERT-Caps With Categorical Cross Entropy (Baseline-2):* This baseline was curated to study the importance of capsule loss for the task.
3) *Original Capsule Model (Baseline-3):* For this baseline, the original capsule model was implemented without the usage and fine-tuning of the BERT model. The results here are reported with the Glove embeddings [34].
4) *Original Capsule Model With Categorical Cross Entropy (Baseline-4):* In baseline-3, instead of capsule loss, we employ the categorical cross-entropy loss to report the results of this baseline.
5) *Proposed Model Without BERT (Baseline-5):* For this baseline, we exclude the usage of the BERT model to fine-tune the embeddings from our model. Instead, we utilize the Glove embeddings and the capsule loss to report the results.
6) *Proposed Model Without BERT and Capsule Loss (Baseline-6):* In baseline-5, instead of capsule loss, we use the categorical cross-entropy loss with the configuration to report the results for this baseline.
7) *Convolutional and Recurrent Model (Baseline-7):* In this baseline, we remove the usage of BERT and capsule layer from the proposed model. Thus, this baseline now has the N-gram convolutional and the bidirectional recurrent layer with categorical cross-entropy loss.
8) *Recurrent Capsule Model (Baseline-8):* From the proposed model, we omit the usage of BERT and the N-gram convolutional layer. Thus, the baseline now contains a bidirectional recurrent layer and the primary capsule layer along with capsule loss.
9) *Recurrent Capsule Model With Categorical Cross Entropy (Baseline-9):* In baseline-8, instead of capsule loss, we employ the categorical cross-entropy loss to report the results.
10) *Bidirectional Recurrent Model (Baseline-10):* In this baseline, we only utilize the bidirectional recurrent layer with categorical cross-entropy loss from our proposed model to report the results. We utilize the Glove embeddings for the same.
11) *N-gram Convolutional Model (Baseline-11):* In this baseline, we only utilize the N-gram convolutional layer with categorical cross-entropy loss from our proposed model to report the results. Similarly, here also, we utilize the Glove embeddings.
12) *Without Extra Features (Baseline-12):* For this baseline, we remove all the additional features from BERT-Caps. This baseline is to highlight the robustness of the proposed model and to observe the influence of extra features.

Table III shows the results in terms of accuracy and F1-score for all the baselines and the proposed model. It is clearly evident from the table that the proposed BERT-Caps produced better results compared with all other baselines by a significant margin. Compared with the best baseline, i.e., baseline-5,

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

Model	Class						
	REQ	QUE	SUG	STM	THT	EXP	OTH
BERT (Baseline-1)	0.57	0.67	0.50	0.65	0.51	0.82	0.60
BERT-Caps + CE loss (Baseline-2)	0.56	0.80	0.30	0.56	0.61	0.80	0.62
Capsule (Baseline-3)	0.36	0.48	0.28	0.60	0.49	0.79	0.58
Capsule + CE loss (Baseline-4)	0.15	0.36	0.21	0.32	0.09	0.69	0.15
Proposed without BERT (Baseline-5)	0.59	0.76	0.53	0.66	0.63	0.84	0.62
Baseline-5 + CE loss (Baseline-6)	0.21	0.76	0.45	0.58	0.50	0.79	0.67
n-gram CNN + Bi-Recurrent (Baseline-7)	0.64	0.74	0.50	0.61	0.58	0.81	0.61
Bi-Recurrent + Capsule (Baseline-8)	0.61	0.76	0.47	0.64	0.64	0.83	0.57
Baseline-8 + CE loss (Baseline-9)	0.68	0.76	0.48	0.60	0.59	0.80	0.59
Bi-Recurrent (Baseline-10)	0.56	0.69	0.49	0.60	0.60	0.75	0.54
n-gram CNN (Baseline-11)	0.54	0.70	0.44	0.59	0.57	0.72	0.57
Proposed without extra features (Baseline-12)	0.51	0.81	0.52	0.61	0.69	0.84	0.61
BERT-Caps	**0.69**	**0.81**	**0.57**	**0.64**	**0.82**	**0.85**	**0.67**

