Leveraging Transformer Self Attention Encoder for Crisis Event Detection in Short Texts

Pantelis Kyriakidis, Despoina Chatzakou, Theodora Tsikrika
Stefanos Vrochidis, Ioannis Kompatsiaris

Information Technologies Institute, Centre for Research and Technology Hellas
{pantelisk,dchatzakou,theodora.tsikrika,stefanos,ikom}@iti.gr

Abstract. Analyzing content generated on social media has proven to be a powerful tool for early detection of crisis-related events. Such an analysis may allow for timely action, mitigating or even preventing altogether the effects of a crisis. However, the high noise levels in short texts present in microblogging platforms, combined with the limited publicly available datasets have rendered the task difficult. Here, we propose deep learning models based on a transformer self-attention encoder, which is capable of detecting event-related parts in a text, while also minimizing potential noise levels. Our models efficacy is shown by experimenting with CrisisLexT26, achieving up to 81.6% f1-score and 92.7% AUC.

Keywords: Self attention, Multihead attention, Crisis event detection

1 Introduction

Over the years, many methods have been introduced in an effort to effectively detect crisis events from online textual content, and thus keep relevant stakeholders and the community at large informed about these events and their aftermath. Such crisis event detection methods were initially utilizing handcrafted feature engineering to enrich their models with semantic and linguistic knowledge [17, 23, 22]. Nevertheless, the handcrafted features are not able to capture the multilevel correlations between words and tend to be overspecialized to the designed domain (e.g. applied on a single language), time consuming, and prone to error propagation. In the last decade, though, Convolutional Neural Networks (CNNs) [21] and Recurrent Neural Networks (RNNs) [16] proved their ability to capture semantic information [19, 33] and, therefore, pioneering works introduced CNNs [7, 31], RNNs [30], and even hybrid approaches [12] to the event detection task. Graph Convolutional Networks (GCNs) which enable the convolution of words that are dependent on each other by using the syntactic representation of a text have also been employed [29]. Finally, attention mechanisms have been applied, either in combination with GCNs [35] or in a more straightforward way [25], to combine attention vectors for words and entities with the original word embeddings before being forwarded to a perceptron layer.

However all these studies focused on the sentence level of large and well organised documents by identifying event trigger words. Another research area for

event detection has been the categorization of short social media posts by their informativeness during a crisis. A publicly available dataset widely used in this context is CrisisLexT26 [32], which contains Twitter posts from 26 different crisis events. CNNs have been extensively studied also for this task (e.g. [6], [28], [5]), with one of the most well known CNN architectures for sentence classification, the Multi-Channel CNN architecture [19], achieving noteworthy results [4].

Although a lot of effort has been put into the event detection task, it appears that no special emphasis has been placed on how to effectively deal with short noisy (informal language, syntactic errors, unordered summarizations etc.) texts, which limits the effectiveness of the currently available methods. For instance, only a limited number of studies have used attention mechanisms towards this direction [18]. To this end, this paper proposes a new way of dealing with short and high noise texts in the context of the crisis-related event detection. The proposed method is based on a state-of-the-art self attention encoder [34], which is utilized as a denoiser for the text before this is further forwarded to other types of layers; to the best of our knowledge, there is no prior work that investigates the effect of the self-attention encoder in the event detection task. We also designed and implemented three models modifying to some extent the way in which the self attention is utilized, while also experimenting with different neural network architectures. Moreover, to be comparable with the state of the art, the Multi-Channel CNN architecture [4] is used as baseline, as well as a variation thereof. Finally, we open-source our implementations for reproducibility and extensibility purposes [1].

2 Methodology

To handle short and noisy texts, we base our models on a self attention method. We assume that the attention will be immune to any temporal inconsistency in the text and be able to reinforce dependencies between relevant words. Before presenting the proposed architectures, we first describe their key components.

