Classification of Incident-related Tweets: Exploiting Word and Sentence Embeddings

Anna Kruspe*, Jens Kersten*, Friederike Klan*

**German Aerospace Center (DLR)*
Institute of Data Science, Jena, Germany
<firstname>.<lastname>@dlr.de

Abstract—**In this paper, we present our five approaches submitted to the Text REtrieval Conference (TREC) Incident Streams (IS) 2019B edition. The goal is to classify crisis-related tweets into a variable set of information classes and to provide an importance score. This multi-class, multi-label and multi-task problem turns out to be even more challenging because of extremely unbalanced training data available. We use recently proposed, publicy available word and sentence embeddings and deep neural network models for this task.**

I. Introduction

The 2019B Text REtrieval Conference (TREC) Incident Streams (IS) track serves as an evaluation for the classification of tweets into 25 incident-related classes. The first edition in 2018 was focusing on assigning a single class to each tweet. However, the complexity of message contents forced a change to a multi-label task.

Similar to the preceding two editions, a class ontology, an annotated training data set, and a test data set without annotations were provided. The ontology comprises 25 classes describing a variety of topics during an incident, such as "Report-ServiceAvailable", "Other-Sentiment", "Request-SearchAndRescue", or "CallToAction-Donations". Additionally, importance labels are defined by four classes: "low", "medium", "high" and "critical".

For training, we used the 2019A edition data that was provided for training and testing. This composed set contains around 24,800 tweets from 21 different crisis events. Due to slight adjustments and improvements of the class ontology over editions, parts of the training data had to be mapped accordingly. Submissions were expected to assign a set of n classes as well as an importance score to each tweet. Furthermore, an incident-wise ranking according to the estimated importance was requested.

We focused on training fully automatic deep neural network (DNN) models in order to contribute to these tasks. This paper describes our five submissions to the challenge. In the next section, a description of our classification approaches is provided. Furthermore, we present an analysis of the results and finish with a small conclusion.

II. Proposed models

In this section, we describe our five proposed models. The first one is the same as last year for comparison. As a new approach, we tested a DNN with pre-trained BERT word embeddings. We then turned to strategies for embedding the whole tweet (qua sentence) instead of separate words. Along these lines, we first tested the commonly used Smooth Inverse Frequency (SIF) approach. Finally, we built two models with pre-trained sentence embeddings, one with Google's Universal Sentence Encoder (USE), and one with mean-max attention autoencoder (Mean-Max AAE) embeddings.

A. Previous year's model: Fusion CNN

In the 2018 TREC-IS edition, we submitted a model based on Kim's CNN structure [6], which is being successfully used in other crisis-related tweet analysis tasks [2]. The model showed good performance, but suffered from a lack of training data for several classes in the TREC-IS 2018 data set. For this reason, we trained similar models on the *CrisisNLP* [5] and *CrisisLexT26* [8], [9] data sets, and combined all three into a fusion model. The concatenated outputs of all three models are fed into a dense layer with 128 nodes and ReLu activation, followed by the output layer. For this year's challenge, we did the same with the new training set. In order to perform multi-class labeling, the original softmax output was replaced with 25 independent sigmoid layers.

For importance scoring, the concatenated outputs are additionally passed to a sequence of two dense layers with ReLu activation (128 and 64 nodes), followed by a linear regression output node. For this, the importance annotations were mapped to numerical values ("low" = 0.0, "medium" =0.33, "high" = 0.66, "critical" = 1.0).

B. DNN with BERT word embeddings

For the fusion CNN model, a pre-trained word embedding specifically trained with crisis-related tweets [5] is utilized. This enables an immediate translation of each individual word in a document or tweet into a fixed numeric vector, capturing the semantic meaning of the word. However, depending on the context, words can have different semantic meanings.

Google's Bidirectional Encoder Representations from Transformer (BERT) [4] not only captures the semantic meaning of single words, but also the contextualized meaning. Here, deep bidirectional representations are learned by masking some percentage of the input tokens randomly, and then predict those masked tokens. We used a pre-trained TensorFlow Hub model[1] to obtain a 768-element vector representation for each tweet. Since this representation has shown to make a CNN obsolete in our experiments, we attached three dense layers (512, 256 and 128 nodes) with ReLu activation to the embedding layer, followed by 25 sub-networks for the information type classification, and one for the importance scoring. Each sub-network consists of a 64-dimensional fully-connected ReLU layer with a consecutive output layer. This output is a sigmoid for the 25 information type classes each, and a 1-dimensional linear node for the importance scoring (regression).