which is the proposed model without the usage and fine-tuning of BERT, the proposed BERT-Caps showed an improvement of almost 2%. This gain is rather intuitive as the proposed BERT-Caps leverages from the joint optimization of the BERT along with the capsule network in the given configuration. In comparison to baseline-1, which is the original BERT model without any fine-tuning, BERT-Caps showed an improvement of more than 4%. This highlights the fact that fine-tuning of the BERT embeddings over task-specific data set improves the performance of the classifier. All the variants with the capsule loss gave better results than its counterpart using categorical cross entropy as is evident from the comparison of the proposed model with baseline-2, where the improvement for the BERT-Caps is almost 6%. Apart from accuracy and F1-score, prediction, i.e., testing times of all the models for the test set (i.e., 1547 tweets), is also provided. As seen, the prediction time of the proposed model is slightly higher than all other baseline models. This is evident as the proposed BERT-Caps has several layers of neural network processing compared with the baselines. Also, prediction times of all the models including baselines and proposed lie in the same range with no significant difference among them. However, with such evolving computational power at hand in recent times, deploying huge DL networks is not too expensive and has the capability of processing an enormous amount of data at a given time. For example, Baseline-5 and Baseline-12 are variants of each other with an improvement in accuracy of 0.69%, with the latter taking 0.085 amount of time more than the former. With a marginal increase in testing time, an increase in accuracy is desirable in recent times. More so, in dealing with platforms, such as Twitter, it becomes even more crucial to monitor tweets with improved accuracy and precision. Thus, to capture fine-grained features from noisy data, developing complex neural networks that can extract such minute details cannot be avoided. In applications such as rumor detection in Twitter [17], [35], identification of tweet acts forms the first step in filtering out tags, such as "STM" and "THT" for further processing in identifying rumors. Thus, tweets acts must be identified even more accurately to support all these related subtasks. In such cases, prediction time can be compromised in order to attain higher accuracy.

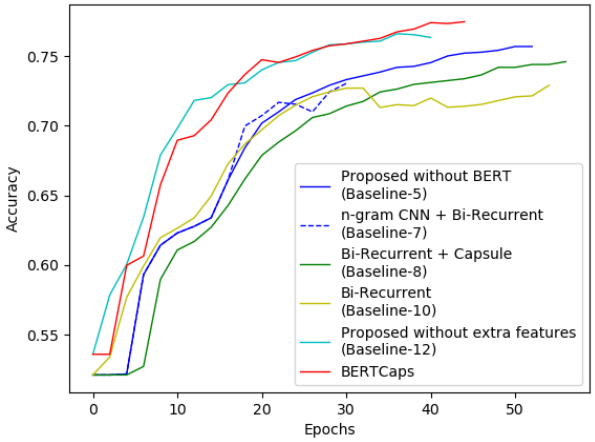

Fig. 3. Learning curve of best five performing baselines and the proposed models in terms of accuracy.

The learning curve of models in terms of accuracy over epochs for five baselines and the proposed model is shown in Fig. 3. As evident, the proposed BERT-Caps attains an improved accuracy of 77.52% in lesser no. of epochs than other baselines that take comparatively more no. of epochs to converge to their maximum values. At about epoch 38, BERT-Caps surpasses Baseline-12 in terms of accuracy and continues to move higher up. This shows that Twitter-specific features assist the model to capture fine-grained difference among numerous tweets, thus stressing the role of incorporating such features into the model. In the case of skewed data set where some of the tags are underrepresented than the others, measures, such as precision, recall, and F1-score, provide a better understanding of the quality of the trained model. These measures quantify that the attained accuracy is solely not because of correct predictions of one or two major classes of the data set. It is equally important for the model to learn the minute difference among all the classes and predict fewer representative tags correctly. Visualization of precision and recall measures of the same five baselines and the proposed model is also shown in Fig. 4. Here, it is evident from the figure that the precision and recall values of the proposed BERT-Caps are higher compared with all other

