Towards inherent Transformers explanations

Comparing Transformers characteristics to human explanations based on sentiment analysis task

Hanna Sobocińska, Emilia Kacprzak, Agnieszka Pluwak (SentiOne)

Problem Statement

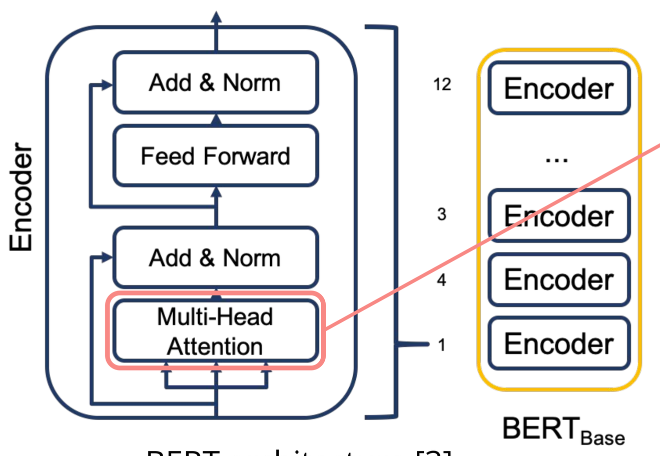
- A good model **explanation**: accurately representing the model's behavior (**fidelity**) and being human-understandable (**comprehensibility**) [1].
- **Question**: Is it feasible to capture both fidelity and comprehensibility using Transformer's inherent attention mechanism?
- **Method**:
  1. A corpus of span- and document-level sentiment annotations was created – the span-level annotations are used as human explanations
  2. An encoder model (**XLM-R** [2]) with a **classification** layer was trained (the model was tested on a sample of 1000 documents and achieved an F1-score of ~0.9)
  3. The attention heads have been extracted from the hidden layers
  4. To capture the explainability potential of the model's hidden features, the **similarities** between various model's **attention heads within layers** and **human explanations** were analyzed.

Dataset

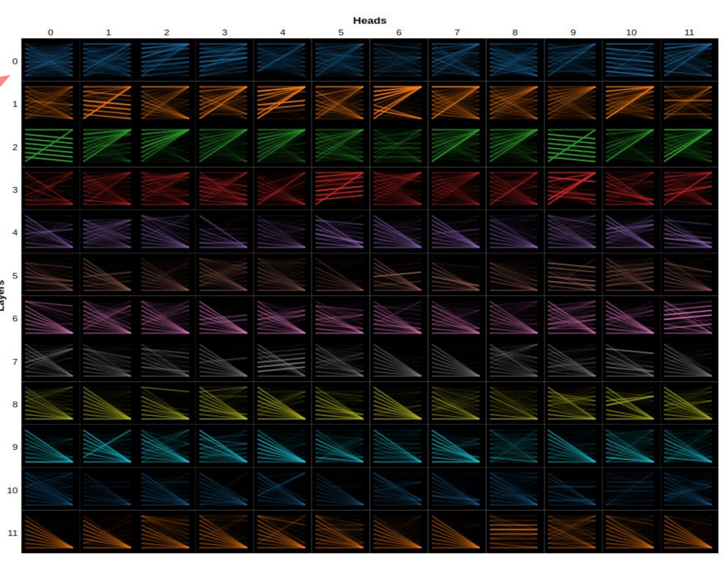
- A large (~30 000 documents), manually annotated by linguistic specialists **sentiment analysis** corpus with annotations on **span-** and **document-level**.
- It contains user-generated content in English, German, Spanish for three domains – utilities, healthcare, and banking.
- **Document** annotations have been used for model's training, **span** annotations have been used as human explanations
- We filtered the dataset to a **binary** scenario (positive-negative) to simplify explanations

Dataset sample

Background – Basic BERT architecture



BERT architecture [3]



Attention Heads Visualization (BertViz) [4]

BERT is based on the multi-head self-attention mechanism which means that each layer consists of multiple attention heads, in which each token is connected to every token in text

Token-level attention definition

Because of the nature of Transformer attention mechanism, it needs to be transformed into token-level attention.

