Transformer-based Methods with #Entities for Detecting Emergency Events on Social Media

Emanuela Boros
University of La Rochelle, L3i
La Rochelle, France
emanuela.boros@univ-lr.fr

Nhu Khoa Nguyen
University of La Rochelle, L3i
La Rochelle, France
nhu.nguyen@univ-lr.fr

Gaël Lejeune
Sorbonne University, STIH/CERES
Paris, France
gael.lejeune@sorbonne-universite.fr

Mickael Coustaty
University of La Rochelle, L3i
La Rochelle, France
mickael.coustaty@univ-lr.fr

Antoine Doucet
University of La Rochelle, L3i
La Rochelle, France
antoine.doucet@univ-lr.fr

ABSTRACT

This paper summarizes the participation of the L3i laboratory of the University of La Rochelle in the TREC Incident Streams 2021. This track aimed at identifying critical information present in social media by categorizing and prioritizing tweets in a disaster situation to assist emergency service operators. For both classifying tweets by information type and ranking tweets by criticality, we proposed a multitask and multilabel learning approach based on representing the tweet text and the event types with pre-trained language models, and by highlighting entities and hashtags. We also experimented with a bag of words representation and classical machine learning methods for the prioritization task. We conclude that, because our multitask approach can take advantage of both tasks, it achieved the best performance in comparison with different proposed ensembles. Our submissions obtained top performance for the prioritization task and surpassed the median performance for the information type classification task.

CCS CONCEPTS

• **Information systems** → **Data stream mining**; **Language models**.

KEYWORDS

Event detection, Named entity recognition, Transformer

1 INTRODUCTION

Natural disasters such as hurricanes, tsunamis, tornadoes, earthquakes, and floods are often a surprise or have little warning, result in deaths, injuries, and destruction, and require swift action to protect people and property. These types of major events and issues are often shared and communicated on Twitter before many other online and offline platforms. TREC[1] Incident Streams campaign aims at producing a series of curated feeds containing social media posts (Twitter), where each feed corresponds to a particular type of information request, aid request, or report containing a particular type of information. The main purpose is creating tools for filtering media stream down to actionable information or to route that information to the appropriate stakeholder (e.g., public health officials, emergency response officers, etc.).

TREC-IS 2021 is the fourth year of promoting research to better support emergency response services' efforts to harness social media, making it an ideal space to investigate new tools and techniques. In this paper, we describe our approach for this edition. We explore methods based on the Transformer architecture and classical machine learning algorithms. Our main approach consists of a stack of Transformer layers with pre-trained and fine-tuned language models trained in a multitask manner to predict both information type and the prioritization level. Furthermore, we take advantage of the presence of entities, the event type, and the title of the tweet. We present relevant related work related to previous TREC Incident Stream editions in Section 2 and the description of the TREC Incident 2021 in Section 3. We detail our methods in Section 4 and the experimental setup along with several ablation studies are presented in Section 5. We draw conclusions about our findings in Section 6.

2 RELATED WORK

TREC-IS was initiated in 2018 [21] and it developed test collections and evaluation methodologies for automatic or semi-automatic filtering approaches for the identification and the social media aid-requests categorization during crisis situations. Relevant crises for TREC-IS include six natural and man-made events: wildfires, earthquakes, floods, typhoons/hurricanes, bombings, and shootings. While the participants in this pilot edition were relatively effective at identifying common information types such as news reports and sentiment, identifying actionable information types like search and rescue requests proved to be still challenging. The participant systems were mostly based on bag of words representation (term frequency-inverse document frequency (TF-IDF) weighting) and classical machine learning techniques such as support vector machine (SVM) [8, 9, 12, 24]. Although they tended to over-estimate, they were more accurate at estimating information criticality.

TREC-IS 2019 [22] included two editions, 2019-A and 2019-B, two Twitter datasets with events retrospectively crawled by TREC-IS organizers (based on manually selected keywords), and modified performance metrics to differentiate between actionable information types and all types. Besides the classical bag-of-words or n-gram

[1] Text REtrieval Conference, http://trec.nist.gov

representation, this year's participants leveraged tweet text during categorization by converting text into a form of word or character sequence embedding (GloVe [29], FastText [1], BERT [10], ELMo [30]). The organizers reported that while deep learning methods were becoming more prevalent and effective, the traditional machine learning remained competitive, and thus, the majority using these approaches (e.g., Naive Bayes, Logistic Regression or Random Forests) were the most effective systems in terms of identifying actionable content for both editions [11, 23, 25].