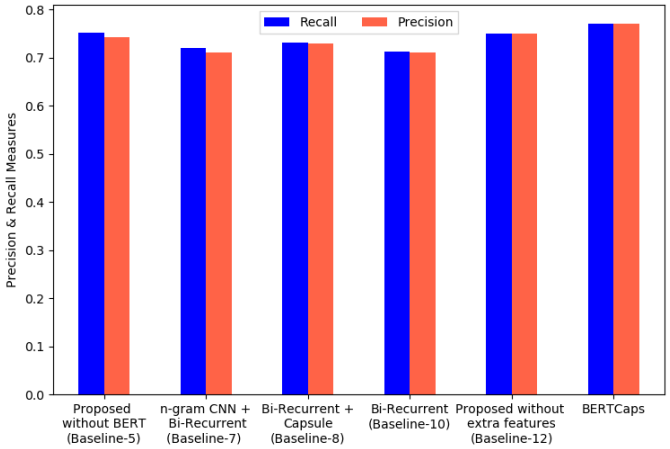

Fig. 4. Precision and recall of best five performing baselines and the proposed models.

TABLE V
CONFUSION MATRIX FOR THE PROPOSED BERT-CAPS

	STM	EXP	QUE	OTH	SUG	REQ	THT
STM	**170**	100	5	4	4	2	8
EXP	47	**719**	20	11	7	0	2
QUE	4	17	**118**	0	0	0	3
OTH	13	20	0	**56**	1	1	0
SUG	4	29	0	3	**34**	1	0
REQ	1	8	7	4	3	**21**	0
THT	5	3	1	0	0	0	**48**

TABLE VI
RESULTS OF THE PROPOSED BERT-CAPS MODEL IN TERMS OF ACCURACY AND F1-SCORE FOR VARYING ITERATIONS

Iterations	Accuracy	F1-score
2	77.32%	0.76
3	76.92%	0.75
5	77.52%	0.77

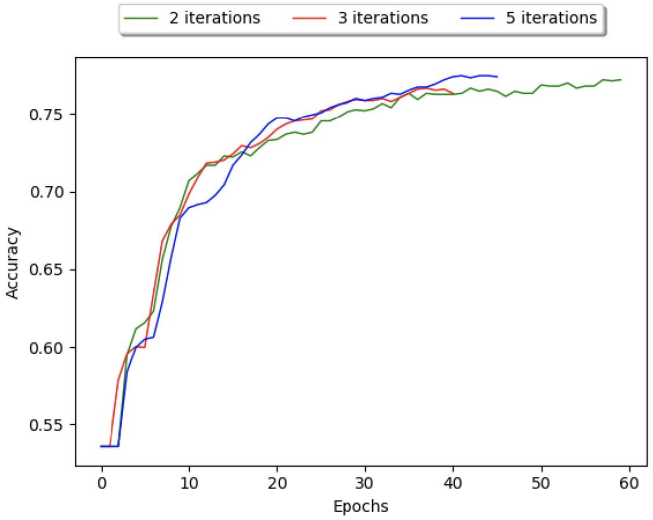

Fig. 5. Influence of routing iteration on BERT-Caps.

baselines, thus highlighting the fact that it is able to learn subtle differences among tweets. The F1-score of individual tags for all the baselines and the proposed model is shown in Table IV. As seen in the table, F1-scores of all the tags are better for the proposed model compared with all other baselines. Tags such as "THT" that have relatively less number of instances in the data set were also learned and classified well. Table V shows the confusion matrix of the proposed BERT-Caps model. All the reported results are statistically significant as we have performed pairwise Welch's t-test [36] at a 5% significant level.

B. Routing Iteration

As discussed earlier, the coupling coefficients u_{ab} are updated by the dynamic routing algorithm that basically helps in determining the connection strength between the capsules. In order to determine the ideal updating iteration for coupling coefficients, we conduct experiments with BERT-Caps in varying iterations (2, 3, and 5). The accuracy and F1-score in these configurations are reported in Table VI. As shown in Fig. 5, the proposed model with five iterations converges faster and gave the best results among all. Thus, a capsule network with five iterations is used in all the experiments as relevant.

