DSHacker at SemEval-2023 Task 3: Genres and Persuasion Techniques Detection with Multilingual Data Augmentation through Machine Translation and Text Generation

Arkadiusz Modzelewski * **Witold Sosnowski** *
Magdalena Wilczynska **Adam Wierzbicki**

Polish-Japanese Academy of Information Technology
Koszykowa 86, 02-008 Warsaw, Poland
firstname.lastname@pja.edu.pl

Abstract

In our article, we present the systems developed for SemEval-2023 Task 3, which aimed to evaluate the ability of Natural Language Processing (NLP) systems to detect genres and persuasion techniques in multiple languages. We experimented with several data augmentation techniques, including machine translation (MT) and text generation. For genre detection, synthetic texts for each class were created using the OpenAI GPT-3 Davinci language model. In contrast, to detect persuasion techniques, we relied on augmenting the dataset through text translation using the DeepL translator. Fine-tuning the models using augmented data resulted in a top-ten ranking across all languages, indicating the effectiveness of the approach. The models for genre detection demonstrated excellent performance, securing the first, second, and third positions in Spanish, German, and Italian, respectively. Moreover, one of the models for persuasion techniques' detection secured the third position in Polish. Our contribution constitutes the system architecture that utilizes DeepL and GPT-3 for data augmentation for the purpose of detecting both genre and persuasion techniques.

1 Introduction

The ability to detect genres [1] and persuasion techniques in the text is crucial in NLP, as it allows to better evaluate the quality of information and categorize it. The SemEval-2023 [2] Task 3 is a shared task that aims to evaluate the performance of systems in detecting genres, framing, and persuasion techniques in multiple languages. This paper presents our systems for detecting genre and persuasion techniques[3], which was prepared for that workshop.

Our methods involve machine translation and text generation techniques to augment the original training data in all participating languages. Specifically, we augmented the genre detection datasets in these languages by creating synthetic texts for each class with the OpenAI GPT-3 Davinci language model. Afterward, we fine-tuned a single multilingual XLM-RoBERTa-large model for genre detection for all languages. Our approach to building systems to detect different persuasion techniques involved a combination of feature engineering and deep learning models. We primarily utilized the DeepL [4] translator for text translation to augment the dataset. Our final systems performed well in the SemEval-2023 Task 3, ranking among the top 10 participants across all languages.

We structured our paper as follows: In Section 2, we give a summary of prior work in the field of detecting genres and persuasion techniques. Sections 3 and 4 describe the methods we used for data augmentation and our systems architecture. Sections 5 and 6 present our experimental setup and the results we obtained. Section 7 presents a multilingual persuasion detection system that outperforms all other systems in that task, submitted after the original deadline. Finally, Sections 8 and 9 present our conclusions, limitations of our approach, and potential avenues for future research.

2 Related Work

2.1 Genre detection

The genre detection task described as distinguishing between opinion, reporting, and satirical news

[*] Equal contribution.

[1] By genre we mean satire, opinion and reporting

[2] The 17th International Workshop on Semantic Evaluation. https://semeval.github.io/SemEval2023/

[3] SemEval-2023 Task 3 included 3 subtasks. Subtask 1 (ST1) was about genre detection and the aim of subtask 3 (ST3) was to detect persuasion techniques

[4] https://www.deepl.com/translator

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1582–1591
July 13-14, 2023 ©2023 Association for Computational Linguistics

