UniBO at CheckThat! 2024: Multi-lingual and Multi-label Persuasion Technique Detection in News with Data Augmentation and Sequence-Token Classifiers

Notebook for the CheckThat! Lab at CLEF 2024

Paolo Gajo¹, Luca Giordano¹ and Alberto Barrón-Cedeño¹

¹DIT, Alma Mater Studiorum - Università di Bologna. Forlì, Italy

Abstract

With the widespread use of the Internet and the rise of algorithmic journalism, consumers of news are exposed more than ever before to manipulative, propagandistic, and deceptive content. As a result, major public events and debates on relevant topics can be significantly influenced. This creates an increasing demand for automated tools that help experts analyze the news ecosystem. We explored persuasion technique detection in multi-lingual news as part of the CheckThat! Lab Task 3. Our pipeline comprises two parts. The first part is a data augmentation module, which uses a BERT-based model fine-tuned for word-alignment to project labels from source texts to machine-translated target texts. The second one is a persuasion technique classification module, leveraging two fine-tuned BERT-based models: a sequence classifier for detecting sentences containing persuasion techniques and a set of 23 token-level classifiers for specific techniques. Our approach, trained on augmented multilingual data with class weighting and a high decision threshold of 0.9, is competitive in all language settings, showing hints of cross-lingual transfer. Despite the research efforts in this direction, exemplified by shared tasks, detecting persuasion techniques, especially across languages, remains challenging due to their implicit and subtle nature.

Keywords

persuasion techniques, multi-lingual, data augmentation, class weighting, decision threshold

1. Introduction

Media language and news discourse have always attracted the attention of applied linguists and sociolinguists, mainly because of four reasons [1]: 1) the media provide an easily accessible source of language data for research and teaching purposes, 2) the media are important linguistic institutions, and their language usage reflects and shapes both language use and attitudes in a speech community, 3) the ways in which the media use language are interesting linguistically in their own right, and 4) the media are important social institutions. They are crucial presenters of culture, politics, and social life, shaping as well as reflecting how these are formed and expressed.

With the widespread use of the Internet and the rise of algorithmic journalism [2, 3, 4, 5, 6], characterized by huge amounts of data, the application of algorithms in all phases of the journalistic process (selection, production, distribution and consumption) and by a high degree of interactivity and direct communication between news producers and consumers, the latter are exposed more than ever before to manipulative, propagandistic, and deceptive content. As a result, major public events and debates on important topics can be significantly influenced. This creates an increasing demand for automated tools that help experts analyze the news ecosystem, detect manipulation attempts, and aid in studying how events, global issues, and policies are portrayed by the media in different countries and languages. There has been a growing interest of the NLP community in trying to detect the use of specific propaganda techniques, as well as the specific span of each instance. This interest is mainly expressed by the organization of shared tasks, including NLP4IF-2019 on Fine-Grained Propaganda Detection [7], SemEval-2020 task 11 on Detection of Persuasion Techniques in News Articles [8], SemEval-2021 task 6 on Detection

paolo.gajo2@unibo.it (P. Gajo); luca.giordano16@studio.unibo.it (L. Giordano); a.barron@unibo.it (A. Barrón-Cedeño)
 https://www.unibo.it/sitoweb/paolo.gajo2 (P. Gajo); https://www.linkedin.com/in/giordanoluca00/ (L. Giordano); https://www.unibo.it/sitoweb/a.barron (A. Barrón-Cedeño)

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

^{© 2024} Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).



Figure 1: Persuasion technique taxonomy.

of Persuasion Techniques in Texts and Images [9], WANLP-2022 Shared Task on Propaganda Detection in Arabic [10], and, most recently, SemEval-2023 Task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup [11]. Task 3 "Persuasion Techniques" of the CheckThat! Lab 2024 [12] is a further effort to advance the state of the art in this research direction.