C. Error Analysis

We also performed a thorough analysis to understand where our proposed model faltered, which are given as follows.

1) *Skewed Data Set:* One of the fundamental reasons being the skewed data set, i.e., contribution of most the tweet acts in the data set is very less with much of the representation from "EXP" and "STM" tags. Thus, much of the confusion were observed between these two classes, as shown in Tables IV and V. For example, tweet such as "RT @SalmanRushdie: 3500 signed pages to be bound into the finished copies of the new novel. #MyWeekend." was wrongly tagged as "EXP" instead of "STM." Compared with the baselines, the proposed system performed pretty fairly for the "THT" tag even though it had lesser instances in the data set.
2) **Fine-Grained Tag:** Tag such as "SUG" suffers massively with respect to F1-score as the majority of the tweets of this category are a subset of "EXP" tags with the former being a fine-grained tag of the latter. For example, "@$jk\_rowling$ should make a sequel to Harry Potter where there's a wizarding College/university" has been misinterpreted as "EXP" instead of "SUG." Also, the number of instances of "SUG" tag is relatively less in the released data set. As seen from the confusion matrix, similar confusions were noticed between classes where tweet such as "#Gaza's Great March of Return: Crisis & Causes via @AlShabaka #GreatReturnMarch in context #OpenGaza. End denial of fundamental rights-persecution-of 2 million Palestinians in Gaza #ICC" is misclassified as "STM" rather than "THT."
3) **Miscellaneous Instances:** Tag such as "OTH" also suffers in terms of F1-score because there is no predefined

TABLE VII
SAMPLE UTTERANCES WITH ITS PREDICTED LABEL FOR THE PROPOSED BERT-CAPS AND BEST TWO PERFORMING (BASELINE-1 AND BASELINE-5) MODELS

Tweet	True Label	BERT-Caps Predicted Label	Proposed without BERT (Baseline-5) Predicted Label	BERT (Baseline-1) Predicted Label
I absolutely love Harry Potter! I wonder what hogwarts would be like with Harry and Ron's children there?	EXP	QUE	EXP	QUE
rt <user>To all men. Go read some # <hashtag>posts. To all women. Take part. It's important we hear it.	SUG	EXP	EXP	STM
rt <user>'Gulabi Gang find abusive husbands & threaten to beat them up with sticks if they hurt their wives again	STM	THT	THT	THT
Rape Victims Shouldn't be the ones that apologise for the actions of the rapist.	EXP	EXP	STM	THT
Mr. Trump, you wanna ban the real terrorists? Start with your supporters that are threatenin' our federal judges.	SUG	SUG	EXP	QUE

nature of the tweet that falls in this category with miscellaneous instances.

Table VII shows the predicted labels for a few sample utterances from the BERT-Caps model as well as the top two best performing baselines, i.e., Baseline-1 and Baseline-5.

D. Comparison With the State of the Art

We also perform a comparative study of the state-of-the-art models [14]–[16], [37], [37] with the results reported in the article [16]. This is because we solely use this data set 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 VIII shows the results of all the state of the art approaches. As is evident from the table, our proposed BERT-Caps model outperformed all other state-of-the-art approaches by a significant margin.

The performance gain of the proposed BERT-Caps can be attributed to the following.

1) The presence of BERT embeddings, which provides deep bidirectional contextual representations by multiple attention heads attending to different sections of input in parallel.
2) One of the primary benefits of a capsule network is that it preserves the object location, whereas the conventional convolutional network loses this information as the pooling layers extract only the most meaningful information.
3) The proposed BERT-Caps leverages from the joint optimization of these two significant layers to learn features pertaining to speech acts and Twitter. Analogous to images where spatial orientation is important to identify objects correctly, for example, the presence and the position of say two simple entities, triangle, and rectangle, (say) work together to identify the presence of a complex entity (say) house.

Similarly, in texts, such kind of orientation in terms of local ordering of words and its semantic representation are crucial to be captured. For example, the presence and position of words, such as "Why" and "?," are important to identify the presence of the tag, such as "QUE." The model aims to capture more fine-grained semantic relationships and its relationships in relation to the ordering of words in a tweet by deciding the importance and agreement of words for a particular prediction task.