Language Modeling. To create a mathematical representation of the input words we use an embedding of the language into a high dimensional euclidean space where neighboring words are also semantically close; input texts are modeled in a D-dimensional vector based on the words' vectors representations. We chose a 300-dimensional Word2Vec model [27] (pretrained on Google news) so as to be comparable with the chosen baseline [4]. Furthermore, to give sequential information to each word embedding, other works (e.g. [34]) add positional encoding to the embedding vector [14]. However, we argue that in short texts (e.g. Twitter posts), positional information would be unnecessary in the embedding layer, as these texts tend to be very unorganized and high in noise.

Self Attention Encoder. Overall, the Self Attention Encoder [34] consists of a block of 2 sub-layers. The first is a *multi-head self attention* and the second is a *position-wise feed-forward* layer, i.e. a fully connected layer with shared parameters over the sequence, applied in each position. Each layer output is

added with a residual connection [15] from the previous layer output followed by Layer Normalization [3] aiming to a more stable and better regularized network.

Intuitively, the attention vector of a query word sums to 1 and spans the "attention" of the query to the most important words (keys); we use as queries (Q), as keys (K), and as values (V) the same input sequence, where every word is a query, a key, and a value, and their dimension is d. The attention mechanism that is applied at the core of each head is a scaled dot product attention (SDPA) which is a normalized version of the simple dot product attention and adheres to the following equations: (1): $E = \frac{(QK^T)}{\sqrt{d}}$; the attention scores are calculated, (2): $A = softmax(E)$; the attention is distributed to every key based on the attention scores, and (3): $C = AV$; the output context vector is calculated as an attention weighted sum of all values.

Multi-head Attention. This is the layer where the attentions are calculated. In order to attend in more than one ways, queries, keys, and values are projected h times through learned projection matrices, where h is the number of attention heads. These h tuples of Q, K, V are forwarded to each head's SDPA. Outputs are concatenated, and once again projected through a learned weight matrix. In the original work, these projections are linear [34]. However, we found that in our case non-linear projection with the Rectifier Linear Unit (ReLU) as an activation function boosted the performance.

GRU and CNN. GRUs are well known neural architectures capable of capturing sequential information [8]. Specifically, an update and reset gate is used to decide what information should be passed to the output. CNNs [21] capture salient features from chunks of information where each chunk is an n-gram of words from the input text where the convolution operation is performed on.

2.1 Proposed Neural Network Architectures

In this section, we introduce the overall design of the three proposed architectures. All models are optimized with ADAM optimizer [20], learning rate of 0.001, and with dropout of 0.5 before the output, for regularization purposes.

Stacked-Self Attention Encoders (Stacked-SAE). The first proposed architecture consists of a stack of 4 (experimentally chosen) self-attention encoders. It is the deepest of the architectures and the most complex one. The output is aggregated with Global Average Pooling [24] and finally projected onto the output layer, i.e. a fully connected layer with softmax as activation function. We expect that a deeper architecture will be able to create better representations and capture more complex patterns.

Attention Denoised Parallel GRUs (AD-PGRU). This and the next architecture are using only one self-attention encoder as a feature extraction mechanism. As a result, the attention weighted output added to the input with the residual connection is expected to reinforce the important words eliminating significant part of the noise. Afterwards and inspired by [19], the signal is passed to a Parallel GRU architecture composed of 3 units that reduces the sequence to

a single vector in the end and concatenates all before forwarding to the output. Each unit is expected to learn a different sequential representation of the input.

Attention Denoised Multi-channel CNN (AD-MCNN). The final architecture uses the same denoiser as the previous one, but instead of a parallel GRU design, we followed the architecture proposed in [19] for sentence classification, i.e. use of three parallel CNN layers operating under different kernel sizes, so as to capture different N-gram combinations from the text, and a max-over-time pooling operation [9]. As for the parameterization, we followed the settings proposed in [4], where the aforementioned architecture was repurposed for the event detection task (experimenting with the CrisisLexT26 dataset).

