NLytics at CheckThat! 2021: Multi-class fake news detection of news articles and domain identification with RoBERTa - a baseline model

Albert Pritzkau[1]

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

Abstract

The following system description presents our approach to the detection of fake news in texts. The given task has been framed as a multi-class classification problem. The multi-class classification problem is one in which a target variable such as the given class label is associated with every input chunk.

In order to assign class labels to the given documents, we opted for RoBERTa (A Robustly Optimized BERT Pretraining Approach) as a neural network architecture 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.

Keywords

Sequence Classification, Deep Learning, Transformers, RoBERTa

1. Introduction

The proliferation of disinformation online, has given rise to a lot of research on automatic fake news detection. CLEF 2021 - CheckThat! Lab [1, 20] considers disinformation as a communication phenomenon. By detecting the use of various linguistic features in communication, it takes into account not only the content but also how a subject matter is communicated.

The shared task [2] defines the following subtasks:

Subtask A Given the "textual content" of an article, specify a credibility level for the content ranging between "true" and "false" including "other".

Subtask B Given the "textual content" of an article, specify a tpical domain covered by the content.

In this work, we covered our approach on both multi-class classification tasks by detecting fake news in the former and assigning a topical domain in the latter task. To build our models, both subtasks only textual content is given as input. Below, we describe the systems built for

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

albert.pritzkau@fkie.fraunhofer.de (A. Pritzkau)
https://www.fkie.fraunhofer.de/ (A. Pritzkau)
0000-0001-7985-0822 (A. Pritzkau)

these two subtasks. At the core of our systems is RoBERTa [3], a pre-trained model based on the Transformer architecture [4].

2. Related Work

The goal of the shared task is to investigate automatic techniques for identifying various rhetorical and psychological features of disinformation campaigns. A comprehensive survey on fake news has been presented by Zhou and Zafarani [5]. Based on the structure of data reflecting different aspects of communication, they identified four different perspectives on fake news: (1) the false knowledge it carries, (2) its writing style, (3) its propagation patterns, and (4) the credibility of its creators and spreaders.

The shared task emphasizes communicative styles that systematically co-occur with persuasive intentions of (political) media actors. Similar to de Vreese et al. [6], propaganda and persuasion is considered as an expression of political communication content and style. Hence, beyond the actual subject of communication, the way it is communicated is gaining importance.

We build our work on top of this foundation by first investigating content-based approaches for information discovery. Traditional information discovery methods are based on content: documents, terms, and the relationships between them [7]. They can be considered as general Information Extraction (IE) methods, automatically deriving structured information from unstructured and/or semi-structured machine-readable documents. Communities of researchers contributed various techniques from machine learning, information retrieval, and computational linguistics to the different aspects of the information extraction problem. From a computer science perspective, existing approaches can be roughly divided into the following categories: rule-based, supervised, and semi-supervised. In our case, we followed the supervised approach by reframing the complex language understanding task as a simple classification problem. Text classification also known as text tagging or text categorization is the process of categorizing text into organized groups. By using Natural Language Processing (NLP), text classifiers can automatically analyze human language texts and then assign a set of predefined tags or categories based on their content. Historically, the evolution of text classifiers can be divided into three stages: (1) simple lexicon- or keyword-based classifiers, (2) classifiers using distributed semantics, and (3) deep learning classifiers with advanced linguistic features.

2.1. Deep Learning for Information Extraction

Recent work on text classification uses neural networks, particularly Deep Learning (DL). Badjatiya et al. [8] demonstrated that these architectures, including variants of recurrent neural networks (RNN) [9, 10, 11], convolutional neural networks (CNN) Zhang et al. [12], or their combination (CharCNN, WordCNN, and HybridCNN), produce state-of-the-art results and outperform baseline methods (character n-grams, TF-IDF or bag-of-words representations).

2.2. Deep Learning architectures

Until recently, the dominant paradigm in approaching NLP tasks has been focused on the design of neural architectures, using only task-specific data and word embeddings such as those mentioned above. This led to the development of models, such as Long Short Term Memory (LSTM) networks or Convolution Neural Networks (CNN), that achieve significantly better results in a range of NLP tasks than less complex classifiers, such as Support Vector Machines, Logistic Regression or Decision Tree Models. Badjatiya et al. [8] demonstrated that these approaches outperform models based on character and word n-gram representations. In the same paradigm of pre-trained models, methods like BERT [13] and XLNet [14] have been shown to achieve state-of-the-art performance in a variety of tasks.