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

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
Sentence Bert-LR (Reimers et al., 2019)	74.63%	0.73
Sentence Bert-MLP (Reimers et al., 2019)	75.07%	0.74
Bert-Capsule (Vlad et al., 2019)	74.81%	0.73
Bert-Bi-LSTM-Capsule (Vlad et al., 2019)	75.99%	0.75
BERT-Caps (Our model)	**77.52%**	**0.77**

VII. CONCLUSION

In this article, we present a BERT-Caps model for the identification of speech acts on Twitter. Our proposed model is based on leveraging from the joint optimization of the pretrained BERT model and capsule layer to learn features pertaining to speech acts and Twitter. Some shallow handcrafted features were also incorporated into the model to boost its robustness. We compare our proposed approach with several strong baselines and outperformed state-of-the-art approaches. Our model attained an overall accuracy and F1 measure of 77.52% and 0.77, respectively.

In the 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.

ACKNOWLEDGMENT

Dr. Sriparna Saha would like to thank 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

[1] Y. Lin. (Jul. 2019). *Twitter Users Statistics 2019 Infographics*. [Online]. Available: https://www.oberlo.in/blog/twitter-statistics
[2] A. C. Pandey, D. S. Rajpoot, and M. Saraswat, "Twitter sentiment analysis using hybrid cuckoo search method," *Inf. Process. Manage.*, vol. 53, no. 4, pp. 764–779, Jul. 2017.
[3] F. Laylavi, A. Rajabifard, and M. Kalantari, "Event relatedness assessment of Twitter messages for emergency response," *Inf. Process. Manage.*, vol. 53, no. 1, pp. 266–280, Jan. 2017.