3 Dataset and Experimental Setup

Ground Truth. For experimentation purposes, the CrisisLexT26 [32] dataset is used; it is publicly available and widely used in related work. It contains 26 different crisis events from 2012 and 2013 and consists of ≈28k tweets (≈1k posts per event). The labels for each tweet concern its: (i) *Informativeness*, whether it is related to the specific crisis or not, (ii) *Information Source*, e.g. government and NGO, and (iii) *Information Type*, e.g. affected individuals.

Text Analysis. Our analysis of the dataset indicates its noisiness as follows. Each tweet contains on average 1.1 hashtags (normalized by the size of the tweet) and 37.4% of all tweets have hashtags in the middle of their text. Similarly, tweets contain on average 0.56 links and 0.81 tags, while the appearances of both in the middle of the text are 15.6% and 55.6%, respectively. A final observation would be that of punctuation, which if used informally may disturb the flow of the sentence, especially when it indicates a more complex sentiment like shock or irony. We found that 11% of the tweets in the dataset include *"."* or *"..."*; likewise one or more —in sequence— *"!"* and *"?"* are found in the middle of the text 8% and 4.5% of the times, respectively.

Experimental Setups. Two experimental setups were designed: (i) *Binary classification:* the focus is on the *Informativeness* category and the objective is to detect the relatedness of a post to a crisis event; and (ii) *Multi-class classification:* with the focus being on the *Information Type* category.

Since the class distribution is highly imbalanced we decided to train our models both for imbalanced and balanced dataset setups. For the balanced setup, we performed oversampling of the minority class using the pretrained BERT model [11] tailored to the Masked Language Model task [2].

We divided the dataset in a stratified fashion with a 0.8-0.2 train-test split, and derived a 10 run average for each experiment creating stochasticity by alternating the global network seed, while all the other randomized parameters, such as the dataset split seed are constant. This is to ensure that all randomness comes from network weights initialization.

Table 1. Experimental results. ("*": statistically significant over baselines)

Binary Classification

	Imbalanced				*Balanced*			
	Precision	Recall	F1-score	AUC	Precision	Recall	F1-score	AUC
MCNN	**0.841**	0.772	0.800	0.921	0.798	0.793	0.793	0.918
MCNN-MA	0.772	0.771	0.770	0.888	0.691	0.768	0.711	0.871
Stacked-SAE	0.821	0.799*	0.808*	0.915	**0.809**	0.784	0.793	0.910
AD-PGRU	0.835	0.799*	0.814*	**0.927***	0.808	0.802	**0.803***	0.921
AD-MCNN	0.834	**0.802***	**0.816***	0.925*	0.804	**0.805**	0.802*	**0.923***

Multiclass Classification

	Imbalanced				*Balanced*			
	Precision	Recall	F1-score	AUC	Precision	Recall	F1-score	AUC
MCNN	**0.671**	0.627	0.640	0.913	0.624	**0.648**	0.632	**0.910**
MCNN-MA	0.616	0.589	0.598	0.883	0.561	0.577	0.563	0.873
Stacked-SAE	0.644	0.630*	0.633	0.906	0.622	0.636	0.626	0.900
AD-PGRU	0.648	0.637*	0.638	0.910	**0.627**	0.640	0.630	0.909
AD-MCNN	0.656	**0.644***	**0.647***	**0.914**	**0.627**	**0.648**	**0.633**	**0.910**

4 Experimental Results

Next, we evaluate the performance of the proposed architectures and compare the results to the baseline. We implemented two baseline architectures: (i) **Multi-channel CNN (MCNN)** [4], which has shown the best performance so far in the CrisisLexT26 dataset; and (ii) **MCNN-MA**: a recently proposed architecture for Sentiment Analysis [13] that uses MCNN and multi-head attention afterwards, which we adapted to our task with suitable parameterization.