TREC-IS 2020 [7] was separated in three tasks. High-Level information type classification was the task where tweet streams from a collection of crisis events should be classified as having one or more of the high-level information types described in an ontology provided by the organizers. The second included a restricted version of the first task that focused only on a subset of the high-level information types. This subset included the top six information types labeled as *actionable* in previous editions (i.e., the types that have, on average, the highest priority) as well as five other types selected from the full set used in the first task. While these two tasks continued the prior TREC-IS work on classifying information for wildfires, earthquakes, floods, typhoons/hurricanes, bombings, and shootings, TREC-IS 2020 proposed the third task to provide assistance to public health officials and emergency response officers with additional tools and evaluation data for future public health emergencies or resurgence of COVID-19. The organizers concluded that the additional complexity of a large information-type label space led to significant performance improvements compared to systems that are trained on a coarser label space [7, 16].

3 TREC-IS 2021

For standardized evaluations of systems, TREC-IS provided participants with training and test datasets, comprised of three components: the ontology of high-level information types, a collection of crisis-event descriptions, and the tweets for each event to be categorized. The participant TREC-IS systems are intended to produce two outputs for crisis-related social media content:

(1) Classifying tweets by *information type*, where each tweet should be assigned as many categories as are appropriate;
(2) Ranking tweets by their criticality (*priority*).

TREC-IS provided multiple Twitter datasets collected from a range of past wildfires, earthquakes, floods, typhoons/hurricanes, bombing, and shooting events. The metadata for each event is provided in an XML topic file as presented in Figure 2.

The information types as either top-level intent, high-level or low-level. For example, Figure 1 is a post regarding an event that happened on 4 August 2020, when a large amount of ammonium nitrate was stored at the port of the city of Beirut, the capital of Lebanon, accidentally exploded. This tweet has an information type of a top-level *Donations* that asks for a service to be provided. The tweet is of high priority with four information type categories: *Donations, Location, ContextualInformation, Sentiment*.

The provided training dataset consisted of a total of 73,499 tweets that covered 75 topics. We did not perform any pre-processing on the data.

> #Beirut #BloodDonors All blood types are needed across all hospitals in Lebanon right now - PLEASE donate if you can. Stay safe.

Figure 1: High priority [Donations, Location, ContextualInformation, Sentiment].

```
<top>
<num>66</num>
<dataset>beirutExplosion2020</dataset>
<title>Beirut Explosion</title>
<type>explosion</type>
<url>https://en.wikipedia.org/wiki/2020_Beirut_explosion</url>
<narr>On 4 August 2020, a large amount of ammonium nitrate
stored at the port of the city of Beirut, the capital of
Lebanon, accidentally exploded, causing at least 180 deaths,
6,000 injuries, US$10−15 billion in property damage,
and leaving an estimated 300,000 people homeless.
</narr>
</top>
<top>
```

Figure 2: Topic example from the tweet streams curated by the TREC-IS organizers.

4 MULTITASK INFORMATION TYPE AND PRIORITY TWEETS CLASSIFICATION

We propose a multitask and multilabel learning approach that consists in a hierarchical architecture with a fine-tuned RoBERTa-based encoder [20] and a stack of Transformer [33] blocks on top of the encoder [2–4, 6]. The multitask prediction layer consists of two separate linear dense softmax layers as presented in Figure 3.