C. DNN with SIF sentence embeddings

The fusion model was focused on embedding each word in a tweet separately, and treating these embeddings independently (bag-of-words). Besides the contextual BERT embedding, several approaches for embedding whole sentences have been developed successfully in the past years. Li et al. give a nice overview over using those for crisis-related tweets in [7].

One such approach that does not require a dedicated pre-training is called Smooth Inverse Frequency (SIF) [1]. The sentence's words are encoded with an arbitrary embedding, on which an average weighted by word frequencies is calculated to represent the sentence. Then, a Principal Component Analysis (PCA) is performed on the sentence vectors, and the projection of the first principal component is subtracted. This method has been shown to beat several more elaborate approaches [1].

We utilized the authors' own Python implementation[2]. Word embeddings are generated using the *CrisisNLP* weights [5]. Word frequencies for English-language Twitter were obtained from *Lexique*[3]. Starting with a 256-node ReLu activation dense layer attached to the embdedding layer, the model is completed in the same manner as our BERT model.

D. DNNs with pre-trained sentence embeddings

In contrast with SIF, many other sentence embeddings function by training a specific sub-model for this task. This requires a large suitable data set from which the model can infer a latent representation. In the TREC-IS context, even the new, bigger data set is somewhat on the small side for this, especially due to the lack of training data for particular classes. Fortunately, several such sentence embeddings that have been pre-trained on very large data sets are available online.

We tested two such pre-trained models. The first one is the so-called Mean-Max Attention Autoencoder (Mean-Max AAE) presented by Zhang et al. [10]. It is based on an encoder-decoder structure with a MultiHead self-attention mechanism. A TensorFlow implementation including a model pre-trained on the *Toronto Books Corpus* by the authors is available online[4].

Second, we employed the Universal Sentence Encoder (USE) published by Cer et al. [3]. They show two versions, one using a Transformer structure and one using Deep Averaging Networks, and train them on a large set of combined data sources. The Transformer version is available from the authors on TensorFlow Hub[5].

In both cases, we attach the same post-network as with the SIF models to adapt the models for information type classification and importance scoring.

III. Results and Discussion

Our approaches' results on the validation set are shown in figures 1, 2, and 3. Over-all, these are roughly in the range of those submitted to the 2019A challenge.

In general, the information type classification task has become more difficult than least year due to the change to multilabeling. As figure 1 shows, our model from last year performs disappointingly in terms of F_1 measure. Fortunately, both the new BERT and USE approaches produce much better results. The other two sentence embedding models, SIF and MeanMaxAAE, are not as useful. We presume that this happens because they were not developed with Twitter data in mind. SIF was designed for other types of text, and the pre-trained MeanMaxAAE model was trained on text from books. For these reasons, they may not adapt well to tweets. On the other hand, USE was trained on a much larger variety of text, and as such seems to be more robust to Twitter text. A logical next step consists of developing dedicated tweet embedding models.

Even though the pre-trained BERT model was trained on *Wikipedia* and *Book Corpus* [11], i.e., data that tends to be quite different compared to tweets, the bidirectional pre-training shows up to be rather task-unspecific.

When comparing the result for all classes to that for only the high importance classes, we see that the high importance classification performs much worse in general. As we saw last year, these tweets occur much less frequently than other classes, and are therefore not highly represented in the training material, leading to difficulties in training models for these classes.

Figure 2a shows the mean squared errors for the priority estimation task. We observe the same trends here: BERT and USE are the best-performing models, and the estimation is more difficult for the high-importance classes. Over-all, the errors are relatively low, but somewhat higher than those achieved by other models in the 2019A iteration of the task. We note that the multi-task approach (i.e. using the same model for both tasks) generally works, but perhaps a refined model for the priority estimation is more suitable.

The Accumulated Alert Worth (AAW) results are shown in figure 2b. We observe a strange effect here: The models that performed better at both the classification and priority estimation tasks produce lower AAW values, and vice versa. A similar trend can be seen in some of the 2019A results;

