CEUR-WS.org/Vol-3496/dipromats-paper1.pdf

Investigating Propaganda Considering the Discursive Context of Utterances*

Albert Pritzkau[1,*,†]

[1] *Fraunhofer Institute for Communication, Information Processing and Ergonomics FKIE, Fraunhofer Str. 20, 53343 Wachtberg, Germany*

Abstract

The following system description presents our approach to the detection of propagandistic techniques in tweets. The given task has been framed as a multi-label classification problem. In a multi-label classification problem, each input chunk—in this case tweet—is assigned one of several class labels. In order to assign class labels to the given utterances, we opted for RoBERTa (A Robustly Optimized BERT Pretraining Approach) for sequence 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. In addition to the content of the message, further features describing the general communication context is taken into account.

Keywords

Pragmatics, Information Extraction, Data Augmentation, Text Classification, RoBERTa

1. Introduction

Political rhetoric, propaganda, and advertising are all examples of persuasive discourse. As defined by Lakoff [1], persuasive discourse is the nonreciprocal "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. DIPROMATS 2023[2] 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 [3] [4] 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.

IberLEF 2023, September 2023, Jaén, Spain

[*] Corresponding author.

[†] These authors contributed equally.

albert.pritzkau@fkie.fraunhofer.de (A. Pritzkau)

0000-0001-7985-0822 (A. Pritzkau)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

2. Background

The central focus of this assignment is manipulative persuasion. The related tasks are to identify and evaluate propaganda and persuasion as found in social media, in particular on Twitter. 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 posts 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 rhetoric (or the use of specific rhetorical devices) are used to characterize public (political) discourse. Referring to the speech act theory [5], 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.

Task descriptions

This work describes the participation in all three subtasks, considering only the English part, respectively only. The challenge for the first subtask is to decide whether a given tweet contains propaganda techniques at all. Accordingly, the task is given as a binary classification problem. Beyond the mere identification of propaganda, a characterization of propaganda is carried out in Task 2 and 3. The messages are to be automatically categorized into different types of propaganda. The latter two subtasks differ in the typology of categorization and propose a coarse-grained categorization with four propaganda classes and a fine-grained categorization with 13 subclasses considered. The fine-grained categorization extends the coarse-grained categorization from subtaks 2 with various subclasses. Since individual messages could also be assigned to more than one class, both subtasks are considered as a multiclass and multilabel classification task.

Exploratory data analysis

The training data as input of this task was provided by 8408 tweets from diplomats of four different international actors: China, Russia, United States, and the European Union in plain text format. In addition to the content of the message and the respective annotation, further information describing the communication context could be extracted from the training data. This includes, for example, engagement features. Other features such as the timestamps and tweet ID were neglected in our case.

Labels were given on message level as one or more of those categories depicted in Figures 1 through 3 for each subtask, respectively.

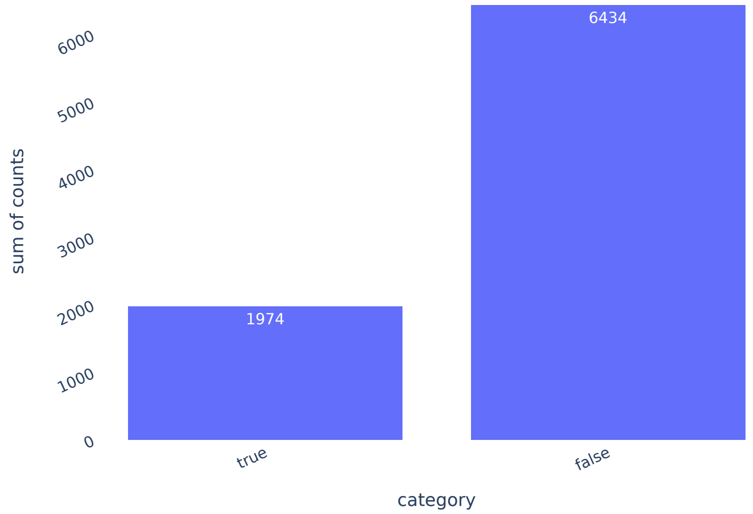

Figure 1: Label distribution for Subtask 1 - training set

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 messages without annotation in subtask 1. In the case of subtask 2, the category *4 appeal to authority* is heavily underrepresented (cf. Figure 2), which makes it difficult for the algorithm to learn anything at all from this category. Finally, a clear skew towards four categories can be seen from the Figure 3 for subtask 3, that account for more than three-quarters (2234) of the total annotations (2643): *1 appeal to commonality - flag waving* (545), *2 discrediting the opponent - name calling* (213), *2 discrediting the opponent - undiplomatic assertiveness/whataboutism* (563), *3 loaded language* (913).