A Transformer is a deep learning architecture based on multi-head attention mechanisms with *sinusoidal position embeddings*. In our implementation, we used *learned absolute positional embeddings* [13] instead, as suggested by [34]. Both versions produced nearly identical results [33]. It is composed of a stack of identical layers. Each layer has two sub-layers. The first layer is a multi-head self-attention mechanism, while the second one is a simple, position-wise fully connected feed-forward network. A residual connection is around each of the two sub-layers, followed by layer normalization. We decided to add a stack of Transformer layers due to the assumption that additional hyperparameters can increase the ability of the architecture to better model long-range contexts and alleviate the number of spurious predicted entities [4]. The output of the RoBERTa encoder consists of the final hidden state vector of [*CLS*] as the representations of the whole sequence and the sequential token representation. We apply the stack of Transformer blocks on this sequence and the output is concatenated with the [CLS] representation which afterwards is fed into two output layers for classification: a *sigmoid* dense layer for *information type* classification (with a 0.5 cut-off) and a *softmax* dense layer for *priority* classification.

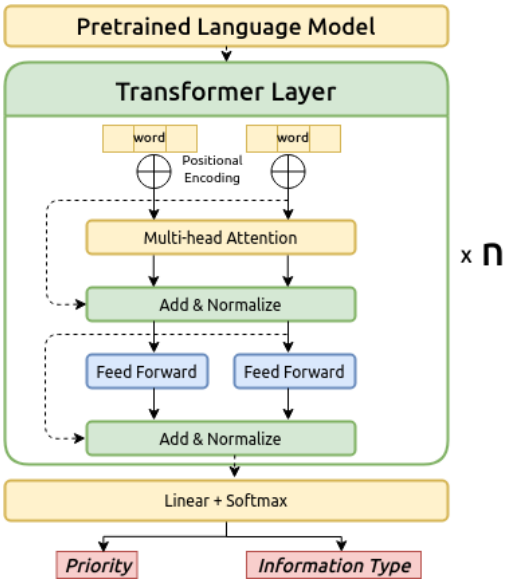

Figure 3: Detailed model for multitask event detection [4, 6].

4.1 Marking Entities and Highlighting Hashtags

Different from previous approaches that primarily focused on the tweet text, we, additionally, took advantage of topic-related information (as in the topic example in Figure 2).

Hashtags. First, since hashtags are one of the main reasons for which Twitter trends emerge rapidly around the day's news, events that unexpectedly gain viral traction, we explore their representation. A hashtag is a combination of keywords or phrases preceded by the # symbol, excluding any spaces or punctuation. Thus, to exploit them, we separate the keywords with a simple rule that tokenizes at the encounter of an uppercase letter (e.g., *#BloodDonation* becomes *# Blood Donation*).

Entities. Second, given that any kind of information can be posted and shared, it is possible to raise the importance of tweets that are related to a person, a product, an organization, or any other entity of interest. These entities can be extremely relevant when aid is being needed in specific locations, specific time frames, etc. For detecting the entity mentions in a document, we used an out-of-the-box already trained model provided by spaCy v3.0+[2] [17]. spaCy features a statistical entity recognition system with default trained pipelines that can identify a variety of named and numeric entities, including companies, locations, organizations, events, and products.

EntityMarkers. For exploiting the hashtags and the entities, we implemented the pre-trained language model with *EntityMarkers* [5, 26, 27, 32]. First, our model extends the RoBERTa [20] model applied to text classification and we add two dense linear layers with softmax activation for the separate tasks: *information type* and *priority*. Next, we augment the input tweet with a series of special tokens. For example, the tweet in Figure 1 becomes enhanced with extra information as in Figure 4.

[2] https://spacy.io/

Events. Additionally, we concatenate the augmented tweet text with the event title and type found in the topic description (e.g., *explosion* and *Beirut Explosion*).

> *<#> <LOC> Beirut </LOC> </#> <#> Blood Donors </#> All blood types are needed across all hospitals in <LOC> Lebanon </LOC> right now - PLEASE donate if you can. Stay safe..*

Figure 4: High priority tweet enhanced with hashtag and entity markers.

4.2 Other Priority Classification Methods

We also explore the bag of words representation (term frequency - inverse document frequency (TF-IDF) weighting) and two classical machine learning approaches for predicting the *priority*.

