NL4IA at SemEval-2023 Task 3: A Comparison of Sequence Classification and Token Classification to Detect Persuasive Techniques

Albert Pritzkau
Fraunhofer Institute for Communication,
Information Processing and Ergonomics
Wachtberg, Germany
albert.pritzkau@fkie.fraunhofer.de

Abstract

The following system description presents our approach to the detection of persuasion techniques in online news. The given task has been framed as a multi-label classification problem. In a multi-label classification problem, each input chunk—in this case paragraph—is assigned one of several class labels. Span level annotations were also provided. In order to assign class labels to the given documents, we opted for RoBERTa (A Robustly Optimized BERT Pretraining Approach) for both approaches—sequence and token classification. Starting off with a pre-trained model for language representation, we fine-tuned this model on the given classification task with the provided annotated data in supervised training steps.

1 Introduction

Political rhetoric, propaganda, and advertising are all examples of persuasive discourse. As defined by Lakoff (1982), persuasive discourse is the non-reciprocal "attempt or intention of one party to change the behavior, feelings, intentions, or viewpoint of another by communicative means". Thus, in addition to the purely content-related features of communication, the discursive context of utterances plays a central role. SemEval-2023 Task 3 (Piskorski et al., 2023) considers persuasion as a communication phenomenon. With this approach, it is assumed that communication depends not only on the meaning of words in an utterance, but also on what speakers intend to communicate with a particular utterance. This concept is from the linguistic subfield of pragmatics. It is not always possible to derive the function of an utterance from its form and additional contextual information is often needed. Recent research like (Tenney et al., 2019) (Jawahar et al., 2019) indicates the possibility that transformer-based networks capture structural information about language ranging from syntactic up to semantic features. Beyond these features, these architectures remain almost entirely unexplored. This task poses an attempt to explore the limits of the prevailing approach, in particular, investigating Transformers ability to capture pragmatic features.

2 Background

The central focus of this assignment is manipulative persuasion. The related task is to identify and evaluate propaganda and persuasion as found in social media. To identify and characterize manipulative persuasion, the context can be stretched arbitrarily far across aspects of epistemology, logic, intent estimates, psychological biases, knowledge of pre-existing narratives, and even physical context. However, to potentially solve this problem in an automated fashion, the prevailing method is to frame the given task as a classification problem. The different propaganda methods are understood as distinguishing criteria. Documents or even subsections of documents are annotated based on these features and thus form the input for training machine learning models, which should then be able to automatically recognize and classify corresponding sections. The undeniable success of this approach for many applications (for example, in NER) is due to the fact that the required features arise directly from the data or are already captured in the data representation used (word embeddings). In fact, current word embeddings already contain representations of a wide range of syntactic and morphological features that can be used to solve many problems. In the following pages, we discuss whether and to what extent the required characteristics are reflected in the training data. In particular, we consider whether and to what extent linguistic structures can be used as a decision criterion. In explaining our findings, a pragmatic perspective is adopted. In general, descriptive, analytical, and linguistic approaches such as speech act theory and

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 794–799

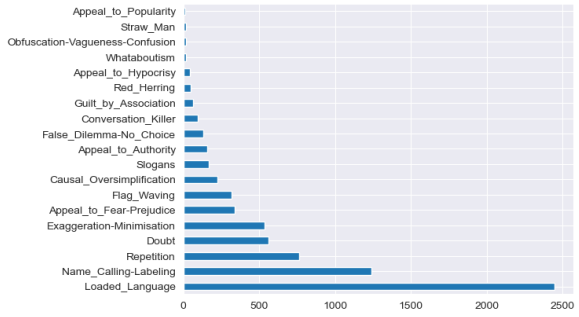

Figure 1: Label distribution - training set

rhetoric (or the use of specific rhetorical devices) are used to characterize public (political) discourse. Referring to the speech act theory (Austin, 1975), linguistic features, also in their form as rhetorical features, are assumed to be identified in the locutionary act. The illocutionary and perlocutionary acts, meanwhile, involve more complex information that might be used in feature engineering, thus incorporating the dimension of discourse.