In addition to the text of the messages, the training dataset also contains engagement features in the form of a numerical value representing the number of reweets or favourites that this tweet exhibited at the time of recording. The statistical analysis of this feature in Table 1 indicates a significant correlation to the existence of a particular propagandistic techniques. A correlation of this kind cannot be seen for the referenced country, as a second metadatum.

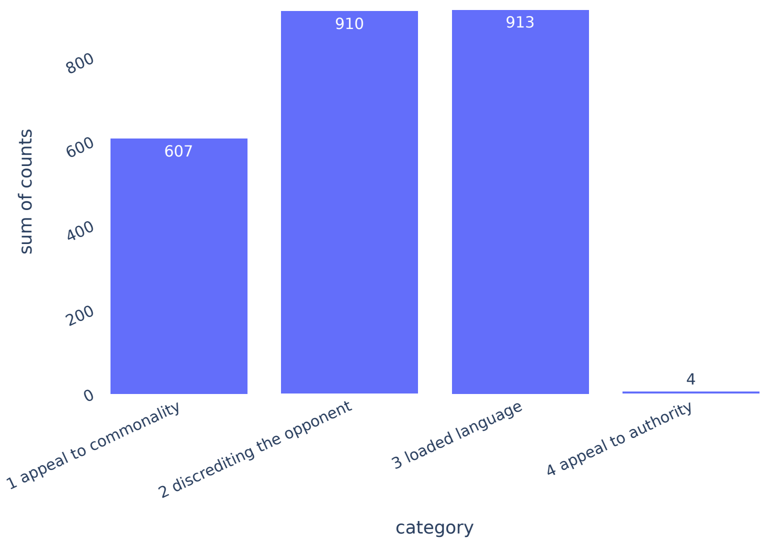

Figure 2: Label distribution for Subtask 2 - training set

category	count	mean	std	min	25%	50%	75%	max
False	6434.0	2474.264	16503.503	0.0	41.00	124.5	524.75	503433.0
True	1974.0	15763.304	47270.240	0.0	82.25	451.5	4299.75	713124.0

Table 1
Descriptive statistics of the engagement features given.

Data augmentation

Aristotle determined the components necessary for persuasion. They are called the three pillars of persuasion - ethos, pathos, and logos. This conceptualization also underlies the annotation of the given dataset. In augmenting the dataset, we focused on the aspect of pathos. Pathos means to persuade by appealing to the emotions of the audience. We consider emotion and sentiment as a natural tool for communication and social influence. These features were extracted before training by applying external extraction models[6] on each tweet. For sentiment as a rather weak characterization the sub-categories were: Positive, Negative, and Neutral. For emotion as rather strong characterization they are: Anger, Joy, Optimism, and Sadness. Instead of making technical changes to the architecture, the original message was augmented with the additional features in text form, e.g. "The phrase contains optimism as emotional content. Its sentiment is positive. The message had 16 interactions. Country of origin is Russia."