[1] https://tfhub.dev/google/bert_uncased_L-12_H-768_A-12/1

[2] https://github.com/PrincetonML/SIF

[3] http://www.lexique.org/?page_id=250

[4] https://github.com/Zminghua/SentEncoding

[5] https://tfhub.dev/google/universal-sentence-encoder-large/3

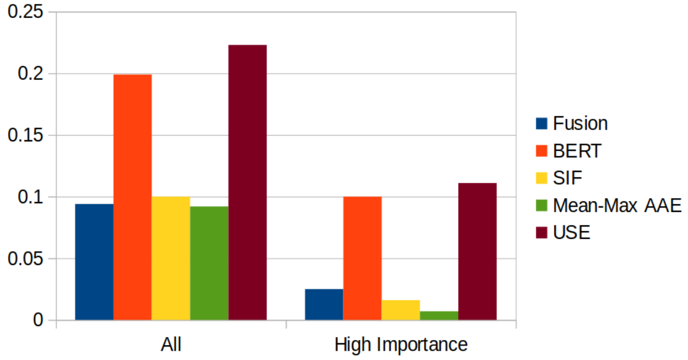
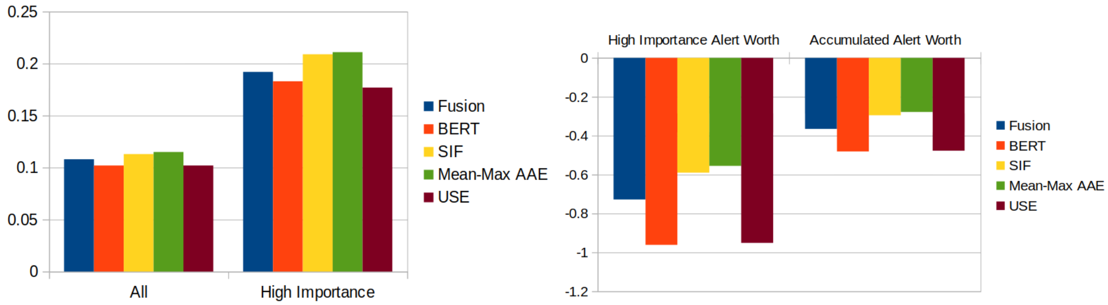


Fig. 1: Overall positive F1 scores for all (a) and high importance classes (b).



(a) Priority estimation error (MSE).

(b) High importance and accumulated alert worth.

Fig. 2: Priority estimation error (a) as well as high importance and accumulated alert worth (b).

we will look more closely into this. We are not sure why this happens; one possible explanation might be that the models performing worse at the priority estimation task may designate fewer tweets as high-importance, which leads to fewer tweets being taken into account for the AAW calculation.

A class-wise analysis of the positive results returned by the models is shown in figure 3, combined with the number of training examples per class given. We observe the same effect as last year: Classes with very little training data are commonly difficult to classify. This is particularly critical because those are often the most important classes. Tweets describing third- and first party observations are represented by a large number of training data, but obviously are hard to classify due to large possible variations of their content.

BERT and USE are a bit more robust to imbalanced training data than the other approaches. For both models, the reason is probably that they are already pre-trained on a lot of semantic knowledge, and therefore are able to generalize even with few training examples on this task. However, there is still a lot of room for improvement. The easiest solution to this problem would be a collection of more training data for these underrepresented classes. Exploiting models trained on a wide variety of other material also seems to be a promising direction.

IV. Conclusion

For the 2019B TREC-IS track, we submitted the results of five fully automatic text classification approaches. Our proposed DNNs with state-of-the-art pre-trained word and sentence embdeddings have shown to provide similar or even better results compared to our 2018 CNN fusion model. SIF and MeanMax AAE were not developed in context of Twitter data analysis. To obtain better results here, the development of dedicated sentence embeddings is required. In contrast, our BERT and USE models produced significantly better results, demonstrating the applicabilty of the involved pre-trained embeddings to a variety of applications and data.

However, this task still remains challenging and offers much room for improvements. One of the main and persistent challenges of this track is the inbalanced availability of training data. The easiest solution for this might be to manually collect additional training data from other sources. Our data augmentation approach from 2018 based on automatic round-trip translation might help to a certain degree. However, the obtained tweets might contain only slight variations and therefore are potentially redundant.

A possible future research direction is the development of dedicated pre-trained tweet embdeddings. Furthermore, taking into account the class ontology hierarchy as well as the inherent dependency of class labels might help to further improve tweet classification and prioritization.

References

[1] S. Arora, Y. Liang, and T. Ma. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations (ICLR)*, Toulon, France, Apr. 2017.