is not widely covered in the literature. However, several publications closely examine this problem. Regarding detecting satirical news, earlier studies employed machine learning methods, such as Support Vector Machines (SVMs), and handcrafted features derived from factual and satirical news headlines and body text, such as bag-of-words, n-grams, and lexical features (Rubin et al., 2016). In contrast, recent studies utilize deep learning approaches to automatically extract features for satire detection. For instance, Yang et al. (2017) proposed a hierarchical model with an attention mechanism and manually created linguistic features to comprehend satire at both the paragraph and article levels. In their work on satire detection, Li et al. (2020) proposed a method that incorporates text and image modalities using the ViLBERT model (Lu et al., 2019). Yang et al. (2017) investigated using paragraph-level linguistic features to identify satire by utilizing a neural network classifier with attention. They compared the effectiveness of paragraph-level features and document-level features on a large English satirical news dataset. In their genre classification task, Haider and Palmer (2017) employed various classification algorithms, namely Linear Discriminant Analysis (LDA), a Naive Bayes Multinominal classifier, Random Forest ensemble classifiers (FOREST), and SVM. Their findings suggest that the FOREST classifier is the most reliable when facing feature changes. However, the best performance is achieved by SVM on the LDA with 200 topic dimensions on content words. In their study on German genre classification, Kim et al. (2017) selected two models to examine their method based on emotional words and three models for the emotion arc-based approach: a random forest classifier (RF), a multi-layer perceptron (MLP), and a CNN. They assessed the performance of their models using a micro-average F1 score. For the original BoW feature set, RF outperforms MLP, whereas it is the opposite for the emotional word-based method. RF and MLP yield equal performance for the emotion arc-based approach, which CNN slightly surpasses. Pei (2022) investigated the value of various semantic and stylistic features for classifying news genres. The study used four models to test the effectiveness of the features and found that genre-exclusive words and synonyms were the most beneficial. In contrast, emotional words had a detrimental effect. The best outcome was a macro-average F1 score, precision, and recall of 0.97 with a feature set that combined preprocessed data and synonym sets classified by the Complement Naive Bayes model.

2.2 Persuasion techniques detection

Analyzing news articles in the context of propaganda was advocated by Rashkin et al. (2017) and Barrón-Cedeno et al. (2019). Rashkin et al. (2017) proposed Long Short-Term Memory (LSTM) model that takes the sequence of words as the input and predicts the Politifact[5] rating. In addition, they compared the language of real news with that of satire, hoaxes, and propaganda to find linguistic characteristics of untrustworthy text. Barrón-Cedeno et al. (2019) proposed *proppy* that was the first publicly available real-world, real-time propaganda detection system for online news. The importance of detecting persuasion techniques in propaganda content was pointed out by Da San Martino et al. (2019). The authors designed a novel multi-granularity neural network and performed fine-grained analysis of texts by detecting all fragments with propaganda techniques and their types.

Persuasion techniques detection was the subject of a similar task described by Da San Martino et al. (2020) during the International Workshop on Semantic Evaluation in 2020. Different approaches to the proposed task in 2020 were primarily based on utilizing large language BERT-based pre-trained models. The winning system by Jurkiewicz et al. (2020) utilized the RoBERTa model with class-dependent rescaling applied to binary cross-entropy and semi-supervised learning methodology. Team aschern proposed a solution with RoBERTa-CRF and transfer learning between two related subtasks in SemEval. Some systems incorporated linguistic features, and ensemble learning (Patil et al., 2020). Different approaches to data augmentation were also explored during SemEval 2020, for instance team WMD applied multiple strategies such as back translation, synonym replacement and TF.IDF replacement. TF.IDF replacement consisted of replacing unimportant words based on TF.IDF score by other unimportant words (Daval-Frerot and Weis, 2020).

Furthermore, it is worth to mention that Da San Martino et al. (2021) presented a survey that examine the state of the art on computational propaganda detection from the perspective of Nat-

[5] PolitiFact.com is an organization that actively investigate and rate the veracity of comments made by public figures, journalists, and organizations.

ural Language Processing and Network Analysis and argued the need for combined efforts between both communities.

3 Dataset

This section provides a brief overview of the data available for task 3 and the methods of data augmentation we used for the ST1 and ST3 in which we participated. As outlined by Piskorski et al. (2023), our task involved training and development datasets, which had known labels and a test dataset without labels. All of the datasets contained news articles. The training and development datasets comprised English, French, German, Italian, Polish, and Russian articles. The test dataset included articles in these languages and three additional languages, Spanish, Greek, and Georgian. The test dataset was used to make final predictions. For a more thorough description of the available data, refer to Piskorski et al. (2023).