4.2.1 Support vector machine (SVM) with mixed word n-grams. Support Vector Machine (SVM) is one of the most practiced machine learning methods for classifying text due to the fact that it is independent of the dimension space of the feature [15]. We used TF-IDF weighting since due to the writing style used on Twitter, social media languages would likely be less informative while comprising a large portion of high-frequency terms, thus should be penalized. We experimented with unigrams, bigrams, and both representations. Our results showed that the combination is the best representation for the SVM classifier, as unigrams by themselves hardly seemed meaningful while it is suspected that they can serve as complements to bigrams. We report one internal SVM-based run as (b) in Table 2.

4.2.2 Logistic regression (LR) with word and character n-grams. Words are often thought to be the traditional "grain", observable for feeding texts into classifying algorithms. Subwords models are often used to overcome the limits of generalization when token forms vary too much [18, 19] or when lemmatization or stemming remains difficult in some languages [14, 28]. Here, we do not consider words but character strings. Since words are just a subset of all character strings of a text, we believe that, with an appropriate choice of n-gram size, it is possible to get some strong baseline without tokenization or even pre-processing. Our experiments showed that character n-grams with $N = 4$ showed the best results, and we report an internal result as (a) in Table 2. It seemed that shorter n-grams failed to store the sequential aspect of language while longer n-grams failed to generalize since they tend to rather give sparse representations (due to anti-monotonicity[31], a 5-gram is at the best as frequent as the 4-grams it contains). We also observed that TF-IDF weighting has a good impact on the results of this type of representation. Among the various machine learning algorithms we experimented, LR was the best performing one (Section *Other Priority Explore Methods* in Figure 2).

5 EXPERIMENTS

In our internal experimental setup, we produced a train-test split, considering that the test set should contain unknown event types. Thus, for the training set, we selected the event types from 0 to 64,

Table 1: Internal evaluation performance scores for the different proposed approaches.

Approach	nDCG	Info-Type			Priority	
	@100	F1(Act)	F1(All)	Acc	F1(Act)	F1(All)
Transformer-based Multitask Explored Models						
sequence_size=256						
RoBERTa-base	0.5052	0.1139	0.2157	0.8877	0.2566	0.2948
(1) RoBERTa-base + #	0.4854	0.1174	0.1872	0.8813	0.3144	0.3343
(2) RoBERTa-base + Entities	0.5057	0.1045	0.1929	0.8791	0.3041	0.3024
(3) RoBERTa-base + # + Entities	0.4750	0.0705	0.1722	0.8845	0.2426	0.2617
(4) RoBERTa-base + 2×Transformer	0.4990	0.2074	0.2399	0.8768	0.2749	0.3046
(5) RoBERTa-base + # + 2×Transformer	0.5237	0.1488	0.2359	0.8825	0.2762	0.3086
(6) RoBERTa-base + # + 4×Transformer	0.4789	0.1101	0.1720	0.8784	0.2518	0.2798
(7) RoBERTa-base + # + 2×Transformer + EventTitle + EventType	0.5397	0.0908	0.2172	0.8851	0.3624	0.3336
sequence_size=280						
RoBERTa-base	0.5123	0.1741	0.2399	0.8809	0.2824	0.2962
(8) RoBERTa-base + #	0.5052	0.1139	0.2157	0.8877	0.2566	0.2948
(9) RoBERTa-base + 2×Transformer	0.5123	0.1741	0.2399	0.8809	0.2824	0.2962
(10) RoBERTa-base + # + 2×Transformer	0.4835	0.2441	0.2610	0.8783	0.2171	0.2702
(11) RoBERTa-base + 2×Transformer + EventTitle + EventType	0.5124	0.2324	0.2837	0.8838	0.2522	0.2987
(12) RoBERTa-base + # + 2×Transformer + EventTitle + EventType	0.5084	0.1930	0.2292	0.8806	0.3011	0.3172
Other Priority Explored Models						
(a) LogReg_char_pond=tf-idf_N(4, 4)	–	–	–	–	0.3672	0.3132
(b) SVM TF-IDF mixed	–	–	–	–	0.2907	0.2968

and from 65 to 75, for the test set. Table 1 presents a selected set of preliminary ensemble experiments with different hyperparameters. To evaluate the performance of such systems, the following two groups of metrics were proposed by the organizers:

- Information Type: its overall classification accuracy, micro-averaged across events and macro-averaged across information types, its overall F1 score, macro-averaged across all information types and micro-averaged across events; and its F1 score among six actionable information types (*GoodsServices, SearchAndRescue, MovePeople, NewSubEvent, EmergingThreats, ServiceAvailable*);
- Prioritization: its overall prioritization error, micro-averaged across events and macro-averaged across all information types and information priority score correlational performance (Pearson correlation).
- Also, nDCG@100 is reported: a normalized, discounted cumulative gain evaluated across the top 100 tweets, micro-averaged across all test events.

