Fine-tuned BERT Model for Multi-Label Tweets Classification

Hamada M. Zahera*
DICE, Paderborn University
Paderborn, Germany
hamada.zahera@uni-paderborn.de

Ibrahim Elgendy
Harbin Institute of Technology
Harbin, China
ibrahim.elgendy@hit.edu.cn

Rricha Jalota
DICE, Paderborn
Paderborn, Germany
rricha.jalota@uni-paderborn.de

Mohamed Ahmed Sherif
DICE, Paderborn
Paderborn, Germany
mohamed.sherif@uni-paderborn.de

Abstract

In this paper, we describe our approach to classify disaster-related tweets into multi-label information types (i.e, labels). We aim to filter first relevant tweets during disasters. Then, we assign tweets relevant information types. Information types can be *SearchAndRescue, MovePeople and Volunteer*. We employ a fine-tuned BERT model with 10 BERT layers. Further, we submitted our approach to the TREC-IS 2019 challenge, the evaluation results showed that our approach outperforms the F1-score of median score in identifying actionable information.

1 Introduction

Social media plays an integral part during tragedies and disasters [1] by providing real-time updates that are crucial for relief and rescue operations. Several studies [2, 3] have been conducted to mitigate disaster impact and fasten response time by leveraging this data. For instance, Toriumi et al. [4] developed a real-time information sharing system to emanate awareness during disasters using a clustering-based tweet classification approach. On the other hand, a semantic-neural network model was proposed by Burel et al. [5] to identify information categories in crisis-related twitter data. They leveraged semantic features (e.g., entity representation) rather than statistical ones to achieve a good classification accuracy.

Considering the large volume of information shared on social media, it is essential to not only effectively monitor it, but also categorize it in a way that can provide better situational awareness to emergency responders. Addressing this issue, TREC-IS[2] aims to classify social media data into different information types to aid public safety personnel during disasters/local incidents and for carrying out post-event analysis. Based on existing crisis management ontologies such as MOAC [3], these information types are modeled as multi-layer ontologies. To know more about the TREC-IS initiative, datasets and ontology, we suggest the interested readers to read [6]. Unlike last year, where the TREC-IS challenge was designed to assign single information type to every tweet, this year, it allowed multiple information types to be assigned to a tweet, thereby, making the problem more challenging. More details about TREC datasets including events, tweets and information types are discussed in section 3.

*Corresponding author
[2] http://dcs.gla.ac.uk/~richardm/TREC_IS/
[3] http://observedchange.com/moac/ns/

The rest of this paper is organized as follows: we first discuss relevant research works in Section 2. Then, we explore the analysis of TREC training dataset in section 3. Afterwards, we describe the details of our approach and official results in sections 4, 5 respectively. In section 6, we conclude the paper with some discussion about future work.

2 Related Work

In recent years, social-networking platforms such as Facebook and Twitter have exploded as a category of online discourse which leads to generate enormous amounts of data [7]. In addition, people pose huge amounts of time-critical and useful information on these platforms during disaster which are helpful to the humanitarian organizations for disaster response efforts. However, nding and detecting the posts and tweets which are related to an ongoing event is not trivial [3]. Many approaches have been developed for detecting the tweets related to the disaster. Some of these approaches have used traditional machine learning techniques while recently, Deep Learning techniques have been adopted as an effective methods for classifying tweets during a disaster situation [8, 9, 10]. In this section, a brief overview of the common models based on the solving methods are given as follows.

Regarding traditional methods, Avvenuti et al., have developed a social media-based system for earthquake detection based on both tweet and reply from Twitter[11]. Then, URL, mention, words, character, punctuation and slang/offensive words are used as features in classification stage to reduce the irrelevant information. Finally, they created a burst detection method for temporal analysis which observes the number of message in time window. Whereas the authors in [12] have mined Twitter data for real-time earthquake detection in which Support Vector Machine (SVM) is used as classifier to remove irrelevant tweet at first. Then the probabilistic model based on poison process is created for temporal analysis which used to estimate the time moment as the earthquake happened. However these systems have some limitations in which the user must pre-define a set of features as a input which will affect on the overall performance. In addition, SVM has been shallow architecture while CNN has become as a popular technology with deep architecture.