As for the **binary classification** (Table 1), we observe that the proposed architectures outperform the baselines in terms of F1 and AUC, both for the balanced and imbalanced datasets, with AD-MCNN being arguably the best performing. This confirms the hypothesis that using the self-attention encoder as a denoiser has a positive impact on the overall detection performance. When MCNN is used as a feature extractor on the embeddings, the convolution window limits the interactions to only neighboring words. So if attention is placed after the CNNs (MCNN-MA), the original signal is altered. On the contrary, the use of attention before MCNN resolves this issue, since it is not restricted neither by the distance between words, nor by the words located between them as an RNN would be. Class-specific results (not reported due to lack of space) on the imbalanced setup, indicate a substantial improvement on the minority class' recall (≈ 0.1 increase) for AD-MCNN. MCNN shows better —though not statistically significant— precision because it lacks the ability to effectively separate the two classes (predicts fewer non-events); a problem which is somewhat addressed in the balanced setup, where the precision is stabilized. As for the **multiclass classification** (Table 1) the overall results follow a similar behavior with the binary one. Although we see comparable performance on the balanced setup, we argue that the AD-MCNN would widen the difference with MCNN if more samples per class were available (on average $\approx 2,8k$ original samples).

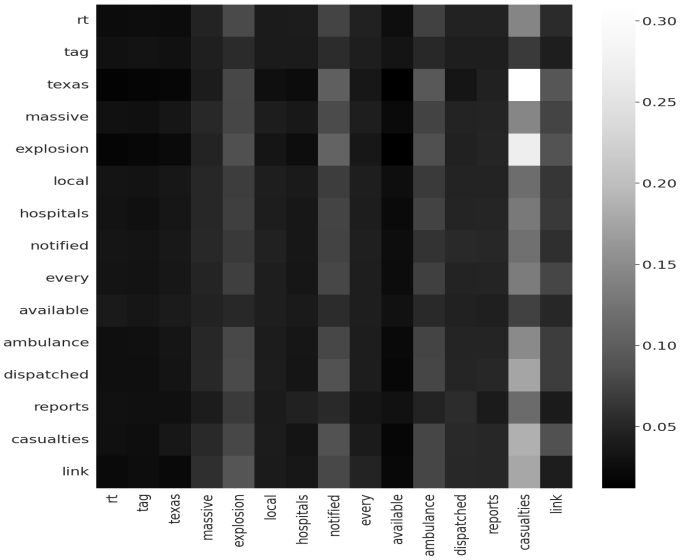

Fig. 1. Attention visualization. The highest scores are observed in the combinations of location ("texas") and type of incident ("explosion") with a very important consequence "casualties".

In Section 1 we claimed that our attention mechanism is utilized as a denoiser for the text before it is further forwarded to other types of layers. To support our argument we provide an illustrative example of the attention scoring of a tweet: *"RT @user: texas: massive explosion u/d - local hospitals notified. every available ambulance dispatched. reports of casualties. HTTP: ..."*. Figure 1 depicts the scores of a single (due to space limitation) attention head, when applied to the above example and indicates that combinations of highly relevant words are being matched up with higher scores, while non-important combinations exhibit low attention scores, resulting in the claimed denoising behavior.

Finally, in Section 2 we argued that positional encoding might not be fit for this specific domain given the informal text used in Twitter, the brevity of the messages, the unordered use of hashtags, etc. Using the AD-MCNN method with the use of positional encoding, we validated this hypothesis, observing 0.01 and 0.004 performance decrease in F1-score and AUC, respectively.

5 Conclusions and Future Work

This study focused on the development of three effective models for the detection of crisis-related events while combatting the inherent noise present in short social media posts. Our hypothesis was that self attention would act as a denoiser, thus reducing unimportant features by paying attention to the ones that matter most, in a way that every vector in the sequence is enhanced with context from other directly related word vectors. We validated our hypothesis by building and evaluating three attention enhanced models that improved performance

against strong baselines. In the future, we intend to further evaluate our models, especially when more data is available. We also plan to evaluate the impact of more recent language models for word embeddings (e.g. [11], [26]), especially multilingual ones [10].