Exploratory data analysis

The training data as input of this task was provided by 446 news and web articles in plain text format with 7201 annotations. Labels were given on paragraph level as one or more of those categories depicted in Figure 1. Imbalance in data can exert a major impact on the value and meaning of accuracy and other well-known performance metrics of an analytical model. Figure 1 depicts a clear skew towards three categories that account for two-thirds of the total annotations: Loaded_Language, Name_Calling-Labeling, and Repetition. Coincidentally, these categories are characterized by a very short annotation length, with a median of 20 characters (cf. Figure 2). Figure 3 on POS tag distribution shows the relevance of syntactic features in the annotated data, e.g. nouns and verbs seem to be most prominently represented within the annotation spans. Figure 4 further describes the relevance of certain POS tags for a specific annotation. For example, while the distribution for Loaded_Language and Repetition is largely uniform across the identified POS tags, Name_Calling-Labeling is characterized by the absence of verbs.

3 System overview

In this study, we evaluate and compare two different approaches: sequence classification and token classification. The comparison is performed at the level of trained models on the same set of data. The different scoring paradigms arise from applying token classifier and sequence classier heads, respectively, on a pre-trained model as the base model. We suggest that contextual information is captured and processed differently in both approaches, leading to a qualitative difference in the scores. Our results show that sequence classification is superior.

3.1 Pre-trained language representation

At the core of each solution of the given task lies a pre-trained language model derived from BERT (Devlin et al., 2018). BERT stands for Bidirectional Encoder Representations from Transformers. It is based on the Transformer model architectures introduced by Vaswani et al. (2017). The general approach consists of two stages. First, BERT is pre-trained on vast amounts of text, with an unsupervised objective of masked language modeling and next-sentence prediction. Second, this pre-trained network is then fine-tuned on task-specific, labeled data. The Transformer architecture is composed of two parts, an encoder and a decoder, for each of the two stages. The encoder used in BERT is an attention-based architecture for NLP. It works by performing a small, constant number of steps. In each step, it applies an attention mechanism to understand relationships between all words in a sentence, regardless of their respective position. By pre-training language representations, the encoder yields models that can either be used to extract high quality language features from text data, or fine-tune these models on specific NLP tasks (classification, entity recognition, question answering, etc.). We rely on RoBERTa (Liu et al., 2019), a pre-trained encoder model, which builds on BERT's language masking strategy. However, it modifies key hyper-parameters in BERT such as removing BERT's next-sentence pre-training objective, and training with much larger mini-batches and learning rates. Furthermore, RoBERTa was also trained on an order of magnitude more data than BERT, for a longer amount of time. This allows RoBERTa representations to generalize even better to downstream tasks compared to BERT.

3.2 Multi-Label Sequence Classification Problem

Model Architecture The task is given as a multi-label classification problem. The models for the experimental setup were based on RoBERTa. For