2.3. Pre-trained Deep Language Representation Model

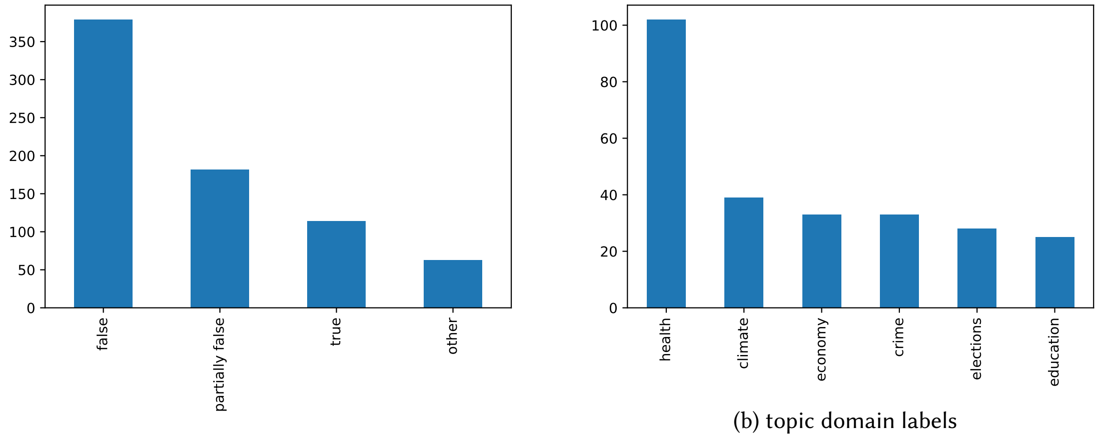
Indeed, the usage of a pre-trained word embedding layer to map the text into vector space which is then passed through a neural network, marked a significant step forward in text classification. The potential of pre-trained language models, as e.g. Word2Vec [15], GloVe [16], fastText [17], or ELMo [18] to capture the local patterns of features to benefit text classification, has been described by Castelle [19]. Modern pre-trained language models use unsupervised learning techniques such as creating RNNs embeddings on large texts corpora to gain some primal "knowledge" of the language structures before a more specific supervised training steps in.

2.4. About BERT and RoBERTa

BERT stands for Bidirectional Encoder Representations from Transformers. It is based on the Transformer model architectures introduced by Vaswani et al. [4]. 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 [3], a pre-trained Encoder model which builds on BERT's language masking strategy. However, it modifies key hyperparameters 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. In this study, RoBERTa is at the core of each solution of the given subtasks.

3. Dataset

The data for the task was developed during the CLEF-2021 CheckThat! campaign [1, 20, 2] and provided by Shahi et al. [21]. The AMUSED framework presented by Shahi [22] was used for data collection. Both subtasked were framed as multi-class classification problem. Class



(a) news labels

(b) topic domain labels

Figure 1: Label distribution - training set

label were provided as credibility levels {false, partially false, true, other} and topical categories {health, economy, crime, climate, elections, and education} for each subtask, respectively. The content parts are distributed between title and body of messages. Both field were concatenated to serve as the input for training.

4. Exploratory data analysis

Unbalanced class distribution Imbalance in data can exert a major impact on the value and meaning of accuracy and on certain other well-known performance metrics of an analytical model. Figure 1 depicts a clear skew towards false information and health information, respectively, in the respective subtask.

Token count Transformer-based models are unable to process long sequences due to their self-attention mechanism, which scales quadratically with the sequence length. BERT-based models enforce a hard limit of 512 tokens, which is usually enough to process the majority of sequences in most benchmark datasets. Statistical summary of token counts in Table 1, however, suggests that most of the sequences of the training set exceed this limit. Thus, anything beyond this limitation will be truncated.

5. Our approach

In this section, we provide a general overview of our approach to both subtasks.

5.1. Experimental setup

Model Architecture Subtasks A and B are both given as a multi-class classification problem. Our model for this subtask is based on RoBERTa. For the classification task, fine-tuning is

Table 1
Statistical summary of token counts on the training set.

	fake news corpus	topic domain corpus
doc count	738	260
mean	729.94	879.42
std	769.98	878.23
min	14	91
25%	305.25	353.75
50%	536.00	670.00
75%	859.75	965.25
max	5828	5655

performed using *RobertaForSequenceClassification*[23] – roberta-base – as the pre-trained model. *RobertaForSequenceClassification* optimizes for Binary Cross Entropy Loss using an AdamW optimizer with an initial learning rate set to 2e-5. Fine-tuning is done on NVIDIA TESLA P100 GPU using the Pytorch [24] framework with a vocabulary size of 50265 and an input size of 512. The model is trained to optimize the objective for 3 epochs. To estimate the performance of the resulting models we have chosen a ratio of 82/18 to split the data into training and validation set.