3.1 Genre detection

3.1.1 Dataset description

In the genre detection task we had to create a model for the multi-class classification at the article level. The first two sets contained articles labeled with the corresponding genres: either satire, reporting, or opinion. The test set consisted of articles for which participants were required to predict the genre and upload the results.

3.1.2 Augmentation

We augmented the datasets in all base languages, namely English, French, German, Italian, Polish, and Russian, by using the Generative Pre-trained Transformer 3 (GPT-3) made available by OpenAI (Brown et al., 2020). Specifically, we utilized the Davinci version of GPT-3, which has a remarkable 175 billion trainable parameters (Brown et al., 2020), to generate news article examples for each genre automatically. For this purpose, we provided the GPT-3 model with the following prompts:

- For satire: *Write a funny satirical article of at least 350 words in <language>.*

- For opinion: *Write an opinion article of at least 350 words in <language>.*

- For reporting: *Write a report article of at least 350 words in <language>.*

To this end, we generated an additional 500 examples for each class of the aforementioned languages, resulting in a total of about 13500 generated articles and added them to the datasets.

3.2 Persuasion techniques detection

3.2.1 Dataset description

The articles available for ST3 were divided into paragraphs. These paragraphs were labeled with persuasion techniques. In total, there were as many as 23 different persuasion techniques. Each paragraph could have included one or more labels in the form of persuasion techniques, but it was also possible that such techniques were not present in the paragraphs.

3.2.2 Augmentation

For detecting persuasion techniques, we applied data augmentation in the form of Machine Translation. Specifically, we utilized the DeepL API to translate the available corpus. We employed this approach to all languages. See Table 1 to understand how we translated different languages.

Original Language	Destination Languages
Italian	French
French	Italian
German	English
Polish	German, Russian

Table 1: Original Language represents the data in the source language, while the Destination Language indicates all languages to which the original paragraphs were translated.

4 Systems Description

4.1 Genre detection

As previously stated, the genre detection task involved a dataset containing news articles in various languages categorized as either opinion, reporting, or satire. The objective was to create a model that could distinguish between these three categories, thereby requiring the development of multi-class text classifiers at the document level. Our solution involved building a single machine learning model for genres detection in all given languages. For this purpose, we used the multilingual RoBERTa-large (Zhuang et al., 2021) encoder provided by the *HuggingFace* as a pre-trained model *XLM-RoBERTa-large*[6]. The text is first tokenized and represented

[6] https://huggingface.co/xlm-roberta-large

as an array of tokens, with *[CLS]* at the start, *[EOS]* at the end, and *[SEP]* separating sentences. This array is then passed through the *XLM-RoBERTa* model, which produces an array of embeddings corresponding to the input tokens. The embedding array is then fed through a fully connected layer having three neurons at the output, one for each class. The output of this layer (logits) is normalized using a softmax function. While training, the resulting normalized output is used with the relevant labels as input for the categorical cross-entropy function to determine the loss. During inference, when making predictions, the predicted class is obtained by selecting the category with the highest probability value by taking the argmax of the normalized output.

4.2 Persuasion techniques detection

We created separate systems for each language available. However, we categorized our systems into two distinct methods. One method was used for analyzing online news data in Polish, while the other was used on remaining languages. Namely, we employed an ensemble consisting of three one-vs-rest classifiers and BERT-based pre-trained models.

4.2.1 Persuasion detection in Polish

For the Polish language, we utilized a classical supervised machine learning approach. Firstly, we generated text embeddings for each paragraph using HerBERT[7], which is a BERT-based Language Model trained on Polish corpora (Mroczkowski et al., 2021). Next, we created another text representation using the StyloMetrix library (Inez Okulska, 2022). This library enables to represent text by quantifying different linguistic features, each in separate dimension. These features belong to the following groups: Grammatical Forms, Inflection, Lexical, Psycholinguistic, Syntactic, and Word Formation and are characterized by three significant aspects: interpretability, reproducibility, and normalization. Additionally, the features are normalized as they express the number of occurrences of a given feature per the number of tokens in the text (Inez Okulska, 2022). This normalization allows us to avoid the scaling effect that could occur in paragraphs of different lengths. We used StyloMetrix features together with HerBERT's embeddings as the input data to three classical machine learning models, namely Logistic Regression (Kleinbaum et al., 2002), eXtreme Gradient Boosting (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017). We then employed ensemble learning as a soft voting classifier to combine outputs from three distinct machine learning classifiers. The soft voting method, utilized in a binary classification scenario, generates the predicted class label by computing the argmax of the summation of the predicted probabilities of each classifier. Following the ensemble architecture, on top of the three classifiers, we applied a meta-estimator [8] to enable multi-class multi-label prediction of persuasion techniques.