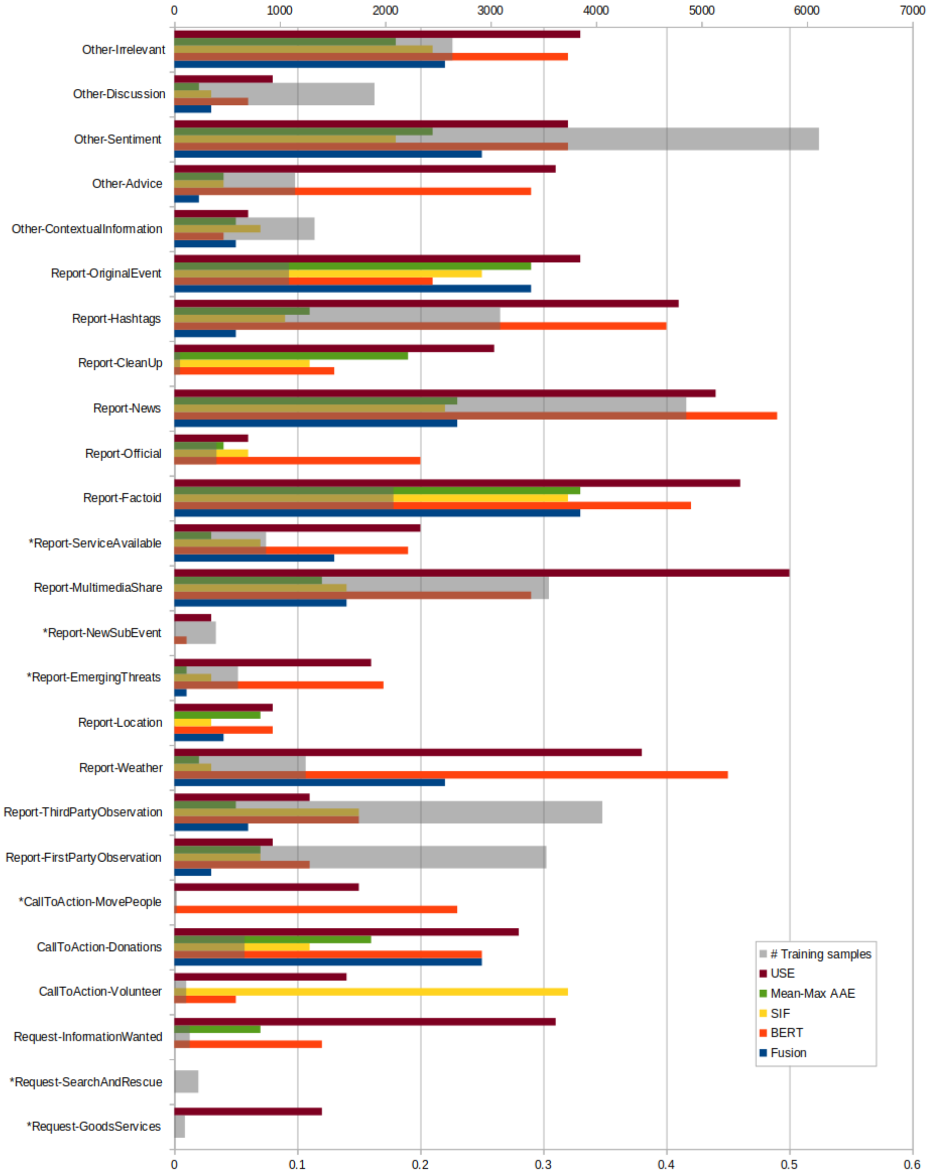


Fig. 3: Class-wise positive F1 scores for the five approaches (colored bars) and numbers of training samples per class (transparent gray bars); actionable (or high importance) classes are marked with an asterisk.

[2] G. Burel and H. Alani. Crisis Event Extraction Service (CREES) - Automatic Detection and Classification of Crisis-related Content on Social Media. In *15th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, Rochester, NY, USA, May 2018.

[3] D. Cer, Y. Yang, S. yi Kong, N. Hua, N. L. U. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cspedes, S. Yuan, C. Tar, Y. hsuan Sung, B. Strope, and R. Kurzweil. Universal sentence encoder. 2018.

[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[5] M. Imran, P. Mitra, and C. Castillo. Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. In *Tenth International Conference on Language Resources and Evaluation (LREC)*, Portoroz, Slovenia, May 2016.

[6] Y. Kim. Convolutional neural networks for sentence classification. In *2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, Oct. 2014.

[7] H. Li, X. Li, D. Caragea, and C. Caragea. Comparison of word embeddings and sentence encodings as generalized representations for crisis tweet classification tasks. In *ISCRAM Asian Pacific Conference*, Wellington, New Zealand, Oct. 2018.

[8] A. Olteanu, C. Castillo, F. Diaz, and S. Vieweg. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *AAAI Conference on Weblogs and Social Media (ICWSM)*, Ann Arbor, MI, USA, June 2014.

[9] A. Olteanu, S. Vieweg, and C. Castillo. What to expect when the unexpected happens: Social media communications across crises. In *Conference on Computer Supported Cooperative Work and Social Computing (ACM CSCW)*, Vancouver, BC, Canada, Mar. 2015.

[10] M. Zhang, Y. Wu, W. Li, and W. Li. Learning universal sentence representations with mean-max attention autoencoder. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Brussels, Belgium, Oct. 2018.

[11] Y. Zhu, R. Kiros, R. S. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *CoRR*, abs/1506.06724, 2015.