Input Embeddings The input embedding layer converts the inputs into sequences of features: word-level sentence embeddings. These embedding features will be further processed by the latter encoding layers.

Word-Level Sentence Embeddings A sentence is split into words $w_1, ..., w_n$ with length of n by the WordPiece tokenizer [25]. The word w_i and its index i (w_i's absolute position in the sentence) are projected to vectors by embedding sub-layers, and then added to the index-aware word embeddings:

$$\hat{w}_i = WordEmbed(w_i)$$

$$\hat{u}_i = IdxEmbed(i)$$

$$h_i = LayerNorm(\hat{w}_i + \hat{u}_i)$$

Attention Layers Attention layers [26, 27] aim to retrieve information from a set of context vectors y_j related to a query vector x. An attention layer first calculates the matching score a_j between the query vector x and each context vector y_j. Scores are then normalized by softmax:

$$a_j = score(x, y_j)$$

$$\alpha_j = exp(a_j)/\Sigma_k exp(a_k)$$

The output of an attention layer is the weighted sum of the context vectors w.r.t. the softmax normalized score: $Att_{X \to Y}(x, \{y_j\}) = \Sigma_j \alpha_j y_j$. An attention layer is called self-attention when the query vector x is in the set of context vectors y_j. Specifically, we use the multi-head attention following Transformer [4].

	fake news corpus	topic domain corpus
accuracy	0.60	0.83
F1	0.43	0.76
accuracy_and_F1	0.52	0.79

Table 2
Evaluation measures on the best performing model checkpoint on the validation set.

Rank	Team	F1-macro
1	sushmakumari	0.8376451772
2	Saud	0.514230825
3	kannanrrk	0.5034290158
4	jmartinez595	0.4680478564
5	hariharanrl	0.448832841
6	cipriancus	0.4463072939
7	Huertas97	0.4142550112
8	pHartl	0.4041478353
9	boby024	0.4013434521
10	nomanashraf712	0.3892308335
11	SaifuddinSohan	0.3822517154
12	**NLytics**	**0.386246366**
13	Ninko	0.3579356596

Table 3
Results on subtask A

Rank	Team	F1-macro
1	hariharanrl	0.8813840965
2	sushmakumari	0.8552061398
3	Ninko	0.8410300885
4	kannanrrk	0.817812671
5	nomanashraf712	0.7896621462
6	architap	0.786037089
7	**NLytics**	**0.7310895828**
8	Huertas97	0.676509793
9	ep	0.4791621206
10	boby024	0.4484680905
11	ashik2580	0.1450648056
12	fazlfrs	0.1450648056
13	azaharudue	0.1282925881

Table 4
Results on subtask B

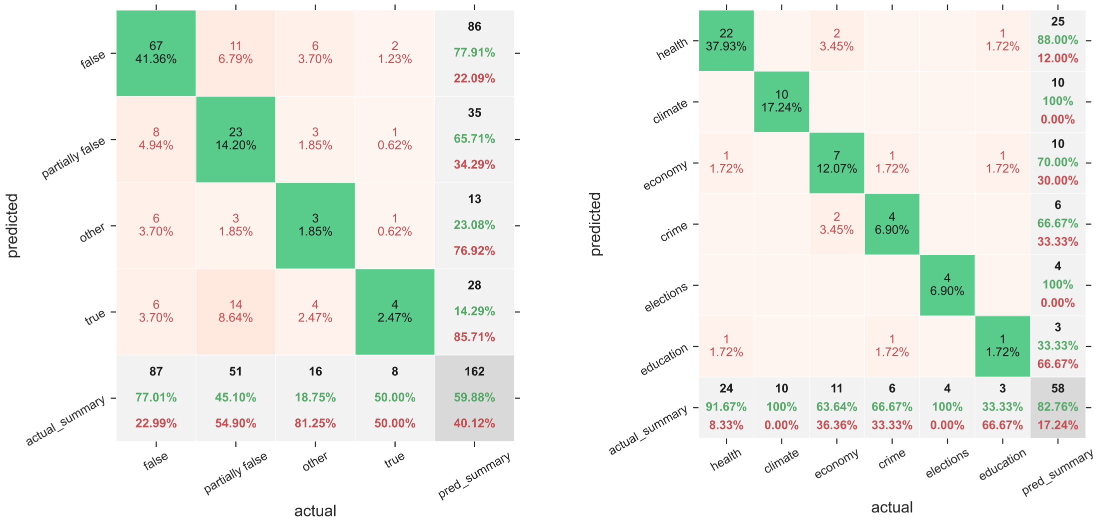
Target Encoding We encode the target labels using a multi-label binarizer as an analog of one-hot aka one-of-K scheme to multiple labels.