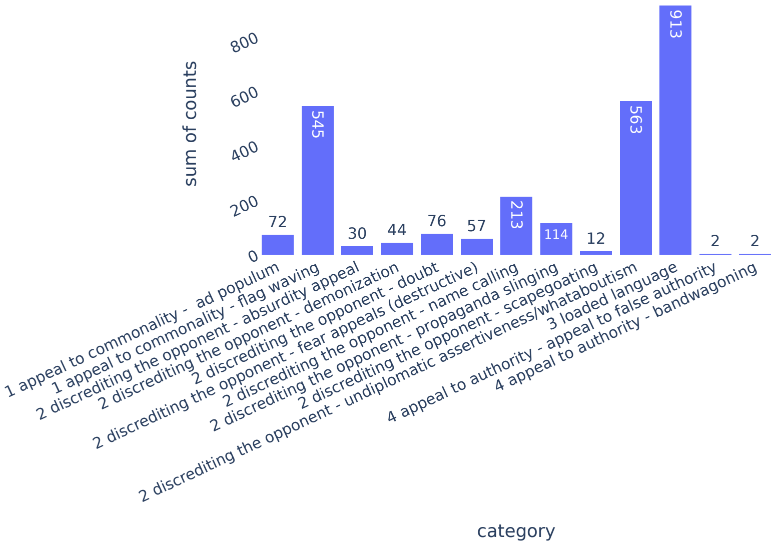

Figure 3: Label distribution for Subtask 3 - training set

3. System overview

In this study, we evaluate and compare a sequence classification approach on the given data with different augmentations. The comparison is performed at the level of trained models on the same set of data. The different scoring paradigms arise from applying sequence classier heads on a pre-trained model as the base model. We suggest that contextual information is leading to a qualitative difference in the scores.

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 [7]. BERT stands for Bidirectional Encoder Representations from Transformers. It is based on the Transformer model architectures introduced by Vaswani et al. [8]. 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 [9], 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 All subtasks are considered as multi-label classification problems. The models for the experimental setup were based on RoBERTa. For the classification task, fine-tuning is initially performed using RobertaForSequenceClassification [10]—$RoBERTa_{LARGE}$—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. After a warmup period during which the learning rate increases linearly from 0 to the initial learning rate, the optimizer is scheduled to decrease the actual learning rate linearly to 0. The training was launched with 20 training epochs each. However, this relatively high number is significantly reduced by an early stopping callback that monitored the performance of the model on validation dataset. A patience of 5 epochs is set for this callback.

4. Experimental setup

In all cases, fine-tuning was done with an NVIDIA TESLA V100 GPU using the Pytorch [11] framework with a vocabulary size of 50265 and an input size of 512. The training data has been randomly split into training and validation subsets with a ratio of 85:15 resulting in 6726 and 1682 post in the respective sets.

5. Results

We participated in all three subtasks on the detection of propagandistic techniques and focused on the English dataset. Official evaluation results on the test set are presented in Table 2,4 and 4. Submissions were optimized for the minimum validation loss to prevent over-fitting of the resulting model. 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. The confusion matrics in Figure 4, 5 and 6 provide a detailed category-level view of the performance of the trained model on the validation dataset. In particular, Figure 6 clearly shows that the model performs particularly well on those categories that are also well represented in the initial distribution, namely *1 appeal to commonality - flag waving*, *2 discrediting the opponent - name calling*, *2*

	Gold	NL4IA	Baseline
False	1.0000	0.9336	0.0000
True	1.0000	0.6570	0.2928
F1 macro	1.0000	0.7953	0.2928
ICM-Hard	0.6611	0.1701	-1.8963
ICM-Hard Norm	1.0000	0.8080	0.0000

Table 2
Official evaluation results by category on Subtask 1 (English)

	Gold	NL4IA	Baseline
1 appeal to commonality	1.0000	0.5363	0.0000
2 discrediting the opponent	1.0000	0.6827	0.0000
3 loaded language	1.0000	0.6489	0.0000
4 appeal to authority	1.0000	0.0000	0.0006
F1 macro	1.0000	0.5591	0.0000
ICM-Hard	0.9296	0.1778	-11.4286
ICM-Hard Norm	1.0000	0.9392	0.0000

Table 3
Official evaluation results by category on Subtask 2 (English)

	Gold	NL4IA	Baseline
1 appeal to commonality - ad populum	1.0000	0.5333	0.0000
1 appeal to commonality - flag waving	1.0000	0.5366	0.0000
2 discrediting the opponent - absurdity appeal	1.0000	0.7143	0.0000
2 discrediting the opponent - demonization	1.0000	0.3000	0.0000
2 discrediting the opponent - doubt	1.0000	0.5909	0.0000
2 discrediting the opponent - fear appeals (destructive)	1.0000	0.3158	0.0000
2 discrediting the opponent - name calling	1.0000	0.5789	0.0000
2 discrediting the opponent - propaganda slinging	1.0000	0.5455	0.0000
2 discrediting the opponent - scapegoating	1.0000	0.0000	0.0000
2 discrediting the opponent - undiplomatic assertiveness/whataboutism	1.0000	0.5809	0.0000
3 loaded language	1.0000	0.6624	0.0000
4 appeal to authority - appeal to false authority	1.0000	0.0000	0.0000
4 appeal to authority - bandwagoning	1.0000	0.0000	0.0000
F1 macro	1.0000	0.4838	0.0000
ICM-Hard	1.0748	0.1227	-11.5747
ICM-Hard Norm	1.0000	0.9247	0.0000

Table 4
Official evaluation results by category on Subtask 3 (English)

discrediting the opponent - undiplomatic assertiveness/whataboutism, and *3 loaded language* (cf. Figure 3).