Participants in the Task 3 "Persuasion Techniques" of the CheckThat! Lab 2024 are given a set of news articles in multiple languages and a list of 23 persuasion techniques (PT), including logical fallacies (e.g., straw man, red herring, bandwagon) and emotional manipulation techniques (e.g., loaded language, appeal to fear, name calling) that might be used to support flawed argumentation. The aim is to identify the spans of texts in which each technique occurs. Text spans assigned with labels might also overlap. Therefore, it is set up as a multi-label sequence-tagging task where each sequence can be assigned more than one class. The evaluation metric is micro-averaged F_1 , modified to account for partial matching between the spans. Furthermore, annotation guidelines were provided by the task organizers, which contain detailed definitions and examples [13].

Our system can be divided in two parts. The first part of the pipeline comprises the data augmentation module: a BERT-based model [14] fine-tuned on a word-alignment task. We use this to project the labels from the source to the target texts, produced by translating the training set into different languages via machine translation (MT). The second part of the pipeline refers to the persuasion technique classification module, i.e. two separate BERT-based models, henceforth referred to as M_1 and M_2 . M_1 is a binary sequence classifier, trained to classify individual sentences as containing a persuasion technique or not. M_2 , is a series of 23 token-level classifiers, $M_{2,1}, \ldots, M_{2,23}$, one per persuasion technique. Leveraging our multilingual MT data augmentation strategy, we trained a single set of multilingual models and used them to infer on all languages of a holdout validation set and of the test set. Accordingly, we submitted runs for all test languages. Our system is competitive in all language settings, showing hints of cross-lingual transfer when training on multi-lingual data and testing on unseen languages. For reproducibility purposes, we release our code and data on our GitHub repository.¹

The rest of the paper is organized as follows. Section 2 provides an overview of related work on persuasion techniques and propaganda detection. In Section 3, we describe the training data provided for the shared task and our data augmentation process. In Section 4 we present our proposed system. Section 5 reports the performance of the system for our official run. Finally, we conclude in Section 6.

¹https://github.com/giorluca/checkthat24_DIT

Language	docs	sentences	Language	docs	sentences
English	536	24,514	Italian	303	7,917
French	211	6,298	Polish	194	9,612
German	177	5,667	Russian	191	5,900
			Total	1,612	59,908

 Table 1

 Statistics of the training data in terms of number of documents and sentences.

2. Related Work

Research on automatic persuasion technique detection in news overlaps to a large extent with work on automatic propaganda detection in news [15, 16, 17]. Early research on propaganda detection focused exclusively on document-level analysis, ignoring the fine-grained aspects of the task.

Rashkin et al. [15] developed the TSHP-17 corpus, annotated in a distant supervised manner (i.e. assigning the label of the news outlet to all articles gathered from that news outlet) at document-level with four classes: *trusted, satire, hoax,* and *propaganda*. However, as can be deduced from the results obtained in the experiment, further verified for reproducibility by Barrón-Cedeño et al. [16] and mentioned by Da San Martino et al. [18], the predictive model trained on this data (logistic regression with *n*-gram representation) failed to generalize, performing well only on articles from sources that the system was trained on and under-performing when evaluated on articles from unseen news sources.

Barrón-Cedeño et al. [16] developed the QProp corpus annotated at document-level with two labels (*propaganda* vs *non-propaganda*) and trained different models (e.g., logistic regression and SVMs) on this data and on the TSHP-17 corpus to predict the two classes, including linguistic features such as writing style and readability indices. Their findings confirmed that using distant supervision might introduce bias in the model and lead to predict the source of the article, rather than to discriminate propaganda from non-propaganda, independently from the news source.

An alternative line of research has focused on detecting the use of specific propaganda techniques in text [19, 20, 17]. Habernal et al. [19, 20] developed a corpus with 1.3k arguments annotated with five fallacies that directly relate to propaganda techniques. A more fine-grained analysis was done by Da San Martino et al. [17], who developed a corpus of news articles annotated with 18 propaganda techniques which was used to train a gated deep neural network for sentence-level propaganda detection. For a survey on computational propaganda detection see Da San Martino et al. [18].