Recently, some approaches used the deep learning techniques as an effective methods for classifying tweets during a disaster situation [13, 14]. For example, Caragea et al. [15] have proposed an approach based on Convolutional Neural Networks to identify informative messages in social media streams during disaster events. This approach has been shown a significant improvement in performance over models that use the "bag of words" and n-grams as features on several datasets of messages from flooding events. Another deep learning model has been introduced in [16]. The semantically-enhanced dual-CNN consists of two layers: a semantic layer that captures the contextual information and a traditional CNN layer. The results show that the dual-CNN model has a comparable performance with a single CNN.

3 Exploratory Data Analysis

The training data, provided in TREC-IS, comprised of 17,682 tweets spread over 25 information categories. Since it was a multi-label classification problem, every tweet belonged to at least one class. For our training, we dropped one information category, *Location*, because of its absence in the earlier TREC-IS runs and datasets. Fig. 1 shows the distribution of tweets in each of the remaining 24 categories. While on one hand, one class (*Sentiment*) had more than 6000 instances, another (*MovePeople*) had as few as 25 tweets only, depicting imbalance in the dataset. The test data consisted of 7,634 tweets. See Table 1 for the statistical information about the train and test datasets.

Table 1: TREC-IS Dataset.

#	Train	Test
Total No. of events	15	6
Total No. of tweets	17682	7634
No. of Information Types	24	25

In contrast to our approach [17] in TREC-2018, we significantly transformed our methodology this year, since the problem transitioned from multi-class to multi-label classification and consequently,

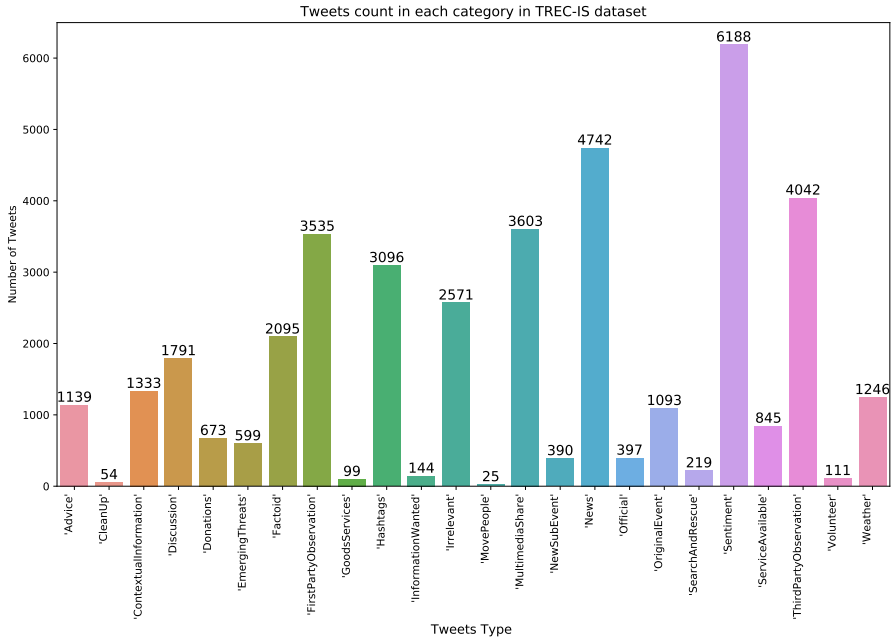

Figure 1: Percentages of Information Types per Tweets in Training and Test Dataset.

the challenge to label tweets correctly increased further. To tackle the problem of imbalanced dataset, we assigned weights to classes in an adaptive manner. That is, well represented classes were assigned lower class weights where as classes with only few instances were given higher weights. For classification of tweets, we considered their contextualized representation in vector space and tried to map this representation with the class labels. The idea was to make the classification model learn the latent semantics of polysemous words correctly.

4 The Approach

In this section, we discuss our proposed approach to categorize disaster-related tweets (collected by TREC-IS) into multi-label information types. First, we describe the preprocessing steps to maintain a normalized and cleaned tweets. Then, we present our fine-tuned BERT model into the task of multi-label classification of disaster tweets.