4.2.2 Persuasion detection in the other languages

Our systems are purely transformer-based for other languages, namely English, Italian, Russian, French, and German. We used different BERT-based language models provided by *HuggingFace*. Table 2 shows pre-trained models for each remaining language. As a preprocessing step, we first drop paragraphs that are duplicated. We then employed one hot encoding for labels. We applied a summarization pipeline using an encoder-decoder pre-trained model to all paragraphs that, after splitting by space, have more than 300 substrings. Specifically, we used *t5-small* pre-trained model provided by *HuggingFace* (Raffel et al., 2020).

Following paragraph summarization, all paragraphs were tokenized and truncated. For each language, we used tokenization appropriate for the specific pre-trained model mentioned in Table 2. The resulting array of tokens is then passed through the appropriate model. As a result, we have an array of embeddings corresponding to the input tokens, which are then passed to a randomly initialized classification head (linear layer) on top. This head is trained with the pre-trained model in fine-tuning process on labeled input data. The amount of available persuasion methods is equivalent to the number of neurons in the model's output. More precisely, 23 outputs neurons for all languages except English, as the English dataset included 19 persuasion techniques. The last layer produces the not normalized logits, a tensor containing the scores for each label. We then apply a sigmoid function independently to every score to normalize the output. Finally, the binary cross-entropy loss generates the difference between the predicted la-

[7] `https://huggingface.co/allegro/herbert-large-cased`

[8] This meta-estimator was a simple strategy that consisted of fitting one ensemble classifier per label.

bels and the ground truth. At the inference stage, the normalized output is subjected to a threshold of 0.3. Specifically, if the output was higher than 0.3, it got labeled 1 and 0 otherwise. We used the same threshold while computing metrics during training.

5 Experimental Setup

5.1 Genre detection

To establish the experiment for detecting genres, we began by creating the training set. Then, we identified the most effective hyperparameters to train the model. Finally, we used the trained model with these hyperparameters to predict the classes of the provided test dataset. To create the training dataset for our experiment, we combined all the provided training datasets for all languages and added articles generated by GPT-3 as previously described in subsection 3.1.2. Likewise, the validation dataset was formed by combining all the validation sets for each language. After establishing the training and validation set, we identified the optimal hyperparameters for training the model. We accomplished this by performing a hyperparameter search for various values of batch size (ranging from 2 to 14), learning rate (ranging from 1e-7 to 1e-4), and weight decay (ranging from 0 to 0.4). Additionally, we implemented a linear warmup for the first 6% of the training steps. We conducted the hyperparameter search using combined validation sets for all languages. The optimal hyperparameters we identified through this process were a batch size of 10, a learning rate of 5e-5, and a weight decay of 0.3.

5.2 Persuasion techniques detection

We randomly divided the data into training and validation data for all languages. In our experiments, we assessed the performance of our models by checking the micro-average F1 score on development data available in the task that was not previously used for training. Any supplementary data generated through various augmentation techniques was also integrated into the training dataset. In the end, in training our final models, we also incorporated the development data.