3. Data

The training data provided for the task is an existing corpus, consisting of 1,612 news articles in 9 languages annotated with 48K instances of 23 persuasion techniques [11]. The persuasion technique taxonomy is pictured in Figure 1. A new test dataset of around 500 news articles in Arabic, Bulgarian, English, Portuguese, and Slovene is provided for this edition [12]. The distribution of training data by language is provided in Table 1.

3.1. Data Preprocessing

While inspecting the datasets in the preliminary stages, we noticed that the gold annotations would sometimes not coincide with the character slices, once visualized as Python strings. This was caused by the fact that some files were saved using carriage return along with the newline character (\r\n), while others only contained newline characters (\n). Once read in Python using the 'r' reading mode, this meant that in files containing '\r\n' the newlines counted as only one character, instead of two; conversely, in those only using '\n' the newlines correctly counted as only one character. In order to solve this issue, we read all files in 'rb' binary mode and decoded prior to data processing, so that \r characters would be correctly counted in order to feed models the correct text spans.

Since the documents are too long to feed in input to the hereby-used models,² we generate smaller training samples by splitting documents at the sentence level, obtaining 59,908 total gold training sentences, as indicated in Table 1. Prior to training, we split the obtained sentence dataset 80/20 into training and validation instances.

3.2. Data Augmentation

In order to increase the amount of available training data we augment the training sentences via MT and label projection. MT is carried out by translating the dataset sentence by sentence from English to the other training languages with the NLLB 3.3B model [21].

Following Nagata et al. [22], the PT annotations are then projected onto the translated text by using mDeBERTa models [23] trained on a word-alignment task with a question-answering classifier head. Given a source sentence A with characters $a_i \in A$, and its translated target sentence B with characters $b_j \in B$, and an alignment between spans $a_{i,i+k}$, labeled as C, and $b_{j,j+l}$, with $i < j \in \mathbb{N}$, $k, l > 0 \in \mathbb{N}$, label projection is the task of assigning the label C to the span $b_{j,j+l}$ [24]. In other words, given a source span, the model is tasked to find the equivalent span in the translated text. Doing this, we obtain synthetic annotated data in the target language, which we use to further train both our sequence and token classifiers. In order to train these word-alignment models, we use XL-WA [25], a multilingual word-alignment dataset [25] built from WikiMatrix [26].³ The dataset has a balanced domain distribution and features 14 EN-XX language combinations. Its training set is composed of silver labels automatically generated, while the development and test sets are manually annotated. We align each source-target combination of machine-translated data (EN-IT, EN-ES, EN-RU, EN-SL, EN-BG, EN-PT), where English is always the source gold data, with a different word-alignment model, trained on the specific language combination contained in XL-WA.

Ultimately, departing from the original 24,514 gold English sentences indicated in Table 1, we generate the same amount for each of the six target languages, for a total of 147,084 extra training sentences. Thus, the total number of training instances amounts to 206,992.

4. Models

4.1. Word Alignment

For the word alignment task, used in the data augmentation step, we adopt the approach proposed by Nagata et al. [22], which treats word alignment as a question answering task using an mDeBERTa [23] model. In this approach, the source word to be aligned is enclosed within rarely used characters, such as ' \bullet ', and the model is fed both the source sequence A and the target sequence B simultaneously. The input to the model at the token level is structured as follows:

$$[CLS]\alpha_1, \dots, tok(\bullet), \alpha_i, \dots, \alpha_{i+k}, tok(\bullet), \dots, \alpha_m$$
$$[SEP]\beta_1, \dots, \beta_j, \dots, \beta_{j+l}, \dots, \beta_n \quad [SEP]$$

Here, the source word to be aligned is represented by the tokens $\alpha_i, \ldots, \alpha_{i+k}$, where $\alpha_i \in tok(A)$. The model is then tasked with predicting the tuple (β_j, β_{j+l}) , where $\beta_j \in tok(B)$, which denotes the boundary indices of the aligned word in the target sequence.

For each language combination involved in the data augmentation process, we train our models for up to 3 epochs on each of XL-WA's languages with a batch size of 16. The optimizer's learning rate is set to 3×10^{-4} , and ϵ is 10^{-8} . We select the best model based on the Exact metric *E* [27]:

$$E = \frac{\sum_{i}^{n} exact(p_i, g_i)}{\|preds\|} , \qquad (1)$$

²Since the models we use are derived from BERT base, they can only handle 512 tokens in input.