We have used **three** methods for the transformation:

- **Sum** – sum of all incoming attentions (all layers and all heads) for each token (as in [5])
- **Mean-Max** – sum of incoming attentions in single attention head, average attention against layers and choose maximum from each head (as in [6])
- **Last Layer** – sum of all incoming attentions from all heads in the last layer (which, by intuition, should be the most related to the classification layer)

In all methods the values have been normalized to <0,1> range. Attention to commas and full stops have been set to 0.

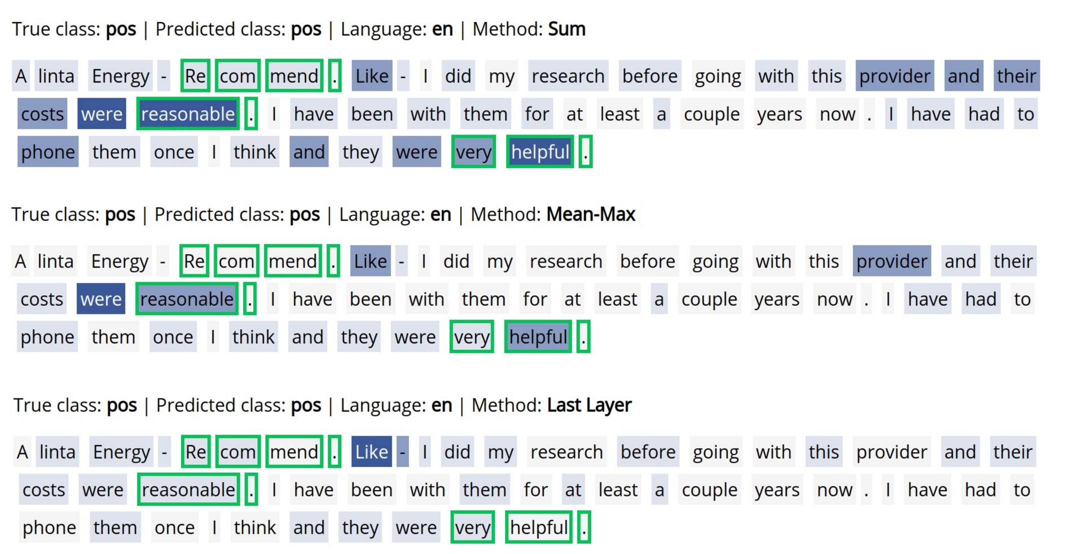
Evaluation metrics

The annotation labels have been transformed into boolean arrays (1 for label and 0 for no label). Based on that we calculated evaluation metrics:

1. **Mean pairwise euclidean distance** between attention value and token label
2. **Hamming distance** between two boolean arrays, where one array is the labels array and the other one is attention array with threshold applied:
   - **High** – attention value > 0.75
   - **Medium** – attention value > 0.5
   - **Low** – attention value > 0.25

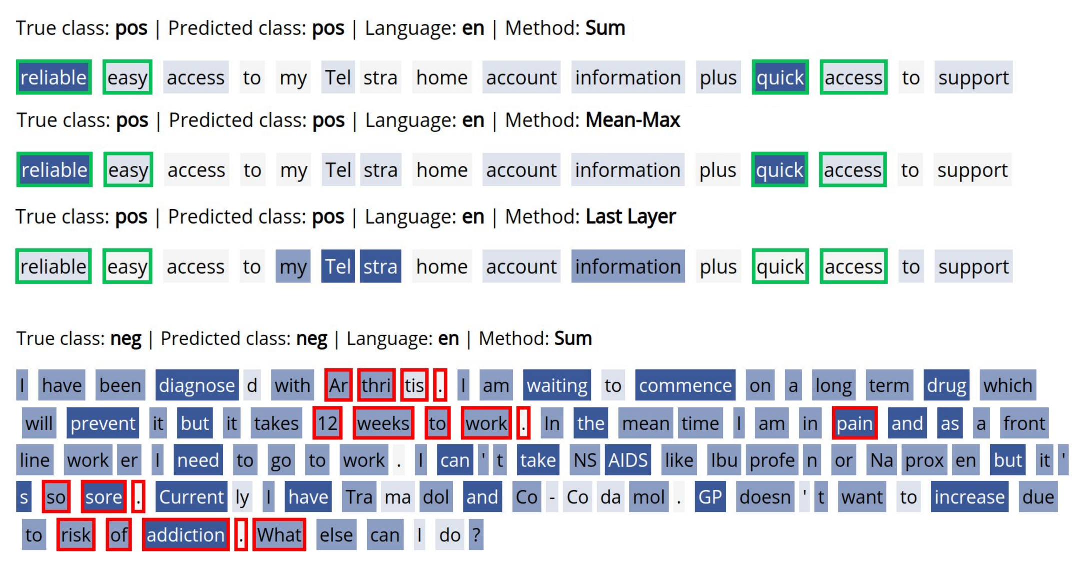
Results – Attention Values vs. Human Explanations