5.2.1 Persuasion detection in Polish

Our approach for Polish dataset employs an ensemble of three classical machine learning models: XGBoost, Logistic Regression, and LightGBM. The data was divided into training and validation sets, where each observation consisted of a paragraph and the associated persuasion techniques labels. For the ensemble approach, we first utilized the *sentence_transformers*[9] library version 2.2.2 to generate embeddings as features for individual paragraphs (Reimers and Gurevych, 2019). Next, we employed the *StyloMetrix*[10] tool, specifically its v0.0.6 release, to generate 100 fully interpretable, normalized, and reproducible features that represent distinct groups of linguistic features. We evaluated our models by checking the F1 micro score on the development set. Due to time constraints, in Polish, we did not perform hyperparameter optimization, however we tested different decision threshold[11] from 0.01 to 0.50, so we had 50 results per training. After we found best threshold (0.08), we trained the final model on all available data in the task, i.e., training and development data. That model was used to infer the test set.

5.2.2 Persuasion detection in the other languages

Initially, the data was divided into data for training and validation. The data for training was augmented to include data generated using translations from other languages, as shown in Table 1. We used DeepL API to generate the translations. In order to interact with DeepL API, we utilized the *DeepL*[12] library of the latest available 1.13.0 version. We conducted experiments separately on each language using the pre-trained models shown in Table 2. As mentioned in our system overview, we performed summarization by utilizing *t5-small*[13] for observations that included paragraphs with more than 300 substrings.

Once we had the datasets available, we thoroughly explored hyperparameters on the models for each language. Our hyperparameters optimization entailed examining the effects of multiple values: batch size (ranging from 3 to 8), learning rate (ranging from 1e-5 to 1e-4) as well as weight decay (ranging from 0 to 0.05). In addition, a linear warmup strategy was applied for the initial training steps (ranging from 10 to 40 steps). We evaluated an individual set of hyperparameters by measuring the micro-average F1 score on the validation set.

[9] https://github.com/UKPLab/sentence-transformers

[10] https://github.com/ZILiAT-NASK/StyloMetrix

[11] The decision threshold is the number below which an observation is classified as having no persuasion technique.

[12] https://github.com/DeepLcom/deepl-python

[13] https://huggingface.co/t5-small

Language	BERT-based model
English	bert-large-cased
Italian	dbmdz/bert-base-italian-xxl-cased
French	dbmdz/bert-base-french-europeana-cased
German	bert-base-german-cased
Russian	DeepPavlov/rubert-base-cased

Table 2: BERT-based pre-trained models for each language provided by *HuggingFace*.

Genre Detection				
	F1 Macro Score			
Language	Baseline	Development Data	Test Data	Official Rank
English	0.28802	0.578	0.59	5
French	0.56806	0.691	0.71	4
German	0.62963	0.785	0.813	2
Italian	0.38940	0.714	0.72	3
Russian	0.39831	0.514	0.558	7
Polish	0.48962	0.860	0.661	7
Spanish	0.15385	n.a.	0.563	1
Greek	0.17054	n.a.	0.593	8
Georgian	0.25641	n.a.	0.597	5

Table 3: Results for subtask 1 on the development and test sets for all languages.

6 Results

Each result presented for genre categorization, and persuasion techniques detection is the average of 5 different runs on random seeds.

6.1 Genre detection

Table 3 shows the evaluation outcomes of our ultimate model on both development and test datasets for each language. The table also presents the model's leaderboard rank. Despite model's straightforward design, the results are noteworthy. In terms of macro-average F1 score, it achieved first place for Spanish, second for German, and third for Italian, among other notable accomplishments.

6.2 Persuasion techniques detection

Our results in Table 4 show that we achieved the highest F1 micro score for the Italian language on the test data. In addition, we obtained the top third official rank in Polish by utilizing an ensemble model comprised of three classical machine learning models. In German and Polish, we got better results on the final test data than during our experiments on development data. Moreover, it seems that our model for the Russian language overfitted on training data. One of the reasons that lead us to make this statement is that the results of test data are much worse than development data. Our systems consistently ranked among the top 10 performers across all languages, as shown in Table 4. In general, our average F1 micro score from our systems made us the sixth best performers among the languages in which we participated.