5.2. Results and Discussion

We participated in both text classification subtasks. Official evaluation results on the test set are presented in Table 3 and Table 4 for each subtask, respectively. We focused on suitable combinations of deep learning methods as well as their 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 and can serve as strong baseline models which, when fine-tuned, significantly outperform training models from scratch. The submission for each subtask is based on the best performing model checkpoint on the validation set as shown in Table 2.

When improving on the pretrained baseline models, class imabalance appears to be a primary challenge. This is clearly reflected in Figure 2, in particular, for the fake news detection subtask. The poor performance especially for the categories *true* and *other*, correlates with distribution of training data across these categories.

A commonly used tactic to deal with imbalanced datasets is to assign weights to each label.



(a) fake news detection

(b) topic domain identification

Figure 2: Confusion matrix for each subtask on the validation set.

Alternative solutions for coping with unbalanced datasets for supervised machine learning are undersampling or oversampling. Undersampling only considers a subset of an overpopulated class to end up with a balanced dataset. With the same goal oversampling creates copies of the unbalanced classes. Overfitting poses the most difficult challenge in this experiment, reducing its generalizability.

With the above findings, we achieve state of the art performance on the text classification datasets. RoBERTa has proven to be powerful language representation model for various natural language processing tasks. As the results of this study show, RoBERTa is also an effective tool for multi-class text classification. In the future, we will probe more insight of BERT on how it works and how to counteract its tendency to overfitting.

To further improve our the trained baseline model, we suggest to use Longformer[28] as a base model. Trained from RoBERTa[3], it addresses the problem of long sequences by replacing the attention matrices by sparse matrices, thus, allowing up to 4096 position embeddings.

6. Conclusion and Future work

In future work, we plan to investigate more recent neural architectures for language representation such as T5 [29] and GPT-3 [30].

Furthermore, we expect great opportunities for transfer learning from the areas such as argumentation mining [31] and offensive language detection [32]. To deal with data scarcity as a general challenge in natural language processing, we examine the application of concepts such as active learning, semi-supervised learning [33] as well as weak supervision [34].

References

[1] P. Nakov, G. Da San Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, N. Babulkov, A. Nikolov, G. K. Shahi, J. M. Struß, T. Mandl, The CLEF-2021 CheckThat! Lab on Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News, in: Proceedings of the 43rd European Conference on Information Retrieval, ECIR˜'21, Lucca, Italy, 2021, pp. 639–649. URL: https://link.springer.com/chapter/10.1007/978-3-030-72240-1{_}75.

[2] G. K. Shahi, J. M. Struß, T. Mandl, Overview of the CLEF-2021 CheckThat! Lab Task 3 on Fake News Detection, in: Working Notes of CLEF 2021—Conference and Labs of the Evaluation Forum, CLEF˜'2021, Bucharest, Romania (online), 2021.

[3] 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, 2019. arXiv:1907.11692.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, volume 2017-Decem, 2017, pp. 5999–6009. arXiv:1706.03762.

[5] X. Zhou, R. Zafarani, Fake News: A Survey of Research, Detection Methods, and Opportunities, ACM Comput. Surv 1 (2018). arXiv:1812.00315.

[6] C. H. de Vreese, F. Esser, T. Aalberg, C. Reinemann, J. Stanyer, Populism as an Expression of Political Communication Content and Style: A New Perspective, International Journal of Press/Politics 23 (2018) 423–438. URL: http://journals.sagepub.com/doi/10.1177/1940161218790035. doi:10.1177/1940161218790035.

[7] J. Leskovec, K. Lang, Statistical properties of community structure in large social and information networks, Proceedings of the 17th international conference on World Wide Web. ACM (2008) 695–704. URL: http://dl.acm.org/citation.cfm?id=1367591.

[8] P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep learning for hate speech detection in tweets, in: 26th International World Wide Web Conference 2017, WWW 2017 Companion, International World Wide Web Conferences Steering Committee, 2017, pp. 759–760. doi:10.1145/3041021.3054223. arXiv:1706.00188.

[9] L. Gao, R. Huang, Detecting online hate speech using context aware models, in: International Conference Recent Advances in Natural Language Processing, RANLP, volume 2017-Septe, Association for Computational Linguistics (ACL), 2017, pp. 260–266. doi:10.26615/978-954-452-049-6-036. arXiv:1710.07395.