³https://ai.meta.com/blog/wikimatrix/

where preds is a list of predictions and $exact(p_i, g_i)$ is the Kronecker delta:

$$exact(p_i, g_i) = \begin{cases} 1, & \text{if } p_i = g_i, \\ 0, & \text{if } p_i \neq g_i. \end{cases}$$
(2)

Before computing E, we lowercase and strip the predicted and gold strings p_i and g_i of excess punctuation and spacing.

4.2. Persuasion Technique Classification

The model we use is composed of two separate Transformer networks [28], both based on mDe-BERTa [23]. The first part of the model, M_1 , is a binary sequence classifier, trained to classify individual sentences as containing any persuasion technique or not. The second part of the model, M_2 , is a series of 23 token-level classifiers, $M_{2,1}, \ldots, M_{2,23}$, one per persuasion technique.

Sequence classifier Upon training, we feed the sequence classifier a balanced subsample of the sentence dataset, obtained as per Section 3. Specifically, we take all sentences containing at least one PT (positive instances) and sample an equal number of negative instances from the rest of the training set.

Token classifiers Prior to training for token classification, we preprocess and label the data using the BIO annotation scheme [29]. In this scheme, the first word of an entity is assigned a B-{class} (beginning) label, subsequent words are assigned an I-{class} (inside) label, and words not part of any entity are assigned an O (outside) label. We follow established methodology by ignoring subword tokens when calculating cross-entropy loss.⁴

Since the 23 token classification models are tailored specifically to each PT, we train them on sentences where only one PT is kept at a time. This means that if a sentence contains a persuasion technique which the model is not supposed to learn to predict, we set the tokens relative to that persuasion technique to the outside 0 label. Just like for the sequence classifier, we balance positive and negative instances for training.

For both M_1 and M_2 , we set the optimizer's learning rate at 5×10^{-5} , while ϵ is 10^{-8} . We train all models for up to 10 epochs with a patience of 2 epochs, keeping the model with the highest performance on the validation set.

4.3. Reducing False Positives

As we are using 23 separate token classifiers, the number of predictions being produced ends up being very large. Since the submission website only accepts TXT files of up to 800 KBytes and our token classifiers produce too many predictions, our full outputs are not suitable for submission in most languages. As such, we opt for reducing the number of positive predictions in order to adhere to the submission size limit. To accomplish this, during training we use a modified, weighted version of the cross-entropy loss function. Specifically, we empirically assign a weight of 0.5 to the 0 majority class (label 0) and a weight of 2.0 to the minority B and I classes (labels 1 and 2). This weighting ensures that the model pays more attention to correctly predicting the minority classes, thus reducing the overall number of positive predictions.

When computing the evaluation metrics, we also apply a threshold to the model's predictions. We use the softmax function to convert the logits into probabilities. Then, we set any probability below the threshold of 0.9 to zero before determining the predicted labels.⁵ This means that the model only makes a prediction if it is at least 90% confident, reducing the number of false positives. We did not experiment with any other parameters, besides function loss weights and the prediction threshold. In addition, since some of the token classifiers obtained an F_1 score of 0 on their class subset of the validation partition obtained from splitting the training set, we exclude the predictions produced by those models.

⁴https://huggingface.co/docs/transformers/en/tasks/token_classification

⁵We attempted different thresholds by increments of 0.1 until the submission files were small enough for submission.

Table 2

Results obtained by the binary M_1 mDeBERTa classifier on the holdout validation set obtained from the 80/20 split of the training set. The reported scores are achieved on all languages of the holdout validation set at once.

Class	Р	R	\mathbf{F}_1
0	0.802	0.662	0.726
1	0.734	0.850	0.788
Macro AVG	0.768	0.756	0.757

Table 3

Average results obtained by the 23 mDeBERTa token classifiers on the holdout validation set from the training set (80/20 split). The macro F₁ scores are achieved on all languages of the holdout validation set at once.