4.1 Tweets Preprocessing

Text from tweets are inherently noisy, cleaning the text before further processing helps to generate better features and semantics. We used *tweettokenize* API to preprocess tweets and create uniform representation, in particular we perform the following preprocessing steps.

- Stop-words, URLs, usernames and unicode-characters are removed.
- Extra white-spaces, repeated full stops, question marks and exclamation marks are removed.
- Emojis were converted to text using the python library `emoji`[4]
- Lemmatization, restoring language vocabulary to general form (can express complete semantics) by `WordNetLemmatizer`[5].
- Finally all tweet tokens are converted to lower-case.

[4] `https://pypi.org/project/emoji/`
[5] `https://www.nltk.org/_modules/nltk/stem/wordnet.html`

4.2 Fine-tuned BERT

BERT [18] is a pretrained language model with transformer architecture [19] that is designed to be easily applied with downstream NLP tasks with fine-tuned manner. After obtaining the sentence vectors from BERT, we build 10 BERT stacked layers on top of the BERT outputs to fine-tune BERT into multi-label classification of tweets. Then, we add an extra *dense* layer with *sigmoid* activation function. Further, we used two loss functions (binary cross-entropy, focal loss [20]) to minimize the errors during training our models.

The model's output are the probabilities of all classes, we hence set a threshold value to pickup a list of most relevant classes to the input tweet as final output.

$$\hat{Y} = \sigma(W_i \times X + b_i) \tag{1}$$

where X_i referred to input tweet i and σ donates the *sigmoid* function abd bias respectively.

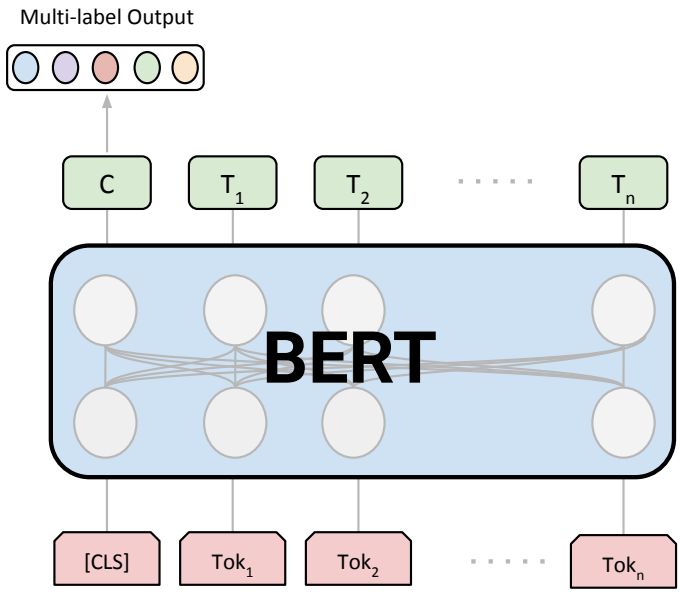

Figure 2: Fine-tuned BERT model

5 Evaluation

5.1 Dataset

The TREC-IS dataset contains 17,682 tweets for training and 7,634 for testing, which has been curated from different types of events (e.g. earthquakes, hurricanes, public or shootings). For each event, the tweets were collected using hashtags and keywords. Human annotators have labeled tweets into the multi-layer ontology of information types. In our experiments, we combined tweets from all the events provided for training, into one dataset.

5.2 Results

The official results are evaluated based on several metrics to demonstrate (i.e, *alert*) tweets with actionable information. Table 2 shows the results of our two submissions (runs) and the median scores. The run UPB-BERT, generated from training our fine-tuned BERT model with binary cross-entropy loss function, while UPB-FOCAL is generate from the same model with focal loss function. The F1 scores from two submissions (0.13, 0.12) are significantly outperform the median F1 score (0.03).

Table 2: Overall Performance from TREC-IS Run B 2019

RUN ID	AW-HP	AW-ALL	F1-Act.	F1-ALL	Acc.	RMSE Act.	RMSE-ALL
UPB-BERT	-0.95	-0.47	0.13	0.14	0.81	0.15	0.09
UPB-FOCAL	-0.93	-0.47	0.12	0.18	0.81	0.14	0.08
Median	-0.91	-0.46	0.03	0.10	0.85	0.17	0.10

However, our proposed models were able two classify TREC-IS tweets with accuracy (0.81) close to media (0.85).