The TREC-IS edition evaluated each participating run across two axes: information type categorization, and information criticality (tweet priority). Table 1 presents our internal evaluation results, with two sets of experiments: one that used a commonly used sequence length of 256, and the other with the maximum tweet tokens, 280. For information type classification, we underlined the first two best results in each set. For prioritization, we underlined the results with a value > 30.

5.1 Submitted Runs

Table 2 compares our submissions with the mean, median, minimum and maximum scores of TREC Incident Streams 2021. The following are our runs that were submitted to this edition of the track:

- (4) + (2): A simple encoder with two additional Transformer layers proved to be efficient regarding the detection of the information type obtaining the highest F1 score on the actionable types (compared to our runs), while entities helped in discerning better between the levels of criticality.
  - *Information Type*: (4) RoBERTa-base + 2×Transformer with a sequence length of 256;
  - *Priority*: (2) RoBERTa-base + E(ntities).
- (5) + (1): The highest F1 score for all information types and F1 for priority detection in actionable types were obtained by augmenting the tweets with our pre-processed hashtags, with and without additional Transformer layers.
  - (5) *Information Type*: RoBERTa-base + # + 2×Transformer with a sequence length of 256;
  - (1) *Priority*: RoBERTa-base + #.
- (10) + (a): For detecting the criticality in tweets, the highest F1 scores were obtained by the logistic regression approach.
  - *Information Type*: (10) RoBERTa-base + 2× Transformer + # with a sequence length of 280;
  - *Priority*: (a) LogReg_char_pond=tf-idf_N(4, 4).
- (11) + (b): This run is characterized by the lowest priority scores which disproves the efficiency of the SVMs in detecting critical tweets.
  - *Information Type*: RoBERTa-base + 2×Transformer + event type + event title with a sequence length of 280;
  - *Priority*: SVM TF-IDF mixed.

Table 2: Evaluation performance scores at TREC Incident Streams 2021.

Runs	nDCG	Info-Type			Priority			
	@100	F1(Act)	F1(All)	Acc	F1(Act)	F1(All)	R(Act)	R(All)
Our Participating Runs								
(4) + (2)	0.5690	**0.2155**	0.2724	0.8896	0.1849	0.2175	0.1862	0.2716
(5) + (1)	0.6050	0.2060	**0.2901**	0.8884	0.2113	**0.2573**	0.3068	0.3143
(10) + (a)	0.3590	0.1948	0.2682	**0.8892**	**0.2250**	0.2478	0.1168	0.1141
(11) + (b)	0.2885	0.1715	0.2804	0.8873	0.0400	0.0967	-0.0177	0.0017
(10) + (12)	0.5955	0.1948	0.2682	**0.8892**	0.2160	0.2429	**0.4349**	**0.3585**
(11) + (12)	0.5963	0.1715	0.2804	0.8873	0.2160	0.2429	**0.4349**	**0.3585**
TREC Incident Streams Track 2021 Results								
Mean	0.5186	0.1906	0.2625	0.8567	0.1832	0.2061	0.1764	0.2003
Median	0.5692	0.2066	0.2823	0.8834	0.1923	0.1754	0.1689	0.2099
Min	0.2885	0.0000	0.0064	0.4999	0.0204	0.0810	-0.0667	0.0017
Max	0.6115	0.2815	0.3211	0.8902	0.3052	0.3125	**0.4349**	**0.3585**

- (10) + (12): The encoders with additional Transformer layers and a sequence length of 280 augmented with hashtags obtained the highest information type accuracy.
  - *Information Type*: RoBERTa-base + 2×Transformer + # with a sequence length of 280;
  - *Priority*: RoBERTa-base + 2×Transformer + event type + event title + # with a sequence length of 280.
- (11) + (12): The Pearson correlation is more of interest for the prioritization task, and the highest values were obtained with our proposed multitask approach with event information and augmented hashtags, encoders with additional Transformer layers and a sequence length of 280 augmented with hashtags.
  - *Information Type*: RoBERTa-base + 2×Transformer + event type + event title with a sequence length of 280;
  - *Priority*: RoBERTa-base + 2×Transformer + event type + event title + # with a sequence length of 280.