Data Aug.	Class Weighting	Threshold	macro F_1	
			0.140	
1			0.317	
1	\checkmark	0.9	0.192	

4.4. Inference

During inference, we produce the submission predictions following a series of steps. First, after models M_1 and M_2 have produced their predictions, we set M_2 's predictions to 0-tensors for those indices where M_1 's predictions are 0. Then, we binarize the predictions to {0, 1}, with the original {1, 2} labels mapping onto 1, and 0 mapping onto 0. Lastly, we assign a character span to each consecutive series of positive (1) predictions in the prediction tensor, based on the characters corresponding to each token.

5. Results

In this section, we first report and discuss the results obtained by our model M_1 on a holdout validation set which contains all training languages, obtained by splitting the training set 80/20 into training and validation data with a set seed (Table 2). Then, we report the results of our 23 token classifiers M_2 on the same holdout validation set, given different training settings (Table 3). We report these results with the intent of providing insight into the contribution of the data augmentation process to the performance of the second classifier and to show how it is affected by class weighting and by setting a decision threshold to reduce false positives. Finally, we present the results achieved by our whole system on the official test set (Table 4). The binary classifier M_1 performs decently, with a macro F_1 of 0.757. Furthermore, as shown by the reported scores in Table 3, the data augmentation more than doubles M_2 's performance. On the other hand, class weighting and setting a decision threshold as high as 0.9, although necessary as shown above, lowers the best performance by 0.12. Since these preliminary tests conducted on the holdout validation split show that data augmentation improves the performance of the M_2 classifier, even when class weighting and a decision threshold of 0.9 are set, we choose to adopt data augmentation also for the final system used to predict on the test set for submission. The rationale behind this decision is based on the assumption that a higher performance on the holdout validation set would also translate onto the test set.⁶

The results for our official test runs are shown in Table 4. Our system performs better than the baseline⁷ across all languages and ranks first for all languages except for Arabic. For that language, Team Mela used a multilingual BERT model which was pre-trained on data in both English and Arabic [30]. We also observe consistent performance across all languages, with micro average F_1 scores ranging from 0.091 for English to 0.123 for Slovene, possibly showing a robust cross-lingual transfer ability when training the model on multi-lingual data and testing it on unseen languages.

⁶Note that our official submission (last row in Table 3) is not the best because it is constrained by the maximum size accepted by the submission website for the produced prediction file. Indeed, the used class weights and prediction threshold are applied in order to reduce the amount of predictions produced by the model.

⁷A token classification model followed by heuristics that was kept private by the organizers [12].

Table 4

Official leaderboard results for all five languages, both in terms of micro and macro F_1 . Best performance in bold. We include the two official participant runs (UniBO and Mela) and the runs by the task organizers (PersuasionMultiSpan and Baseline).^{*a*}

Rank Team	F1 micro	F1 macro	Rank Team	F1 micro	F1 macro	
English			Portuguese			
1 UniBO	0.092	0.061	 PersuasionMultiSpar 	0.132	0.120	
– PersuasionMultiSpan	0.078	0.086	1 UniBO	0.107	0.073	
2 Baseline	0.009	0.001	2 Baseline	0.002	0.000	
Bulgarian			Slovenian			
- PersuasionMultiSpan	0.132	0.128	 PersuasionMultiSpar 	0.153	0.127	
1 UniBO	0.114	0.081	1 UniBO	0.123	0.075	
2 Baseline	0.009	0.002	2 Baseline	0.003	0.002	
Arabic						
1 Mela	0.301	0.080				
2 UniBO	0.108	0.068				
– PersuasionMultiSpan	^a 0.028	0.059				
3 Baseline ^b	0.021	0.006				

^{*a*} PersuasionMultiSpan is an XML-RoBERTa [31] model trained on the SemEval 2023 corpus [11]. It was announced by the organizers only after the run submission deadline [32].

