On Multitask Hierarchical BERT Base Neural Network for Multi-label Persuasion Techniques Detection

POLISH-JAPANESE ACADEMY OF INFORMATION TECHNOLOGY

Katarzyna Baraniak and Marcin Sydow
Polish-Japanese Academy of Information Technology
kbaraniak@pjwstk.edu.pl, msyd@pjwstk.edu.pl

Motivation & Research Goals

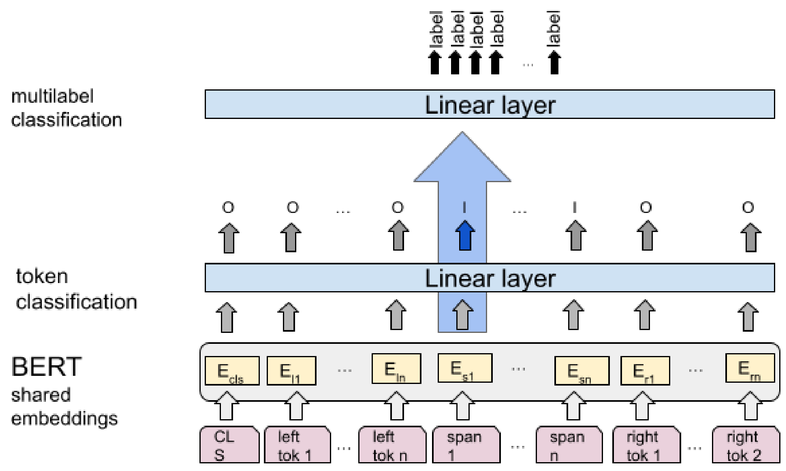
News articles are expected to present reliable information, however quite often contain some kind of manipulation. Unconscious reader can be unable to spot all kind of persuasion that he is exposed on. We created a system that combines multitask learning and hierarchical neural networks to detect the persuasion techniques in paragraphs of news articles. Our system was presented as a solution of a task 3 at a semeval 2023 competition [1] We tested our solution on languages: English, French, German, Italian, Polish and Russian.

Methods

Data. The input data is a single paragraph, the output is a list of techniques in a given paragraph. Additionally, for each paragraph the list of spans containing start character index, end character index and persuasion technique is provided. There are 23 unbalanced classes.

Persuasion techniques: Doubt, Whataboutism, Appeal to Hypocrisy, Causal Oversimplification, Appeal to Authority, Guilt by Association, Slogans, Flag Waving, Loaded Language, Red Herring, False Dillema-No Choice, Appeal to Popularity, Convers. Killer, Name Calling-Labelling, Appeal to Fear-Prejudice Exaggeration-Minimisation, Repetition, Straw Man, Obfuscation-Vaguity-Confusion

Model architecture:
Our solution is based on multitask hierarchical networks. Multitask networks share the same part or full architecture to solve several tasks being trained at once. Hierarchical networks are formed as an acyclic graph, what means that the tasks are learned by the networks' modules in some order. The results of previous modules influence the next modules. In our solution we created a network that solves two tasks: span identification and persuasion techniques classification.



The first layer of a model is a BERT layer with dropout. It is followed by the first linear layer for tag classification of persuasion span.
The index of the first token of the detected span is used to find the BERT embedding from the first layer, which is then passed to the second linear layer. The second linear determines whether the sample belongs to any of persuasion class or not.In contrast to [2] our main aim is only persuasion technique detection, we do not use special tokens and we use one multitask network.
Loss function:

$$Loss = 0.5 * Loss_1 + Loss_2 \quad (1)$$

The $Loss_1$ is cross entropy that calculates the loss for span identification. The $Loss_2$ is a binary cross entropy for a multi-label classification with added weight for each class.
The code of our solution is available at `https://github.com/Katarzynaa/persuasion_detection`

Selected Results

We present mean scores after 3 times run with random initialisation achieved on the dev set for all language.

	Our model		BERT		baseline	
L.	fmicro	fmacro	fmicro	fmacro	fmicro	fmacro
Eng	0.408±0.00	0.160±0.03	0.381±0.01	0.145±0.00	0.161	0.217
Po	0.367±0.01	0.211±0.02	0.362±0.02	0.218±0.01	0.125	0.057
Fr	0.416±0.01	0.283±0.01	0.386±0.01	0.285±0.01	0.293	0.135
It	0.435±0.02	0.221±0.01	0.402±0.03	0.205±0.01	0.389	0.104
Ru	0.446±0.00	0.159±0.02	0.399±0.02	0.147±0.02	0.253	0.043
Ge	0.412±0.00	0.237±0.02	0.396±0.00	0.236±0.01	0.331	0.100

Baseline results come from the leader board and present svm model with n-grams. Our model outperforms the basic BERT and baseline for all languages according to both measures.

Error analysis We analysed which classes in the English devset are the easiest/hardest to be recognized.

Technique	N.dev	N. train	precision	recall	f1
Doubt	187	51	0.29	0.25	0.27
Whataboutism	2	16	0.00	0.00	0.00
Appeal to Hypocr.	8	40	0.00	0.00	0.00
Causal Oversimp.	24	213	0.06	0.08	0.07
Appeal to Author.	28	154	0.10	0.04	0.05
Guilt by Associat.	4	59	**0.60**	0.75	**0.67**
Slogans	28	153	0.26	0.25	0.25
Flag Waving	96	287	0.46	0.50	0.48
Loaded Lang.	483	1809	0.49	**0.89**	0.63
Red Herring	19	44	0.00	0.00	0.00
False Dil.-NoCh.	63	122	0.30	0.05	0.08
App. to Popular.	34	15	0.00	0.00	0.00
Convers. Killer	25	91	0.00	0.00	0.00
Name Call.-Lab.	250	979	0.39	0.70	0.50
A.to Fear-Prejud.	137	310	0.26	0.15	0.19
Exaggerat.-Mini.	115	466	0.19	0.38	0.26
Repetition	141	544	0.19	0.04	0.06
Straw Man	9	15	0.00	0.00	0.00
Obf.-Vag.-Conf.	13	8	0.00	0.00	0.00

Example of errors:

- technique hard to detect based on a single paragraph, without context of the whole article

  "Red Herring" - introducing irrelevant information

  – *"Melania paired the mid-length half price frock with Christian Loubotin heels"*

- lack of broader context, world knowledge

  "Appeal to Hypocrisy"

  – *Of course, Sir Kim would have had plenty of targets had he decided to pass judgement on the present incumbent of the White House.*

- very short sentence

  "Conversation Killer" - a short and rather obvious statement or hidden in some long paragraph

  – *"Everybody knows it."*

Conclusions
We discovered that simple change of the index in Bert embedding may help to improve the persuasion classification. Moreover, we are able to identify spans and perform classification on a limited data using described networks. Our system works better than classic BERT for sequence classification.

References

[1] SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup
Piskorski, Jakub and Stefanovitch, Nicolas and Da San Martino, Giovanni and Nakov, Preslav, 2023

[2] ApplicaAI at SemEval-2020 task 11: On RoBERTa-CRF, span CLS and whether self-training helps them.
Dawid Jurkiewicz, Łukasz Borchmann, Izabela Kos- mala, and Filip Graliński In Pro- ceedings of the Fourteenth Workshop on Semantic Evaluation, pages 1415–1424, Barcelona (online). International Committee for Computational Linguis- tics

[3] Semeval-2020 task 11: Detection of propaganda techniques in news articles
Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. In Proceedings of the Fourteenth Workshop on Semantic Evalu-