LEVERAGING TRANSFORMER SELF ATTENTION ENCODER FOR CRISIS EVENT DETECTION IN SHORT TEXTS

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

CERTH
CENTRE FOR RESEARCH & TECHNOLOGY HELLAS

CENTRE FOR RESEARCH & TECHNOLOGY - HELLAS
Information Technologies Institute

CONTACT INFO
pantelisk@iti.gr

ACKNOWLEDGMENTS

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.

INTRODUCTION

- Analysis of social media content for early detection of crisis-related events
  - Timely action
  - Mitigation/prevention of the effects of a crisis
- Problem:
  - High noise levels in short texts present in social media posts
  - Limited publicly available datasets
- The current SOTA on the task (MCNN)
  - Cannot extract effective features for correlated words that are far apart in the sentence
  - Includes a lot of noisy words in the convolutions

Contributions

- Use of Transformer self-attention encoders
  - Detection of event-related parts in a text
  - Minimization of potential noise levels
- Up to 81.6% f1-score and 92.7% AUC on CrisisLexT26 dataset

Hypothesis

Attention will be immune to any temporal inconsistency and distance between related words and be able to reinforce useful correlations.

METHODOLOGY

Language Modeling (LM)

- Word2Vec model, 300 dims/word
- Pretrained on Google news

Positional Encoding: It is used by the Transformer model to compensate for the lack of sequential modeling. However, we argue that for short texts, it would be unnecessary, as these texts tend to be very brief and unorganized.

Self-Attention Encoder: We employ the SOTA attention encoding method from Transformers as a feature extractor after LM. The attention is performed in the Multi-Head Attention Layer.

- Transformations of input for each head is performed non linearly

Baselines

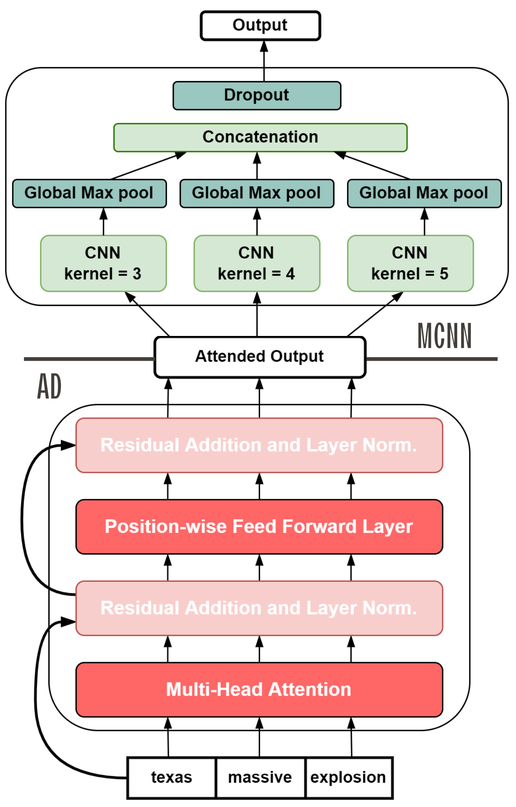
- *Multi-channel CNN (MCNN)*: 3 parallel CNN layers operating with different kernel sizes, so as to capture different N-gram combinations from the text, and a max-over-time pooling operation
- *MCNN-MA*: MCNN, followed by Multihead-Attention proposed for Sentiment Analysis

Proposed neural architectures

- *Stacked Self Attention Encoders (Stacked-SAE): 4 SAE stacked followed by Global Average Pooling*
- *Attention Denoised Parallel GRUs (AD-PGRU): 1 SAE followed by 3 parallel GRUs learning different sequential representations of the input*
- *Attention Denoised Multi-channel CNN (AD-MCNN): 1 SAE followed by MCNN*

***Data Augmentation (oversampling)**: Use of a pretrained BERT model tailored to the Masked Language Model task.*

ARCHITECTURE



AD-MCNN

DATA

Dataset: CrisisLexT26

- ~28k Twitter posts
- 26 Crisis events 2012-2013
- ~1k posts per event
- Labels:
  - *Informativeness related/unrelated*
  - *Information source* e.g. NGO, government
  - *Information type* e.g. affected individuals

Experimental setup

- Binary classification
  - Tweet informativeness
- Multiclass Classification
  - Information type (7 types)
- Imbalanced/Balanced
  - 88-12 % for binary classification
  - Balanced setup: minority class augmentation

Experiment details

- Train-test split: 0.8-0.2
- 10 run average (alternating network seed)

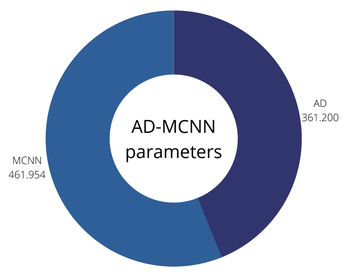
RESULTS

Binary Classification

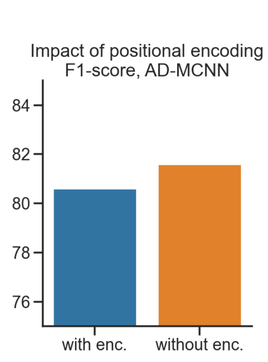
- The proposed architectures outperform the baselines, especially in terms of Recall
- MCNN underperforms when recalling minority class examples

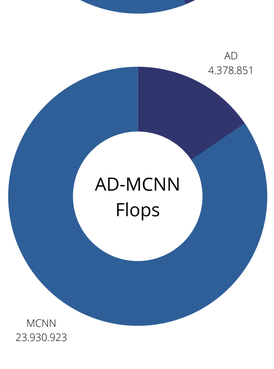
Multiclass Classification

- The overall results follow a similar behavior to the binary one, however, they are less pronounced (less original samples per class ~2.8k).