6. Conclusions

In this paper, we presented our experiments and findings on persuasion technique detection in news in multiple languages, which was part of the CheckThat! Lab Task 3. Our system is divided in two parts. The first part of the pipeline comprises data augmentation: a combination of machine translation and cross-language label projection. The second part of the pipeline refers to the persuasion technique classification module, comprising two separate BERT-based models. The first acts as a binary sequence classifier, trained to classify individual sentences as containing a persuasion technique or not. The second comprises 23 token-level classifiers, one per persuasion technique.

We submitted runs for five all test languages. Our final model, trained on the shared task's training data augmented via MT and tuned with class weighting and a high decision threshold of 0.9, is competitive in all language settings and shows hints of cross-lingual transfer capabilities when trained on multi-lingual data and testing on unseen languages.

Automatically detecting persuasion techniques in news in a multi-lingual setting still proves to be a challenging task, given how implicitly such techniques manifest and how subtle their usage can be. This leaves space for future work. For example, investigating the existence of more explicit predictive features, including but not limited to textual and linguistic indicators or online dissemination patterns, that prove to be typical of propagandistic news, independently of news source and common across different languages, which could be leveraged alongside transformer models and data augmentation.

References

- [1] A. Bell, Language and the media, Annual review of applied linguistics 15 (1995) 23-41.
- [2] C. W. Anderson, Towards a sociology of computational and algorithmic journalism, New media & society 15 (2013) 1005–1021.
- [3] M. Coddington, Clarifying journalism's quantitative turn: A typology for evaluating data journalism, computational journalism, and computer-assisted reporting, Digital journalism 3 (2015) 331–348.
- [4] A. Graefe, Guide to automated journalism, Tow Center for Digital Journalism Publications, Columbia University, 2016.
- [5] N. Thurman, Personalization of news, in: T. P. Vos, F. Hanusch, D. Dimitrakopoulou, M. Geertsema-

Sligh, A. Sehl (Eds.), The International Encyclopedia of Journalism Studies, Wiley-Blackwell, Massachusetts, USA, 2018.

- [6] J. M. Túñez-López, C. Toural-Bran, A. G. Frazão-Nogueira, From data journalism to robotic journalism: The automation of news processing, Journalistic metamorphosis: media transformation in the digital age (2020) 17–28.
- [7] G. Da San Martino, A. Barron-Cedeno, P. Nakov, Findings of the NLP4IF-2019 shared task on fine-grained propaganda detection, in: Proceedings of the second workshop on natural language processing for internet freedom: censorship, disinformation, and propaganda, 2019, pp. 162–170.
- [8] G. Da San Martino, A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, P. Nakov, SemEval-2020 task 11: Detection of propaganda techniques in news articles, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020, pp. 1377–1414.
- [9] D. Dimitrov, B. B. Ali, S. Shaar, F. Alam, F. Silvestri, H. Firooz, P. Nakov, G. Da San Martino, SemEval-2021 task 6: Detection of persuasion techniques in texts and images, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), 2021, pp. 70–98.
- [10] F. Alam, H. Mubarak, W. Zaghouani, G. Da San Martino, P. Nakov, Overview of the WANLP 2022 shared task on propaganda detection in arabic, in: Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), 2022, pp. 108–118.
- [11] J. Piskorski, N. Stefanovitch, G. Da San Martino, P. Nakov, 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), 2023, pp. 2343–2361.
- [12] A. Barrón-Cedeño, F. Alam, T. Chakraborty, T. Elsayed, P. Nakov, P. Przybyła, J. M. Struß, F. Haouari, M. Hasanain, F. Ruggeri, et al., The CLEF-2024 CheckThat! Lab: Check-Worthiness, Subjectivity, Persuasion, Roles, Authorities, and Adversarial Robustness, in: European Conference on Information Retrieval, Springer, 2024, pp. 449–458.
- [13] J. Piskorski, N. Stefanovitch, V.-A. Bausier, N. Faggiani, J. Linge, S. Kharazi, N. Nikolaidis, G. Teodori, B. De Longueville, B. Doherty, et al., News categorization, framing and persuasion techniques: Annotation guidelines, European Commission, Ispra, JRC132862 (2023).
- [14] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.
- [15] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, Y. Choi, Truth of varying shades: Analyzing language in fake news and political fact-checking, in: Proceedings of the 2017 conference on empirical methods in natural language processing, 2017, pp. 2931–2937.
- [16] A. Barrón-Cedeño, I. Jaradat, G. Da San Martino, P. Nakov, Proppy: Organizing the news based on their propagandistic content, Information Processing & Management 56 (2019) 1849–1864.
- [17] G. Da San Martino, Y. Seunghak, A. Barrón-Cedeno, R. Petrov, P. Nakov, et al., Fine-grained analysis of propaganda in news article, in: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), Association for Computational Linguistics, 2019, pp. 5636–5646.
- [18] G. Da San Martino, S. Cresci, A. Barrón-Cedeño, S. Yu, R. Di Pietro, P. Nakov, A survey on computational propaganda detection, in: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, 2021, pp. 4826–4832.
- [19] I. Habernal, R. Hannemann, C. Pollak, C. Klamm, P. Pauli, I. Gurevych, Argotario: Computational argumentation meets serious games, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2017, pp. 7–12.
- [20] I. Habernal, P. Pauli, I. Gurevych, Adapting serious game for fallacious argumentation to german: Pitfalls, insights, and best practices, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.
- [21] N. Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi,