7 Post Deadline Multilingual Persuasion Techniques Detection

Due to constraints on available computing power, we were compelled to decide for which task to use larger and multilingual pre-trained models. We assumed that the classification of persuasion techniques would be more language-specific. As a result, in the ST3, we used different models for each language. However, we proceeded with experiments following the final call for submissions and fine-tuned *XLM-RoBERTa-large*. This approach produced superior results to our final solution to this competition for each language.

7.1 Systems description

The input data consisted of articles from all languages. Additionally, we added development data and all translated articles between languages as stated in Table 1. The goal was to develop a multilingual system that would detect persuasion techniques. For that purpose, we used the same pre-

Persuasion techniques detection				
	F1 Micro Score			
Language	Baseline	Development Data	Test Data	Official Rank
English	0.19517	0.3414	0.32004	9
French	0.24014	0.4048	0.38771	7
German	0.31667	0.3668	0.40816	9
Italian	0.39719	0.5279	0.49563	7
Russian	0.20722	0.3764	0.25714	10
Polish	0.17928	0.3755	0.38958	3

Table 4: Results for subtask 3 on the development and test sets for all languages.

Persuasion techniques detection - Post Deadline			
	F1 Micro Score		
Language	Baseline	Test Data	Rank
English	0.19517	0.39029	1
French	0.24014	0.46249	3
German	0.31667	0.52006	1
Italian	0.39719	0.56480	1
Russian	0.20722	0.38588	2
Polish	0.17928	0.40753	4
Spanish	0.24843	0.37605	3
Greek	0.08831	0.27827	1
Georgian	0.13793	0.42657	3

Table 5: Results for subtask 3 on test sets for all languages. Results post deadline as of 15.04.2023.

trained model for genre detection, namely *XLM-RoBERTa-large*. All paragraphs were first tokenized and truncated to the first 512 tokens. We used the resulting array of tokens to produce embeddings by utilizing *XLM-RoBERTa-large*. The array of embeddings was then passed through a fully connected layer with a randomly initialized classification head with 23 neurons on top. We normalized produced logits with the sigmoid function, and then we used the binary cross-entropy loss to generate the difference between the predicted and the actual labels. During inference, we utilized the same threshold of 0.3 as for before described BERT-based systems. We assigned the presence of appropriate persuasion techniques for scores higher than 0.3, and in the other cases, we assigned 0, i.e., no persuasion technique.

7.2 Experimental setup

For our experiment, we divided data into training and validation datasets so that both included paragraphs from each language. The English dataset with articles included 19 persuasion techniques, and the rest of the languages included 23 techniques. To incorporate all of the data for training and maintain a consistent target variable, we added four missing labels denoting persuasion techniques to the English articles. These labels were marked for English as not appearing in any paragraph. We did not perform optimization of hyperparameters in the post-deadline experiment. Instead of optimization, we used predetermined values. See Table 6

Hyperparameter	Value
`learning rate`	`5e-6`
`batch size`	`8`
`epochs`	`6`
`warmup steps`	`30`
`weight decay`	`0.03`

Table 6: Hyperparameters and their values assumed for multilingual model for persuasion detection

7.3 Results

First of all, it is essential to mention that we did not have ground truth available for the test data at the time of developing this system. We received results on test data thanks to the organizers providing a post-deadline leaderboard for further submissions.

We show our results of multilingual post-deadline experiments on persuasion techniques' detection in Table 5. Except for Polish, our solutions would be in the top three in each language. Our post-deadline results are superior to our final submission results for this task.

8 Conclusion

In this paper, we present our solutions for two SemEval subtasks: news genre categorization and persuasion techniques' detection.

We developed a single multilingual system utilizing the pre-trained RoBERTa model to classify online news genres. To improve the performance of the system, we added to the training dataset texts generated by the GPT3-Davinci language model.

Regarding the persuasion techniques' detection task, we provided individual solutions for each language with available training data. For the Polish language we created a system that employs ensemble learning of three classical machine learning models: LightGBM, XGBoost, and Logistic Regression. For that system, we generated two types of features from data, specifically, HerBERT-based embeddings and fully interpretable linguistic features. All other systems for persuasion techniques detection employ BERT-based pre-trained models and use summarization applied to longer paragraphs in online news articles.