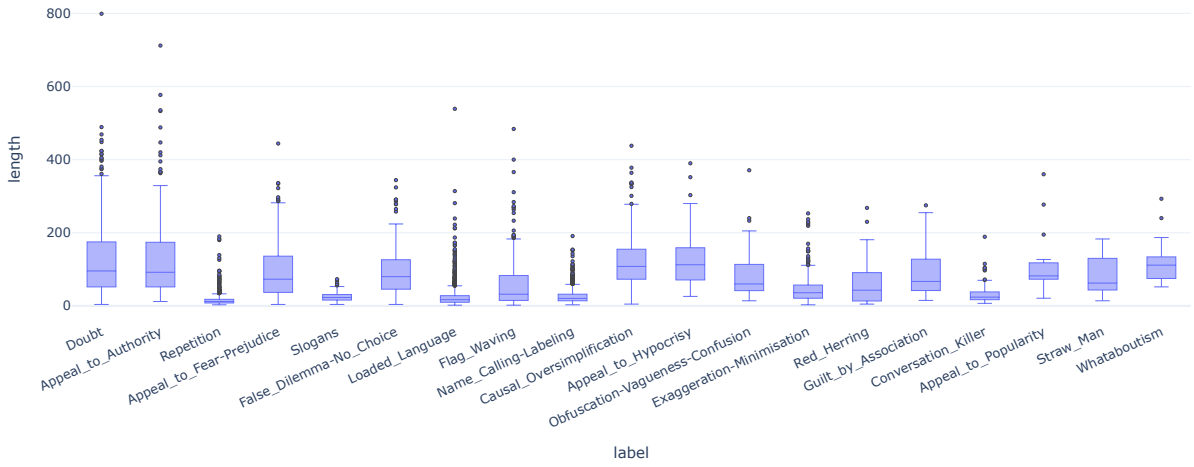

Figure 2: Length distribution of annotation spans - training set

the classification task, fine-tuning is initially performed using RobertaForSequenceClassification (Wolf et al., 2020)—$RoBERTa_{LARGE}$—as the pre-trained model. RobertaForSequenceClassification optimizes for a regression loss (Binary Cross-Entropy Loss) using an AdamW optimizer with an initial learning rate set to 2e-5.

3.3 Multi-Label Token Classification Problem

Tagging format We transformed the initial span markup into the IOB tagging format (Inside, Outside, Begin). As we have 19 possible entity classes, each token can be assigned one of the 39 tags given by an O-tag, and the I-tag and B-tag of the various techniques, respectively.

Model Architecture We fine-tuned a RoBERTa model to predict the above IOB tags for each token in the input sentence. In the default configuration each token is classified independently of the surrounding tokens. Although the surrounding tokens are taken into account in the contextualized embeddings, there is no modeling of the dependency between the predicted labels: for example, an I tag logically cannot follow an O tag. Since RoBERTa does not model the dependencies between the predicted tokens, we further added a linear-chain Conditional Random Field (CRF) model (Lafferty et al., 2001) as an additional layer, in order to model the dependency between the predicted labels of individual tokens. Since the sequence of an O-tag following an I-tag does not appear in the training set, it assigns by observation a very low probability to the transition from an O-tag to an I-tag. The CRF receives the logits for each input token, and makes a prediction for the entire input sequence, taking into account the dependencies between the labels, similarly to Lample et al. (2016). Note that RoBERTa works with byte pair encoding (BPE) units, while for the CRF itis necessary to work with complete words. Thus, only head tokens were used as input to the CRF, and any word continuation tokens were omitted.

4 Experimental setup

In both cases, fine-tuning was done with an NVIDIA TESLA V100 GPU using the Pytorch (Paszke et al., 2019) framework with a vocabulary size of 50265 and an input size of 512. The model was trained to optimize the objective for 50 epochs.

5 Results

We participated in the persuasion techniques detection task and focused on the English dataset. Official evaluation results on the test set are presented in Table 1. During the training phase, we focused on finding the best combinations of deep learning methods and optimized the corresponding hyperparameter settings. Finetuning pre-trained language models like RoBERTa on downstream tasks has become ubiquitous in NLP research and applied NLP. Even without extensive pre-processing of the training data, we already achieve competitive results. The resulting models serve as strong baselines, which, when fine-tuned, significantly outperform models trained from scratch.

Our submission is based on the trained model af-

	Team	F1 micro	F1 macro
1	APatt	0.37562	0.12919
2	SheffieldVeraAI	0.36802	0.17194
3	Appeal for attention	0.36299	0.16621
4	KInITVeraAI	0.36157	0.13324
5	NLUBot101	0.36058	0.19722
6	FTD	0.34637	0.08765
7	TeamAmpa	0.32457	0.15768
8	QCRI	0.32004	0.13251
9	DSHacker	0.32004	0.13983
10	CLAC	0.30933	0.07122
11	**NL4IA**	**0.30761**	**0.14160**
12	Unisa	0.29758	0.10871
13	MaChAmp	0.29476	0.14940
14	Riga	0.28045	0.06163
15	NAP	0.26294	0.08174
16	SATLab	0.25887	0.10291
17	ReDASPersuasion	0.25053	0.04476
18	UnedMediaBiasTeam	0.24070	0.07846
19	Baseline	0.19517	0.06925

Table 1: Official Ranking on Task 3 (English)

ter 100 training epochs in the case of the multi-label sequence classification. We were able to improve the official F1-micro score of 0.30761 to 0.34280 by evaluating the intermediate checkpoints. Based on this difference, we can assume a significant overfitting in our resulting model. With a maximum F1-micro score of 0.31325 in the case of the token classification approach, results remained in all cases below those of the sequence classification approach. Compared to the sequence-level assignment, it seems that token-level assignment requires an increased discriminatory power between the individual categories, which is clearly not sufficiently satisfied in this case.