When comparing Table 1 and Table 2, we first observe that our internal results have the same distribution as the official TREC results. For the prediction of the information type, the F1 scores for all information types, as well as for the actionable ones, are generally lower when no additional Transformer layer is used, and the highest when over two layers are added. Also, when comparing the importance of entities, hashtags, or events, we notice that augmenting the tweets with the event title and type outperforms the models that use entities. Moreover, the pre-processing of the hashtag information further increases the performance when applied together with several Transformer layers. When detecting the priority types, we observe the same tendencies when it comes to the number of Transformer layers. However, higher scores were observed when entities and hashtags were added to the base model also. Regarding the maximum length of the tweets, a marginal increase in performance was observed when using the maximum length of a tweet (280), nonetheless, they were not statistically significant.

Finally, it is clear, from our internal experimental setup, and then from the final TREC results, that augmenting the text with entities and hashtags can bring an important increase in the detection of tweet criticality, and can be further improved when the event type and title are known.

6 CONCLUSIONS AND FUTURE WORK

For the participation of our team (L3i) in the TREC Incident Streams 2021, we proposed two approaches, a multitask information type and a prioritization Transformer-based model with entities, hashtags, and event types, and a classical machine learning-based method for the prioritization task. We conclude that taking advantage of learning both tasks with highlighted features (e.g., entities, hashtags) outperforms all our other methods in our experimental setup. The machine learning methods obtained high scores. However, they could not compete with the Transformer-based methods. Finally, our submissions obtained top performance for the prioritization task and scores above the median for the information type classification task.

ACKNOWLEDGMENTS

This work has been supported by the European Union's Horizon 2020 research and innovation program under grants 770299 (NewsEye) and 825153 (Embeddia), and by the ANNA and Termitrad projects funded by the Nouvelle-Aquitaine Region.

REFERENCES

[1] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.

[2] Emanuela Boros and Antoine Doucet. 2021. Transformer-based Methods for Recognizing Ultra Fine-grained Entities (RUFES). *arXiv preprint arXiv:2104.06048* (2021).

[3] Emanuela Boros, Ahmed Hamdi, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Jose Moreno, Nicolas Sidere, and Antoine Doucet. 2021. Atténuer les erreurs de numérisation dans la reconnaissance d'entités nommées pour les documents historiques. In *COnférence en Recherche d'Informations et Applications-CORIA 2021, French Information Retrieval Conference,*.

[4] Emanuela Boros, Ahmed Hamdi, Elvys Linhares Pontes, Luis-Adrián Cabrera-Diego, José G Moreno, Nicolas Sidere, and Antoine Doucet. 2020. Alleviating digitization errors in named entity recognition for historical documents. In *Proceedings of the 24th Conference on Computational Natural Language Learning*. 431–441.

[5] Emanuela Boros, José G. Moreno, and Antoine Doucet. 2021. Event Detection with Entity Markers. In *Advances in Information Retrieval*, Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani (Eds.). Springer International Publishing, Cham, 233–240.

[6] Emanuela Boros, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Ahmed Hamdi, José Moreno, Nicolas Sidère, and Antoine Doucet. 2020. Robust named entity recognition and linking on historical multilingual documents. In *Conference and Labs of the Evaluation Forum (CLEF 2020)*, Vol. 2696. CEUR-WS Working Notes, 1–17.

[7] Cody Buntain, Richard McCreadie, and Ian Soboroff. 2020. Incident Streams 2020: TRECIS in the Time of COVID-19. (2020).

[8] Won-Gyu Choi, Seung-Hyeon Jo, and Kyung-Soon Lee. 2018. CBNU at TREC 2018 Incident Streams Track.. In *TREC*.