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 propagandistic 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 per se, they must be made available to the training process in some other way. In the present case, the metadata already contained basic engagement features that could be incorporated into the training. Additional context features (sentiment, emotion) could be derived from the content using external estimators. Other features of interest may be derived from research in pragmatics.

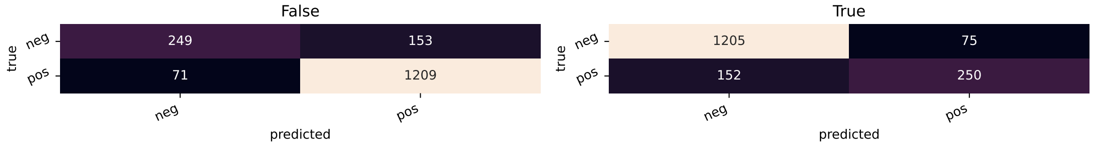

Figure 4: Multilabel confusion matrix on Subtask 1 - training vs. validation set

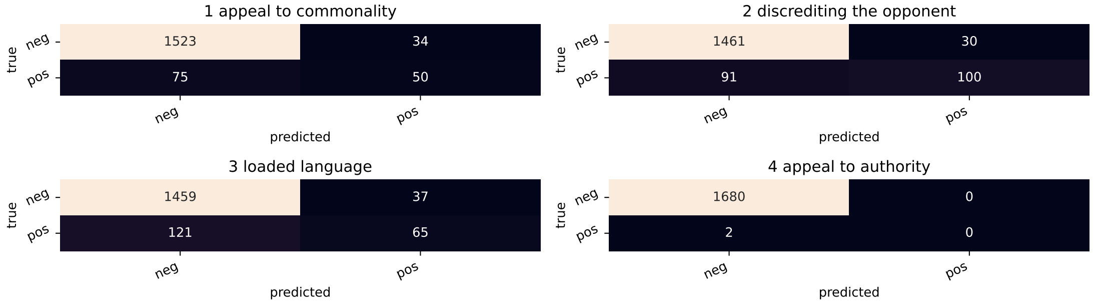

Figure 5: Multilabel confusion matrix on Subtask 2 - training vs. validation set

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.