Figures 3 and 4 show the performance of our submissions against all information types. It depicts that we got a higher precision and recall for categories (*News, Sentiment, MultimediaShare, Factoid*) that were better represented in the training dataset than the ones with fewer examples (*GoodServices, SearchAndRescue, CleanUp, Volunteer*, etc). This clearly shows that our models learned the more prominent information types better and could not generalize well over others.

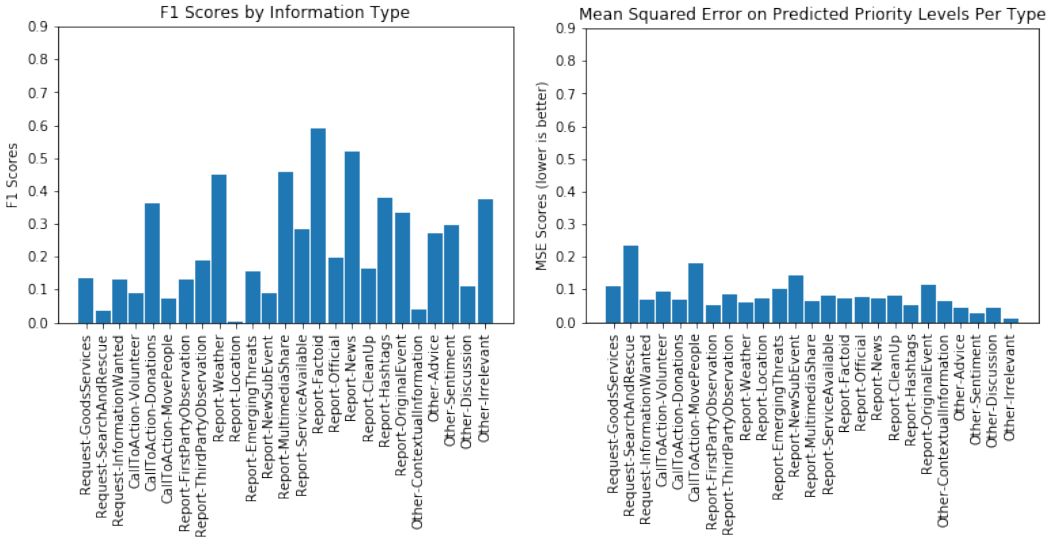

Figure 3: Performance evaluations (F1 and RMSE) from our submissions UPB-FOCAL

6 Conclusion and Future Work

In this paper, we described our two submissions to the TREC-IS 2019 track. Our approach employs contextualied word embedding from pre-trained BERT model to represent tweets features. We proposed two fined-tuned BERT models, first model (UPB-BERT) minimize training errors using binary cross-entropy loss function, while our second model (UPB-FOCAL) employs focal loss function. In the future, we plan to re-evaluate our approach with more training data and handle classes imbalances in multi-label classification problem.

Acknowledgements

This work has been supported by the BMVI projects LIMBO (project no. 19F2029C), and also by the German Federal Ministry of Education and Research (BMBF) within 'KMU-innovativ: Forschung für die zivile Sicherheit' in particular 'Forschung für die zivile Sicherheit' and the project SOLIDE (no. 13N14456).