J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, J. Wang, M. Ai, No Language Left Behind: Scaling Human-Centered Machine Translation, 2022.

- [22] M. Nagata, K. Chousa, M. Nishino, A supervised word alignment method based on cross-language span prediction using multilingual BERT, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 555–565.
- [23] P. He, J. Gao, W. Chen, DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing, in: The Eleventh International Conference on Learning Representations, 2022.
- [24] A. Jain, B. Paranjape, Z. C. Lipton, Entity projection via machine translation for cross-lingual NER, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 1083–1092. URL: https://aclanthology.org/D19-1100. doi:10.18653/v1/D19-1100.
- [25] F. Martelli, A. S. Bejgu, C. Campagnano, J. Čibej, R. Costa, A. Gantar, J. Kallas, S. Koeva, K. Koppel, S. Krek, M. Langemets, V. Lipp, S. Nimb, S. Olsen, B. S. Pedersen, V. Quochi, A. Salgado, L. Simon, C. Tiberius, R.-J. Ureña-Ruiz, R. Navigli, XL-WA: a Gold Evaluation Benchmark for Word Alignment in 14 Language Pairs, in: F. Boschetti, N. N. Gianluca E. Lebani, Bernardo Magnini (Eds.), Proceedings of the Ninth Italian Conference on Computational Linguistics (CLiC-it 2023), volume 3596, CEUR-WS, Venice, Italy, 2023.
- [26] H. Schwenk, V. Chaudhary, S. Sun, H. Gong, F. Guzmán, WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia, in: P. Merlo, J. Tiedemann, R. Tsarfaty (Eds.), Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 1351–1361.
- [27] P. Rajpurkar, R. Jia, P. Liang, Know what you don't know: Unanswerable questions for SQuAD, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 784–789.
- [28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is All you Need, in: Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017.
- [29] L. A. Ramshaw, M. P. Marcus, Text Chunking Using Transformation-Based Learning, Springer Netherlands, Dordrecht, 1999, pp. 157–176.
- [30] S. Nabhani, M. A. R. Riyadh, Mela at CheckThat! 2024: Transferring persuasion detection from English to Arabic a multilingual BERT approach, in: [33], 2024.
- [31] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451.
- [32] J. Piskorski, N. Stefanovitch, F. Alam, R. Campos, D. Dimitrov, A. Jorge, S. Pollak, N. Ribin, Z. Fijavž, M. Hasanain, N. Guimarães, A. F. Pacheco, E. Sartori, P. Silvano, A. V. Zwitter, I. Koychev, N. Yu, P. Nakov, G. Da San Martino, Overview of the CLEF-2024 CheckThat! lab task 3 on persuasion techniques, in: [33], 2024.
- [33] G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CLEF 2024, Grenoble, France, 2024.