When improving on the pretrained baseline models, class imbalance appears to be a primary challenge. With a highest ranked F1-micro score of 0.37562, it is necessary to discuss other causes for the low discriminatory power.

Possible challenges related to neural architectures arise either from under-specification of the objective function or from general difficulties of feature engineering. Difficulties with the objective function arise when the target variables, in our case the individual persuasion techniques, conceptually cannot be well separated. Issues with feature engineering are to be expected when required features cannot be captured from the training data. Tenney et al. (2019) suggest that transformer-based networks are able to glean structural information–both syntactic and semantic–from language. If this is so, we expect that further important features may be hidden in the broader context, especially when it comes to manipulative communication. Since these features do not emerge from the training data, they must be made available to the training process in some other way. Features of interest may be derived from research in pragmatics.

6 Conclusion

The use of neural architectures in the field of pragmatics remains largely unexplored. The results of the given task demonstrate the limitations of this method. In the future, we would like to extend the current approach to features of the extended communicative context. Our research concerns the specification of a consistent objective function aligned with the discursive context of manipulative communication. We hypothesize that the target variables of this function in the form of different discourse elements will respond to different features of the given communicative context. If the required features cannot be derived from the linguistic structure of the utterances, they have to be obtained from the extended context of the communication. We are investigating ways to make external features available to the training process. In order to identify pragmatic features and how to exploit them, XAI methods might come to help.

References

John Langshaw Austin. 1975. *How to do things with words*. Oxford University Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ISBN: 1810.04805v2.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? Technical report.

John Lafferty, Andrew Mccallum, and Fernando Pereira. 2001. Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data Abstract. 2001(June):282–289.

Robin Tolmach Lakoff. 1982. Persuasive discourse and ordinary conversation, with examples from advertising. *Analyzing discourse: Text and talk*, pages 25–42. Publisher: Georgetown, Georgetown University Press.