The color of the **gradient** background is the attention value (with three thresholds: **0.25**, **0.5**, **0.75**). The **green** and **red** borders show the human explanation labels.



"Reasonable" and "helpful" are highlighted in "Sum" method and not at all in "Last Layer" method. The method of aggregation is crucial for results.

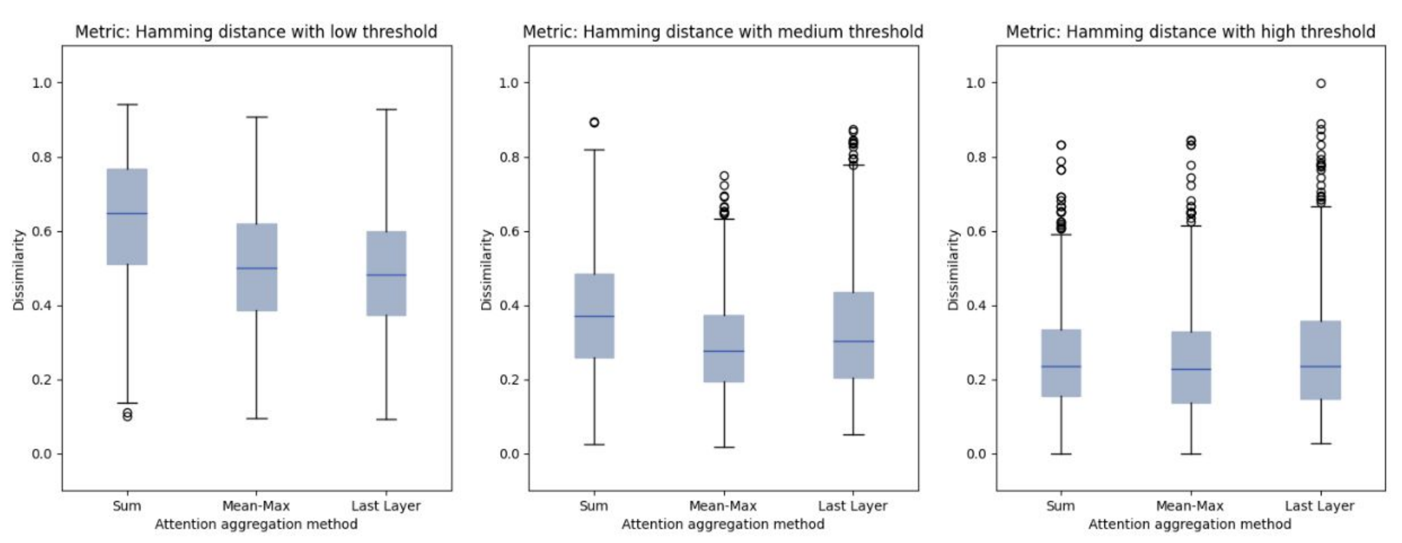
Model is attending to some explanatory words (issues) but also negations (cannot), articles (the), or conjunctions (and).



In case of short texts model has higher chance of focusing on correct tokens in opposition to long texts.