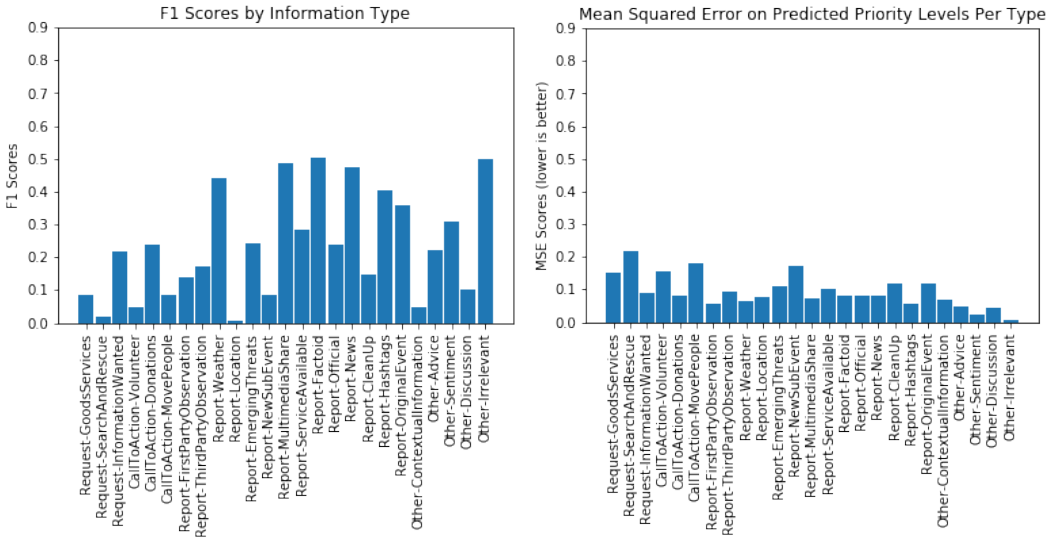

Figure 4: Performance evaluations (F1 and RMSE) from our submissions UPB-BERT

References

[1] Tomer Simon, Avishay Goldberg, and Bruria Adini. Socializing in emergencies—a review of the use of social media in emergency situations. *International Journal of Information Management*, 35(5):609–619, 2015.

[2] J Brian Houston, Joshua Hawthorne, Mildred F Perreault, Eun Hae Park, Marlo Goldstein Hode, Michael R Halliwell, Sarah E Turner McGowen, Rachel Davis, Shivani Vaid, Jonathan A McElderry, et al. Social media and disasters: a functional framework for social media use in disaster planning, response, and research. *Disasters*, 39(1):1–22, 2015.

[3] Peter M Landwehr and Kathleen M Carley. Social media in disaster relief. In *Data mining and knowledge discovery for big data*, pages 225–257. Springer, 2014.

[4] Fujio Toriumi and Seigo Baba. Real-time tweet classification in disaster situation. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11-15, 2016, Companion Volume*, pages 117–118, 2016.

[5] Grégoire Burel, Hassan Saif, and Harith Alani. Semantic wide and deep learning for detecting crisis-information categories on social media. In *International Semantic Web Conference*, pages 138–155. Springer, 2017.

[6] Richard Mccreadie, Cody Buntain, and Ian Soboroff. Trec incident streams: Finding actionable information on social media. 2019.

[7] Arshdeep Kaur. Analyzing twitter feeds to facilitate crises informatics and disaster response during mass emergencies. 2019.

[8] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, 47(4):67, 2015.

[9] Shamik Kundu and PK Srijith. *Classification of Short-Texts Generated During Disasters: Traditional and Deep learning Approach*. PhD thesis, Indian Institute of Technology Hyderabad, 2018.

[10] Sudha Subramani, Hua Wang, Huy Quan Vu, and Gang Li. Domestic violence crisis identification from facebook posts based on deep learning. *IEEE access*, 6:54075–54085, 2018.

[11] Marco Avvenuti, Stefano Cresci, Mariantonietta N La Polla, Andrea Marchetti, and Maurizio Tesconi. Earthquake emergency management by social sensing. In *2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS)*, pages 587–592. IEEE, 2014.

[12] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[13] Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176, 2015.

[14] Xiaocheng Feng, Bing Qin, and Ting Liu. A language-independent neural network for event detection. *Science China Information Sciences*, 61(9):092106, 2018.

[15] Cornelia Caragea, Adrian Silvescu, and Andrea H Tapia. Identifying informative messages in disaster events using convolutional neural networks. In *International Conference on Information Systems for Crisis Response and Management*, pages 137–147, 2016.

[16] Gregoire Burel and Harith Alani. Crisis event extraction service (crees)-automatic detection and classification of crisis-related content on social media. 2018.

[17] Hamada M Zahera, Rricha Jalota, and Ricardo Usbeck. Dice@ trec-is 2018: Combining knowledge graphs and deep learning to identify crisis-relevant tweets. In *TREC*, 2018.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.