Most final systems were trained on augmented data obtained through various techniques, including GPT-3-based text generation, machine translation utilizing the DeepL API. Our systems consistently ranked among the top 10 solutions, exceeding the baseline the organizers set. In future work it would be interesting to further develop the multilingual approach to detect persuasion techniques.

Limitations

Our study has a few limitations that must be considered when interpreting the results. Initially, we faced difficulties in meeting the strict requirements of the SemEval format, which limited the scope and depth of our analysis. As a result, we had to prioritize certain methods over others, and we were unable to test all potential combinations of data augmentation techniques.

For instance, in ST1, we could have extended the training dataset by translating each language into every other language using different machine translation models. Similarly, in ST3, due to the aforementioned limitations, we only translated certain languages into others for training purposes and used StyloMetrix features only in Polish.

Furthermore, the OpenAI policy regarding their language models prevented us from generating syntactic data for ST3 (persuasion technique detection) due to the potential for misuse and impact.

Overall, a more thorough investigation into the impact of various augmentation techniques on model performance could help determine whether augmentation improves the model's outcomes and which technique provides the most performance gain.

Despite these limitations, our study provides a valuable contribution to the field by demonstrating the effectiveness of our system architecture, which employs DeepL and GPT-3 for data augmentation for specific tasks.

Acknowledgements

This work has been supported from the grant of the Polish National Center for Research and Development titled: "Development and verification of original methods of vertical artificial intelligence for automatic and precise detection of disinformation"(project number: INFOSTRATEG-I/0010/2021-00).

References

Alberto Barrón-Cedeno, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. 2019. Proppy: A system to unmask propaganda in online news. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9847–9848.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. Semeval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2021. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4826–4832.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. 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)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Guillaume Daval-Frerot and Yannick Weis. 2020. Wmd at semeval-2020 tasks 7 and 11: Assessing humor and propaganda using unsupervised data augmentation. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1865–1874.

Thomas Haider and Alexis Palmer. 2017. Modeling communicative purpose with functional style: Corpus and features for German genre and register analysis. In *Proceedings of the Workshop on Stylistic Variation*, pages 74–84, Copenhagen, Denmark. Association for Computational Linguistics.

Anna Zawadzka Inez Okulska. 2022. Styles with benefits. the stylometrix vectors for stylistic and semantic text classification of small-scale datasets and different sample length. *Proceedings of the 3rd Polish Conference on Artificial Intelligence, April 25-27, 2022, Gdynia, Poland.*

Dawid Jurkiewicz, Łukasz Borchmann, Izabela Kosmala, and Filip Graliński. 2020. Applicaai at semeval-2020 task 11: On roberta-crf, span cls and whether self-training helps them. *arXiv preprint arXiv:2005.07934*.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Evgeny Kim, Sebastian Padó, and Roman Klinger. 2017. Investigating the relationship between literary genres and emotional plot development. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 17–26, Vancouver, Canada. Association for Computational Linguistics.

David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein. 2002. *Logistic regression*. Springer.

Lily Li, Or Levi, Pedram Hosseini, and David A Broniatowski. 2020. A multi-modal method for satire detection using textual and visual cues. *arXiv preprint arXiv:2010.06671*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. 2021. HerBERT: Efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10, Kiyv, Ukraine. Association for Computational Linguistics.

Rajaswa Patil, Somesh Singh, and Swati Agarwal. 2020. Bpgc at semeval-2020 task 11: Propaganda detection in news articles with multi-granularity knowledge sharing and linguistic features based ensemble learning. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1722–1731.

Ziming Pei. 2022. The impact of semantic and stylistic features in genre classification for news.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. 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*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Victoria L Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the second workshop on computational approaches to deception detection*, pages 7–17.

Fan Yang, Arjun Mukherjee, and Eduard Dragut. 2017. Satirical news detection and analysis using attention mechanism and linguistic features. *arXiv preprint arXiv:1709.01189*.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized bert pre-training approach with post-training. In *Proceedings of the 20th chinese national conference on computational linguistics*, pages 1218–1227.