[4] S. M. Mohammad, X. Zhu, S. Kiritchenko, and J. Martin, "Sentiment, emotion, purpose, and style in electoral tweets," *Inf. Process. Manage.*, vol. 51, no. 4, pp. 480–499, Jul. 2015.
[5] J. L. Austin, *How to do Things With Words*. Oxford, U.K.: Oxford Univ. Press, vol. 88, 1975.
[6] J. R. Searle and J. R. Searle, *Speech Acts: An Essay in the Philosophy of Language*. Cambridge, U.K.: Cambridge Univ. Press, 1969, vol. 626.
[7] J. R. Searle, "A taxonomy of illocutionary acts," in *Language, Mind and Knowledge*. Minneapolis, MN, USA: Univ. of Minnesota, 1975, pp. 344–369.
[8] A. Stolcke *et al.*, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Comput. Linguistics*, vol. 26, no. 3, pp. 339–373, Sep. 2000.
[9] H. Khanpour, N. Guntakandla, and R. Nielsen, "Dialogue act classification in domain-independent conversations using a deep recurrent neural network," in *Proc. COLING-26th Int. Conf. Comput. Linguistics, Tech. Papers*, 2016, pp. 2012–2021.
[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Minneapolis, MN, USA, vol. 1, Jun. 2019, pp. 4171–4186.
[11] Q. Chen, Z. Zhuo, and W. Wang, "BERT for joint intent classification and slot filling," 2019, *arXiv:1902.10909*. [Online]. Available: http://arxiv.org/abs/1902.10909
[12] G. Castellucci, V. Bellomaria, A. Favalli, and R. Romagnoli, "Multilingual intent detection and slot filling in a joint BERT-based model," 2019, *arXiv:1907.02884*. [Online]. Available: http://arxiv.org/abs/1907.02884
[13] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3856–3866.
[14] R. Zhang, D. Gao, and W. Li, "What are tweeters doing: Recognizing speech acts in Twitter," in *Proc. AAAI Workshop Analyzing Microtext*, 2011, pp. 86–91.
[15] S. Vosoughi and D. Roy, "Tweet acts: A speech act classifier for twitter," in *Proc. ICWSM*, 2016, pp. 711–715.
[16] T. Saha, S. Saha, and P. Bhattacharyya, "Tweet act classification: A deep learning based classifier for recognizing speech acts in Twitter," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.
[17] S. Vosoughi, "Automatic detection and verification of rumors on Twitter," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, USA, 2015.
[18] C. Cerisara, S. Jafaritazehjani, A. Oluokun, and H. T. Le, "Multi-task dialog act and sentiment recognition on mastodon," in *Proc. 27th Int. Conf. Comput. Linguistics (COLING)*, Santa Fe, NM, USA, Aug. 2018, pp. 745–754.
[19] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proc. ICASSP-IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Mar. 1992, pp. 517–520.
[20] S. Grau, E. Sanchis, M. J. Castro, and D. Vilar, "Dialogue act classification using a Bayesian approach," in *Proc. 9th Conf. Speech Comput.*, 2004.
[21] H. Kumar, A. Agarwal, R. Dasgupta, and S. Joshi, "Dialogue act sequence labeling using hierarchical encoder with CRF," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.
[22] D. Verbree, R. Rienks, and D. Heylen, "Dialogue-act tagging using smart feature selection; results on multiple corpora," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2006, pp. 70–73.
[23] N. Kalchbrenner and P. Blunsom, "Recurrent convolutional neural networks for discourse compositionality," in *Proc. Workshop Continuous Vector Space Models Compositionality*. Stroudsburg, PA, USA: Association Computational Linguistics, 2013, pp. 119–126. [Online]. Available: http://aclweb.org/anthology/W13-3214
[24] Y. Liu, K. Han, Z. Tan, and Y. Lei, "Using context information for dialog act classification in DNN framework," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 2170–2178.
[25] 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," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.
[26] C. Sankar and S. Ravi, "Deep reinforcement learning for modeling chit-chat dialog with discrete attributes," 2019, *arXiv:1907.02848*. [Online]. Available: http://arxiv.org/abs/1907.02848
[27] M. Jeong, C.-Y. Lin, and G. G. Lee, "Semi-supervised speech act recognition in emails and forums," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Stroudsburg, PA, USA: Association Computational Linguistics, vol. 3, 2009, pp. 1250–1259.
[28] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 5998–6008.
[29] Y. Zhu *et al.*, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 19–27.
[30] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*. [Online]. Available: http://arxiv.org/abs/1609.08144
[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
[32] L. Xiao, H. Zhang, W. Chen, Y. Wang, and Y. Jin, "MCapsNet: Capsule network for text with multi-task learning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 4565–4574.
[33] P. J. Stone, D. C. Dunphy, and M. S. Smith, *The General Inquirer: A Computer Approach to Content Analysis*. Oxford, U.K.: MIT Press, 1966, p. 351.
[34] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
[35] S. Vosoughi and D. Roy, "A human-machine collaborative system for identifying rumors on Twitter," in *Proc. IEEE Int. Conf. Data Mining Workshop (ICDMW)*, Nov. 2015, pp. 47–50.
[36] B. L. Welch, "The generalization of student's' problem when several different population variances are involved," *Biometrika*, vol. 34, nos. 1–2, pp. 28–35, 1947.
[37] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Stroudsburg, PA, USA: Association Computational Linguistics, 2019, pp. 1–11. [Online]. Available: http://arxiv.org/abs/1908.10084
[38] G.-A. Vlad, M.-A. Tanase, C. Onose, and D.-C. Cercel, "Sentence-level propaganda detection in news articles with transfer learning and BERT-BiLSTM-Capsule model," in *Proc. 2nd Workshop Natural Lang. Process. Internet Freedom, Censorship, Disinformation, Propaganda*, 2019, pp. 148–154.

Tulika Saha (Graduate Student Member, IEEE) received the M.Tech. degree in software engineering from the National Institute of Technology Durgapur, Durgapur, India, in 2016.