Acknowledgements

This research has received funding from the European Union's H2020 research and innovation programme as part of the INFINITY (GA No 883293) and AIDA (GA No 883596) projects.

References

1. Crisis-event-detection-in-short-texts - implementation, https://github.com/M4D-MKLab-ITI/Crisis-Event-Detection-in-Short-Texts
2. Bert for masked lm (2020), shorturl.at/drRV4
3. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. stat **1050**, 21 (2016)
4. Burel, G., Alani, H.: Crisis event extraction service (crees)-automatic detection and classification of crisis-related content on social media (2018)
5. Burel, G., Saif, H., Alani, H.: Semantic wide and deep learning for detecting crisis-information categories on social media. In: International semantic web conference. pp. 138–155. Springer (2017)
6. Caragea, C., Silvescu, A., Tapia, A.H.: Identifying informative messages in disaster events using convolutional neural networks. In: International conference on information systems for crisis response and management. pp. 137–147 (2016)
7. Chen, Y., Xu, L., Liu, K., Zeng, D., Zhao, J.: 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). pp. 167–176 (2015)
8. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
9. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. Journal of machine learning research **12**(ARTICLE), 2493–2537 (2011)
10. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451 (2020)
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186 (2019)
12. Feng, X., Qin, B., Liu, T.: A language-independent neural network for event detection. Science China Information Sciences **61**(9), 1–12 (2018)

13. Feng, Y., Cheng, Y.: Short text sentiment analysis based on multi-channel cnn with multi-head attention mechanism. IEEE Access **9**, 19854–19863 (2021)
14. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning. In: International Conference on Machine Learning. pp. 1243–1252. PMLR (2017)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
16. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
17. Hong, Y., Zhang, J., Ma, B., Yao, J., Zhou, G., Zhu, Q.: Using cross-entity inference to improve event extraction. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. pp. 1127–1136 (2011)
18. Kabir, M.Y., Madria, S.: A deep learning approach for tweet classification and rescue scheduling for effective disaster management. In: Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. pp. 269–278 (2019)
19. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1746–1751. Association for Computational Linguistics, Doha, Qatar (Oct 2014). https://doi.org/10.3115/v1/D14-1181, https://www.aclweb.org/anthology/D14-1181
20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015 (2015)
21. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
22. Li, Q., Ji, H., Hong, Y., Li, S.: Constructing information networks using one single model. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1846–1851 (2014)
23. Li, Q., Ji, H., Huang, L.: Joint event extraction via structured prediction with global features. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 73–82 (2013)
24. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)
25. Liu, S., Chen, Y., Liu, K., Zhao, J.: Exploiting argument information to improve event detection via supervised attention mechanisms. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1789–1798 (2017)
26. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
27. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
28. Nguyen, D., Al Mannai, K.A., Joty, S., Sajjad, H., Imran, M., Mitra, P.: Robust classification of crisis-related data on social networks using convolutional neural networks. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 11 (2017)
29. Nguyen, T., Grishman, R.: Graph convolutional networks with argument-aware pooling for event detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)

30. Nguyen, T.H., Cho, K., Grishman, R.: Joint event extraction via recurrent neural networks. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 300–309 (2016)
31. Nguyen, T.H., Grishman, R.: Event detection and domain adaptation with 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 2: Short Papers). pp. 365–371 (2015)
32. Olteanu, A., Vieweg, S., Castillo, C.: What to expect when the unexpected happens: Social media communications across crises. Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work Social Computing (2015)
33. Tai, K.S., Socher, R., Manning, C.D.: Improved semantic representations from tree-structured long short-term memory 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). pp. 1556–1566 (2015)
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 6000–6010 (2017)
35. Yan, H., Jin, X., Meng, X., Guo, J., Cheng, X.: Event detection with multi-order graph convolution and aggregated attention. 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). pp. 5770–5774 (2019)