[9] Abu Nowshed Chy, Umme Aymun Siddiqua, and Masaki Aono. 2018. Neural Networks and Support Vector Machine based Approach for Classifying Tweets by Information Types at TREC 2018 Incident Streams Task.. In *TREC*.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186.

[11] Alexis Dusart, Gilles Hubert, and Karen Pinel-Sauvagnat. 2019. Irit at trec 2019: Incident streams and complex answer retrieval tracks. In *Text REtrieval Conference*. National Institute of standards and Technology (NIST).

[12] Miguel Ángel García-Cumbreras, Manuel Carlos Díaz-Galiano, Manuel García-Vega, and Salud María Jiménez-Zafra. 2018. SINAI at TREC 2018: Experiments in incident streams. *Weather* 38, 3 (2018), 14.

[13] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122* (2017).

[14] Dhaou Ghoul and Gaël Lejeune. 2019. MICHAEL: Mining Character-level Patterns for Arabic Dialect Identification (MADAR Challenge). In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*. Association for Computational Linguistics, Florence, France, 229–233. https://doi.org/10.18653/v1/W19-4627

[15] B S Harish, Devanur Guru, and Manjunath Shantharamu. 2010. Representation and Classification of Text Documents: A Brief Review. *International Journal of Computer Applications,Special Issue on RTIPPR* 1 (01 2010), 110 – 119.

[16] Alexander J Hepburn and Richard McCreadie. 2020. University of Glasgow Terrier Team (uogTr) at the TREC 2020 Incident Streams Track. *Image* 8 (2020), 5.

[17] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *Unpublished software application* (2017). https://spacy.io

[18] Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226* (2018).

[19] Gaël Lejeune, Romain Brixtel, Antoine Doucet, and Nadine Lucas. 2015. Multilingual event extraction for epidemic detection. *Artificial Intelligence in Medicine* (aug 2015). https://doi.org/10.1016/j.artmed.2015.06.005 doi: 10.1016/j.artmed.2015.06.005.

[20] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[21] Richard McCreadie, Cody Buntain, and Ian Soboroff. 2019. Trec incident streams: Finding actionable information on social media. (2019).

[22] Richard McCreadie, Cody Buntain, and Ian Soboroff. 2020. Incident Streams 2019: Actionable Insights and How to Find Them. In *Proceedings of the International ISCRAM Conference*.

[23] Akanksha Mishra and Sukomal Pal. 2019. IIT BHU at TREC 2019 Incident Streams Track.. In *TREC*.

[24] Taro Miyazaki, Kiminobu Makino, Yuka Takei, Hiroki Okamoto, and Jun Goto. 2018. NHK STRL at TREC 2018 Incident Streams track.. In *TREC*.

[25] Taro Miyazaki, Kiminobu Makino, Yuka Takei, Hiroki Okamoto, and Jun Goto. 2019. Label embedding using hierarchical structure of labels for twitter classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 6317–6322.

[26] José G Moreno, Emanuela Boros, and Antoine Doucet. 2020. TLR at the NTCIR-15 FinNum-2 Task: Improving Text Classifiers for Numeral Attachment in Financial Social Data. In *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo Japan*. 8–11.

[27] José G Moreno, Antoine Doucet, and Brigitte Grau. 2021. Relation Classification via Relation Validation. In *Proceedings of the 6th Workshop on Semantic Deep Learning (SemDeep-6)*. 20–27.

[28] Kyubyong Park, Joohong Lee, Seongbo Jang, and Dawoon Jung. 2020. An Empirical Study of Tokenization Strategies for Various Korean NLP Tasks. *arXiv preprint arXiv:2010.02534* (2020).

[29] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global Vectors for Word Representation.. In *EMNLP*, Vol. 14. 1532–1543.

[30] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018)*. Association for Computational Linguistics, New Orleans, Louisiana, 2227–2237.

[31] Luc De Raedt and Albrecht Zimmermann. 2007. Constraint-based pattern set mining. In *proceedings of the 2007 SIAM International conference on Data Mining*. SIAM, 237–248.

[32] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. *arXiv preprint arXiv:1906.03158* (2019).

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

[34] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv* abs/1910.03771 (2019).