She is currently a Research Scholar with the Department of Computer Science and Engineering (CSE), IIT Patna, Patna, India. She has authored research articles in *Cognitive Computation*, *PLOS One*, *Expert Systems with Applications*, and *Multimedia Tools and Applications* and in some peer-reviewed international conferences, such as the Association for Computational Linguistics (ACL) and the International Joint Conference on Neural Networks (IJCNN). Her research interests broadly include addressing several aspects of natural language processing, especially dialog systems and social media-based conversations, deep learning, and reinforcement learning.

Srivatsa Ramesh Jayashree received the bachelor's degree in computer science and engineering from IIT Patna, Patna, India, in 2020.

He is currently a Software Development Engineer with OYO Rooms, India. His current research interests include deep learning, multiobjective optimization, clustering, and natural language processing. His other areas of interest include cloud computing, distributed systems, and product management.

Sriparna Saha (Senior Member, IEEE) received the master's and Ph.D. degrees in computer science from the Indian Statistical Institute, Kolkata, India, in 2005 and 2011, respectively.

She is currently an Associate Professor with the Department of Computer Science and Engineering, IIT Patna, Patna, India. She has authored or coauthored more than 120 articles. Her current research interests include machine learning, pattern recognition, multiobjective optimization, language processing, and biomedical information extraction.

Dr. Saha was a recipient of several awards, including the Lt Rashi Roy Memorial Gold Medal from the Indian Statistical Institute for outstanding performance in M.Tech. (computer science), the Google India Women in Engineering Award in 2008, the NASI Young Scientist Platinum Jubilee Award in 2016, the BIRD Award in 2016, the IEI Young Engineer's Award in 2016, the SERB Women in Excellence Award in 2018, the SERB Early Career Research Award in 2018, the Humboldt Research Fellowship, the Indo–U.S. Fellowship for Women in STEMM (WISTEMM) Women Overseas Fellowship Program in 2018, and the CNRS Fellowship. She was a recipient of the India4EU Fellowship of the European Union to work as a Post-Doctoral Research Fellow at the University of Trento, Trento, Italy, from September 2010 to January 2011, and the Erasmus Mundus Mobility with Asia (EMMA) Fellowship of the European Union to work as a Post-Doctoral Research Fellow at Heidelberg University, Heidelberg, Germany, from September 2009 to June 2010.

Pushpak Bhattacharyya received the B.Tech. degree from IIT Kharagpur, Kharagpur, India, in 1984, the M.Tech. degree from IIT Kanpur, Kanpur, India, in 1986, and the Ph.D. degree from IIT Bombay, Mumbai, India, in 1994.

He is currently a Distinguished Alumnus of IIT Kharagpur and an Abdul Kalam National Fellow. He is also a Computer Scientist and a Professor with Computer Science and Engineering Department, IIT Bombay, where he also heads Natural Language Processing Research Group, Center for Indian Language Technology (CFILT) Laboratory, and the AI-NLP-ML Laboratory, IIT Patna, Patna, India. He is also the Former Director of IIT Patna. He has led several government and industry projects of international and national importance. He has authored or coauthored more than 100 articles in top-tier NLP conferences, such as ACL, International Conference on Computational Linguistics (COLING), North American Chapter of the ACL (NAACL), Conference on Natural Language Learning (CoNLL), Empirical Methods in Natural Language Processing (EMNLP), and International Conference on Language Resources and Evaluation (LREC). He has authored the textbook, *Machine Translation*. His current research interests include natural language processing, artificial intelligence, machine learning, psycholinguistics, eye tracking, information retrieval, and Indian language WordNets—IndoWordNet. A significant contribution of his research is multilingual lexical knowledge bases, such as IndoWordNet and Projection.

Prof. Bhattacharyya is also a fellow of the Indian National Academy of Engineering (FNAE). He was a recipient of faculty grants from IBM, Microsoft, Yahoo, and the United Nations. For sustained contribution to technology, he received the P. K. Patwardhan Award of IIT Bombay in 2008, the Manthan Award of the Ministry of IT in 2009, and the VNMM Award of IIT Roorkee in 2014. He was the President of the Association for Computational Linguistics from 2016 to 2017 and the Vijay and Sita Vashee Chair Professor.