[10] J. Pavlopoulos, P. Malakasiotis, I. Androutsopoulos, Deeper attention to abusive user content moderation, in: EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings, Association for Computational Linguistics, Stroudsburg, PA, USA, 2017, pp. 1125–1135. URL: http://aclweb.org/anthology/D17-1117. doi:10.18653/v1/d17-1117.

[11] G. K. Pitsilis, H. Ramampiaro, H. Langseth, Effective hate-speech detection in Twitter data using recurrent neural networks, Applied Intelligence 48 (2018) 4730–4742. doi:10.1007/s10489-018-1242-y. arXiv:1801.04433.

[12] Z. Zhang, D. Robinson, J. Tepper, Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network, in: Lecture Notes in Computer Science (including sub-

series Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 10843 LNCS, Springer Verlag, 2018, pp. 745–760. doi:10.1007/978-3-319-93417-4_48.

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

[14] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, Q. V. Le, XLNet: Generalized Autoregressive Pretraining for Language Understanding, Technical Report, 2019. arXiv:1906.08237.

[15] T. Mikolov, Q. V. Le, I. Sutskever, Exploiting Similarities among Languages for Machine Translation (2013). arXiv:1309.4168.

[16] J. Pennington, R. Socher, C. D. Manning, GloVe: Global vectors for word representation, in: EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 2014, pp. 1532–1543. doi:10.3115/v1/d14-1162.

[17] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, in: 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference, volume 2, 2017, pp. 427–431. URL: https://github.com/facebookresearch/fastText. doi:10.18653/v1/e17-2068. arXiv:1607.01759.

[18] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep Contextualized Word Representations, Association for Computational Linguistics (ACL), 2018, pp. 2227–2237. doi:10.18653/v1/n18-1202. arXiv:1802.05365.

[19] M. Castelle, The Linguistic Ideologies of Deep Abusive Language Classification, 2019, pp. 160–170. doi:10.18653/v1/w18-5120.

[20] P. Nakov, G. Da San Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, N. Babulkov, A. Nikolov, G. K. Shahi, J. M. Struß, T. Mandl, S. Modha, M. Kutlu, Y. S. Kartal, Overview of the CLEF-2021 CheckThat! Lab on Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News, in: Proceedings of the 12th International Conference of the CLEF Association: Information Access Evaluation Meets Multiliguality, Multimodality, and Visualization, CLEF˜’2021, Bucharest, Romania (online), 2021.

[21] G. K. Shahi, J. M. Struß, T. Mandl, Task 3: Fake News Detection at CLEF-2021 CheckThat!, CLEF˜’2021, Zenodo, Bucharest, Romania (online), 2021. doi:10.5281/zenodo.4714517.

[22] G. K. Shahi, AMUSED: An Annotation Framework of Multi-modal Social Media Data (2020). arXiv:2010.00502.

[23] 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. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le 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. arXiv:1910.03771v5.

[24] 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. arXiv:1912.01703.

[25] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Ł. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, J. Dean, Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation (2016). arXiv:1609.08144.

[26] D. Bahdanau, K. H. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, International Conference on Learning Representations, ICLR, 2015. arXiv:1409.0473.

[27] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: 32nd International Conference on Machine Learning, ICML 2015, volume 3, International Machine Learning Society (IMLS), 2015, pp. 2048–2057. arXiv:1502.03044.

[28] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The Long-Document Transformer (2020). arXiv:2004.05150.

[29] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, arXiv 21 (2019) 1–67. arXiv:1910.10683.

[30] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020. arXiv:2005.14165.

[31] M. Stede, Automatic argumentation mining and the role of stance and sentiment, Journal of Argumentation in Context 9 (2020) 19–41. URL: https://www.jbe-platform.com/content/journals/10.1075/jaic.00006.ste. doi:10.1075/jaic.00006.ste.

[32] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, Predicting the type and target of offensive posts in social media, in: NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, volume 1, Association for Computational Linguistics, Stroudsburg, PA, USA, 2019, pp. 1415–1420. URL: http://aclweb.org/anthology/N19-1144. doi:10.18653/v1/n19-1144. arXiv:1902.09666.

[33] S. Ruder, B. Plank, Strong Baselines for Neural Semi-supervised Learning under Domain Shift, ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers) 1 (2018) 1044–1054. arXiv:1804.09530.

[34] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, C. Ré, Snorkel: rapid training data creation with weak supervision, in: VLDB Journal, volume 29, Springer, 2020, pp. 709–730. doi:10.1007/s00778-019-00552-1.