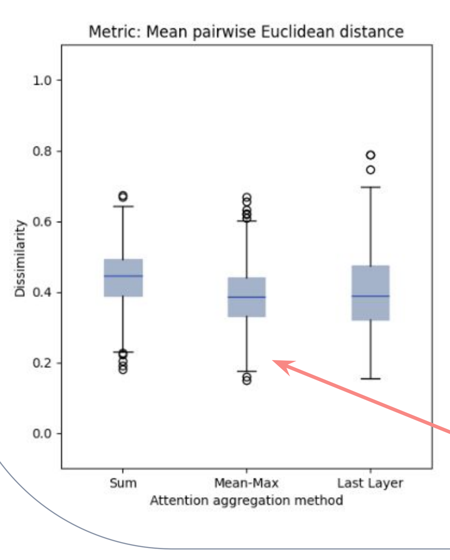
Results – Dissimilarities between Attention and Human Labels

Hamming distance



Hamming Distance with high threshold achieved the highest similarity to the original labels for all attention transformation methods

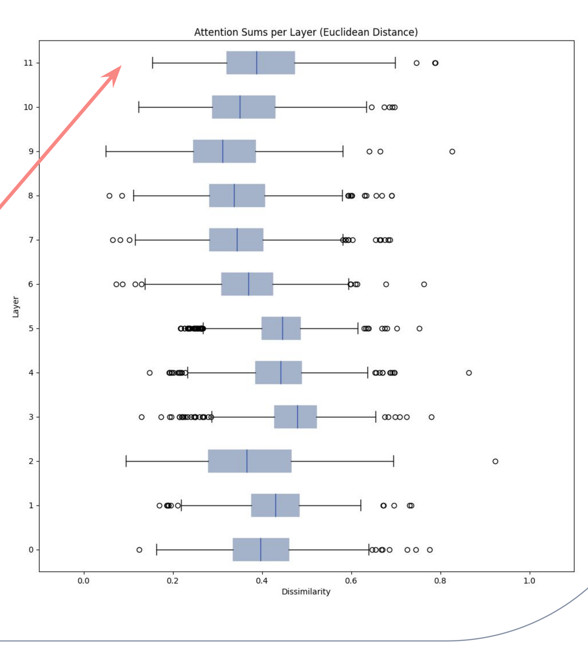
Euclidean distance



Mean pairwise euclidean distance is the similar for the mean-max and last layer method, yet mean-max distribution has a smaller variance

Attention per layer (mean pairwise euclidean distance)

Contradictory to first hypothesis it is not true that the last layer is the most similar to the human explanations (they are all similar)



Conclusions and future work

- Attention seems to look like in **typical MLM BERT encoder** – it is not much different because it's for sentiment analysis task
- Attention focuses the **sentence structure – not semantical meaning** of a text
- There is high attention on **punctuation, articles** and **conjunctions** (setting full stops and commas to zero change the attention values significantly)
- Due to the nature of the corpora, short sequences consist mostly of an opinion; because of that **explanations seem to work better for short sequences**.
- The method of **transforming multi-layer, multi-head, self-attention into per-token attentions is very important** – maybe applying some other aggregation would lead to more interpretable results
- After our analysis, we lean towards the conclusion that **attention is not usable for interpretability for classification task**
- **Future work**: other aggregation methods (e.g. attention flows), using other human explanations (e.g. opinion subjects), training a multiclass classifier (using a neutral or mixed sentiment class)

Contact

Hanna Sobocińska
➔ hanna.d.sobocinska@gmail.com
Emilia Kacprzak
➔ emilia.kacprzak@sentione.com
Agnieszka Pluwak
➔ agnieszka.pluwak@gmail.com

References

[1] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51(5), 1-42.
[2] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020, July). Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 8440-8451).
[3] Evtimov, R., Falli, M., & Maiwald, A. (2020). BERT Architecture. URL: https://humboldt-wi.github.io/blog/img/seminar/bert/bert_architecture.png (last access: 20.10.2023)
[4] Vig, J. (2019). A Multiscale Visualization of Attention in the Transformer Model. ACL 2019, 37.
[5] Vishwamitra, N., Hu, R. R., Luo, F., Cheng, L., Costello, M., & Yang, Y. (2020, December). On Analyzing COVID-19-related Hate Speech Using BERT Attention. In 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 669-676). IEEE.
[6] Akula, R., & Garibay, I. (2021). Interpretable multi-head self-attention architecture for sarcasm detection in social media. Entropy, 23(4), 394.

Acknowledgement. This work has been carried out as part of the Project "SentiDeepFusion: deep learning approach to automation of multilingual information extraction from colloquial texts for conversational interfaces and customer experience analytics" (POIR.01.01.01-00-0923/20), cofinanced by the European Regional Development Fund under the Smart Growth Programme 2014 – 2020.