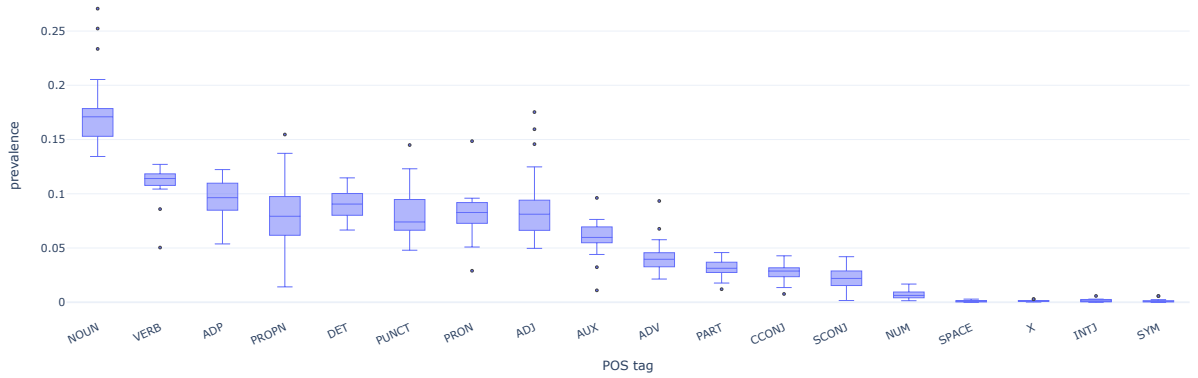

Figure 3: POS distribution - training set

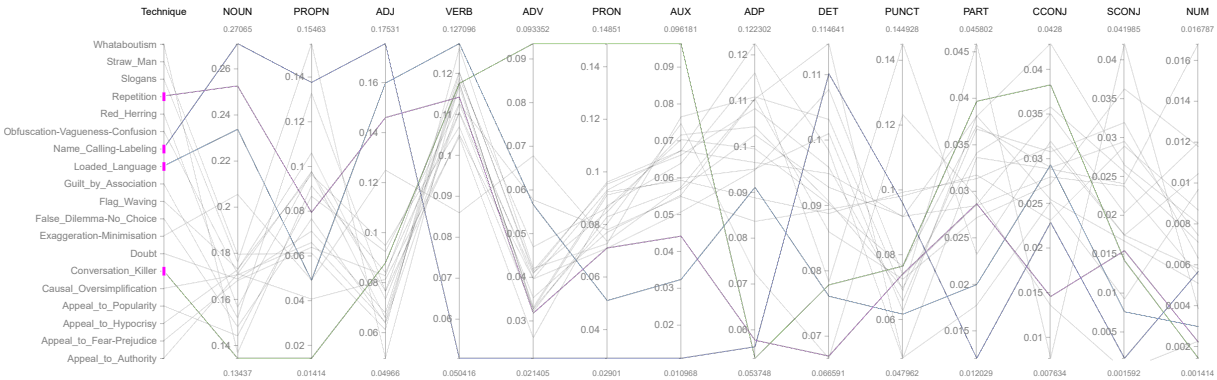

Figure 4: Parallel Coordinates of POS distributions - training set

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, pages 260–270. Association for Computational Linguistics (ACL).

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 e-prints*, pages arXiv–1907.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Neural information processing systems foundation. ISSN: 10495258.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, SemEval 2023, Toronto, Canada.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations.

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*, volume 2017-Decem, pages 5999–6009. ISSN: 10495258.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *arxiv.org*, pages 38–45.

A Classification reports

	precision	**recall**	**f1-score**	**support**
Appeal_to_Authority	0.29	0.29	0.29	96
Labeling	0.38	0.34	0.36	284
Repetition	0.24	0.13	0.17	156
Doubt	0.28	0.25	0.26	199
Loaded_Language	0.29	0.28	0.29	486
Flag_Waving	0.46	0.43	0.44	249
No_Choice	0.68	0.61	0.64	103
Causal_Oversimplification	0.15	0.17	0.16	126
Minimisation	0.35	0.32	0.34	191
Conversation_Killer	0.04	0.02	0.03	46
Prejudice	0.58	0.56	0.57	260
Slogans	0.20	0.20	0.20	41
Guilt_by_Association	0.07	0.04	0.05	23
Red_Herring	0.00	0.00	0.00	21
Whataboutism	0.83	0.62	0.71	8
Appeal_to_Hypocrisy	0.00	0.00	0.00	4
Straw_Man	0.00	0.00	0.00	1
Appeal_to_Popularity	0.00	0.00	0.00	2
Confusion	0.00	0.00	0.00	1
avg / total	**0.35**	**0.32**	**0.34**	**2297**

Table 2: Classification report on task 3 (en) for the best checkpoint on the development set (token classification).

	precision	**recall**	**f1-score**	**support**
Appeal_to_Authority	0.50	0.18	0.26	39
Labeling	0.71	0.65	0.68	192
Repetition	0.36	0.31	0.33	97
Doubt	0.55	0.32	0.40	92
Loaded_Language	0.67	0.67	0.67	333
Flag_Waving	0.60	0.67	0.63	45
No_Choice	0.27	0.29	0.28	21
Causal_Oversimplification	0.34	0.34	0.34	38
Minimisation	0.58	0.41	0.48	79
Conversation_Killer	0.00	0.00	0.00	17
Prejudice	0.47	0.46	0.47	61
Slogans	0.77	0.52	0.62	33
Guilt_by_Association	0.17	0.08	0.11	12
Red_Herring	0.00	0.00	0.00	7
Whataboutism	0.00	0.00	0.00	1
Appeal_to_Hypocrisy	0.50	0.33	0.40	3
Straw_Man	0.00	0.00	0.00	3
Appeal_to_Popularity	0.00	0.00	0.00	4
Confusion	0.00	0.00	0.00	0
avg / total	**0.34**	**0.28**	**0.30**	**1077**

Table 3: Classification report on task 3 (en) for the best checkpoint on the development set (sequence classification).