- Positional encoding lowers the performance of the model
  - Noisy, unorganized form of text.
- Regarding complexity, although Attention Denoiser (AD) adds a lot of parameters to the model size it has less effect to the overall operations performed.
- Each attention head is computed in parallel, contrary to the slow sequential computation of RNNs.

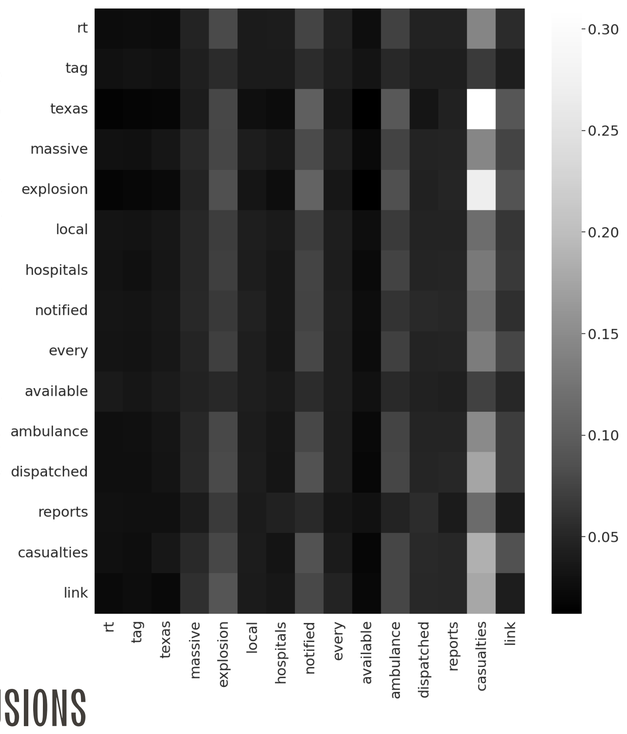


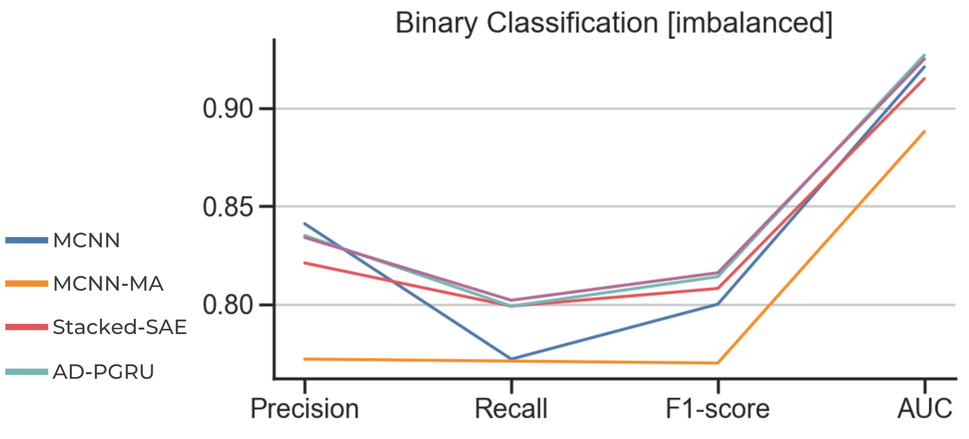


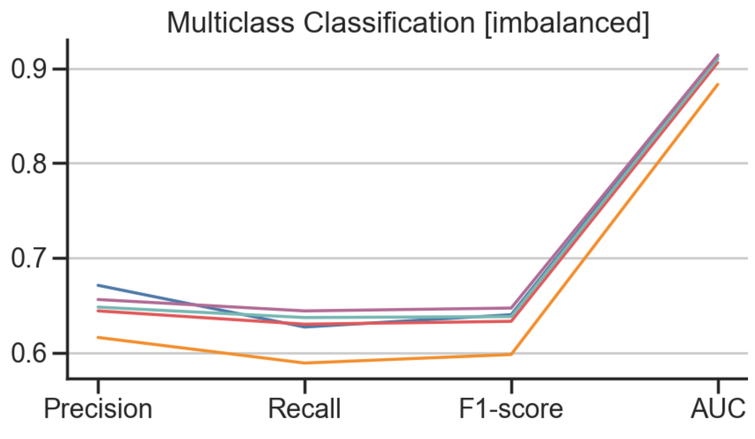
CASE STUDY

Attention head heatmap

- The highest scores are observed in the combinations of location ("texas") and type of incident ("explosion") with their consequences, "casualties".
- Attention acts as a denoiser for the text.
- Relevant words are being matched up with higher scores, while non-important combinations exhibit low attention scores, resulting in the claimed denoising behavior.







CONCLUSIONS

- AD before MCNN is neither restricted to word n-grams (CNN) nor by the distance between words (RNN) and thus performs better as a feature extractor.
- Positional encoding might not be useful in short social media texts.
- Multi-head attention seems to perform better when input transformations are performed **non-linearly**.