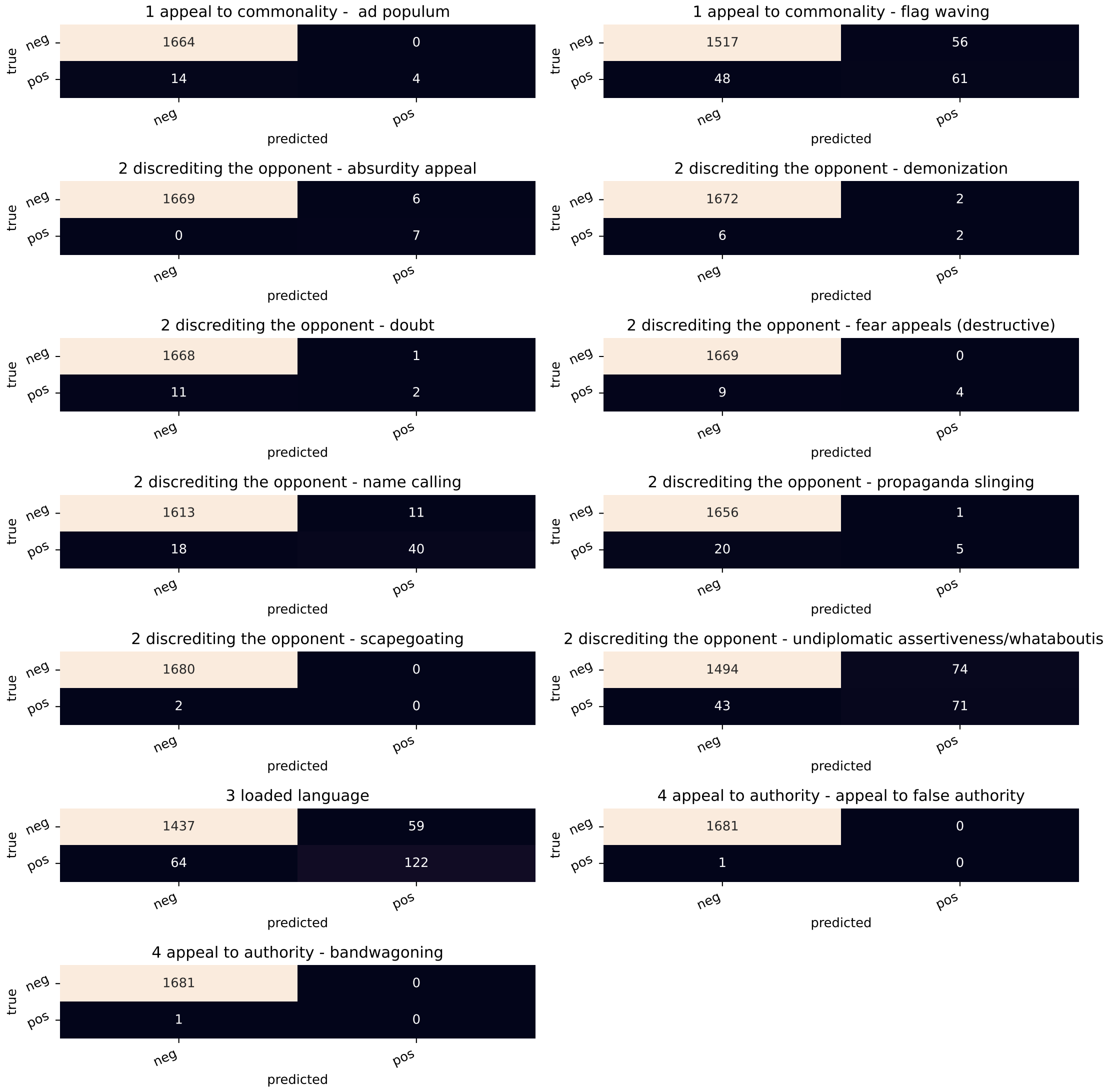

Figure 6: Multilabel confusion matrix on Subtask 3 - training vs. validation set

In order to identify pragmatic features and how to exploit them, XAI methods might come to help.

References

[1] R. T. Lakoff, Persuasive discourse and ordinary conversation, with examples from advertising, Analyzing discourse: Text and talk (1982) 25–42. Publisher: Georgetown, Georgetown University Press.

[2] Pablo Moral, Guillermo Marco, Julio Gonzalo, Jorge Carrillo-de-Albornoz, Iván Gonzalo-Verdugo, Overview of DIPROMATS 2023: automatic detection and characterization of propaganda techniques in messages from diplomats and authorities of world powers, Procesamiento del Lenguaje Natural 71 (2023).

[3] I. Tenney, P. Xia, B. Chen, A. Wang, A. Poliak, R. T. McCoy, N. Kim, B. V. Durme, S. R. Bowman, D. Das, E. Pavlick, What do you learn from context? Probing for sentence structure in contextualized word representations (2019). URL: http://arxiv.org/abs/1905.06316.

[4] G. Jawahar, B. Sagot, D. Seddah, What does BERT learn about the structure of language?, Technical Report, 2019. URL: https://hal.inria.fr/hal-02131630.

[5] J. L. Austin, How to do things with words, Oxford University Press, 1975.

[6] F. Barbieri, J. Camacho-Collados, L. Neves, L. Espinosa-Anke, Tweeteval: Unified benchmark and comparative evaluation for tweet classification, arXiv preprint arXiv:2010.12421 (2020).

[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018). ISBN: 1810.04805v2.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, volume 2017-Decem, 2017, pp. 5999–6009. ISSN: 10495258.

[9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv e-prints (2019) arXiv–1907.

[10] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. v. Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-Art Natural Language Processing, in: arxiv.org, 2020, pp. 38–45. URL: https://github.com/huggingface/. doi:10.18653/v1/2020.emnlp-demos.6.

[11] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, volume 32, Neural information processing systems foundation, 2019. URL: http://arxiv.org/abs/1912.01703, iSSN: 10495258.