


Manipulative Style Recognition of Czech News Texts using Stylometric Text Analysis

Radoslav Sabol and Aleš Horák 

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00, Brno, Czech republic
{xsabol, ha.les}@fi.muni.cz

Abstract. The rampant spread of manipulative texts purporting propaganda, disinformation or surveillance, requires the readers to take heed of the actual reasoning behind and the real purpose of the newspaper texts. The capability to recognize a malignant content asks for more and more concentration and background knowledge. A support offered by automated content analysis tools forms an important part of such protective approaches.

In the presented text, we introduce a new approach to detecting a set of manipulative stylistic techniques in Czech newspaper texts by exploiting stylometric methods in conjunction with deep learning text classification. We show that the stylometric analysis with almost 20,000 features allows to improve the results for most of the techniques. The results are evaluated with a previously published Czech Propaganda dataset.

Keywords: stylometry, propaganda detection, manipulative style analysis, Propaganda dataset, Czech

1 Introduction

The current accessibility and popularity of the Web, along with the ease of creating new content, takes freedom of expression to a whole new level, which is considered a positive development. An adverse side effect of this makes it straightforward to create websites and online news outlets that publish manipulative content. Disinformation through online news outlets creates an illusion of the information being reliable, affecting a much broader population than from the other sources [7]. Due to the immense and dynamic nature of the Internet, manual detection is difficult to grasp, and automated tools are desired to protect or warn readers of the manipulative content.

In 2019, a shared task was held in *Workshop on NLP4IF: censorship, disinformation, and propaganda*¹ [4]. The task consisted of two different problems based on the **Propaganda Techniques Corpus** [5] dataset. The dataset consists of fine-grained annotations that range from techniques that *leverage emotions* (for example *Loaded Language*, an act of using phrases with strong connotations) to *logical*

¹ <http://www.netcopia.net/nlp4if/2019/>

fallacies (like *Straw Man*, where writer refutes an argument not presented by the opposition). From 25 submitted approaches, the best-performing ones utilized BERT contextual embeddings. Other successful approaches exploited contextual embeddings of RoBERTa, ELMo, or context-independent representations based on lexical, sentiment, or TF-IDF features [4].

In the current paper, we present recent results in **manipulative style recognition** of Czech texts. In Section 2, we describe the specifics of the currently used benchmark dataset. Section 3 proposes a set of stylometric text features crafted using Czech linguistic tools. Section 4 presents a deep neural architecture based on XLM Roberta [3] that combines both the contextless stylometric features and the context-specific representation based on transformer models. In the last section, we evaluate and compare the approach that uses the proposed features and a similar model that does not.

2 Dataset Description

The *Propaganda* benchmark dataset, originally proposed by Baisa et al. [2], is a collection of 8,644 documents extracted from Czech news outlets that were previously investigated for spreading Russian propaganda [1]. The benchmark dataset is annotated with 21 diverse attributes, where 16 of them are relevant for analysis presented in this paper. The scope of annotation ranges from document-wide attributes to span level attributes that mark a specific segment of the text as an occurrence of a specific stylistic technique. The documents were tokenized and morphologically annotated using `unitok` [8], `majka` [11] and `desamb` [13] tools.

Among the annotated attributes, eight of them refer to *manipulative techniques*. The techniques are labeled at both the document and the span level.

- **Argumentation**: author presents an argument (yes/no)
- **Blaming**: author is blaming someone (yes/no)
- **Demonization**: author refers to individuals, groups, or political bodies as evil (yes/no)
- **Emotions** author uses emotive writing (fear, anger, indignation, compassion, other, missing)
- **Fabulation**: author spreads false rumors and exaggerates problems (yes/no)
- **Fear Mongering**: author appeals to fear, uncertainty, or certain threat (yes/no)
- **Labelling**: authors labels an entity with a short, pejorative phrase (yes/no)
- **Relativizing**: author either relativizes negative actions of Russia or positive actions of the opponent (yes/no)

Document-level attributes describe the expected structure and content of the document, so it is not reasonable to annotate them on the span level. These attributes are **Genre** (3 categories), **Topic** (13 categories), **Scope** (4 categories), **Location** (8 categories), and **Overall Sentiment** (3 categories).

Other attributes have annotations present on the span level, but they are not described as manipulative techniques. The listed attributes are **Expert** (yes/no), **Opinion** (yes/no), and **Russia** (5 categories).

3 Stylometric Text Features

In this section, we describe the proposed features that can be observed in Table 1. Overall, there is a lack of consensus for an ideal, universal set of stylometric features as they depend on the currently analyzed task and domain. The process of extracting such features requires linguistic analysis at various levels: *lexical*, *syntactic*, *semantic*, *structural*, *content-specific*, and *idiosyncratic* [6]. The current feature extraction implementation is inspired by the features proposed by [9].

Word and **Sentence length** distributions are implemented in three ways for both tokens and lemmas. The **naive** version is not adjusted to the real distribution present in the dataset and directly assumes word lengths 1 through 15 and more. **Improved** analysis creates bins of variable length derived from the data. The **N-Gram** approach analyzes naive word length n -grams.

Word Class N-Gram frequencies are extracted using annotations by majka and desamb. The N parameter ranges between 1 and 5, and only the n -grams with a relative frequency above 0.1% are considered. **Morphological Tags N-gram** frequencies consider more information than word classes. The **full** version uses the entire morphological tag, whereas the **simplified** omits the infrequent parts of the tag.

Word Repetition metrics are analyzed on multiple levels. **Average repetition** features compute frequency histogram for each unique token/lemma in the document, where the mean relative frequency is the resulting feature. **Bag of Words** repetition turns documents into TF-IDF normalized bag of words representation where stopwords and words with too low relative frequency are omitted. **Word Class Repetition** is a normalized word class histogram where for each token and its corresponding word class, the word class count is incremented for each sentence the token is repeated in. **Probabilistic Word Class Repetition** computes the probability of word class being repeated from the referential corpus and returns the difference between the referential probabilities and the probabilities from the provided document.

Letter Casing features are composed of two different methods. The first method computes n -grams of capital letters according to their position in the word and sentence using fixed rules. The indexed version also considers the exact position of the token in the sentence.

The **parametrized** version of **word suffixes** computes relative frequencies of the last n characters of each token document-wide. The **stemmed** version attempts to guess the word suffix based on the provided word and its corresponding lemma.

Word Richness considers two methods of computing vocabulary richness. *Simpson's Diversity Index* [10] is computed on all alphanumerical and alpha

tokens. *Coefficient of Colligation*, as known as *Yule's K* [12] is computed with the k values of 10, 100, 1,000 and 10,000.

Punctuation frequency examines the presence of various punctuation marks in the document. The **position frequency** version also considers the placement of punctuation marks. Finally, the n -gram version observes 100 most common punctuation n -grams with a relative frequency above 0.2%.

Fixed Typography is an idiosyncratic feature that checks for typography rules violations and various patterns related to typography that are checked using 11 regular expressions. **Dynamic Typography** observes the n -gram frequencies of non-alphanumeric tokens.

Character N-gram frequencies are extracted for at most 1,000 unique items with preferred document frequency around 50%. **Emoticon Presence** checks for the presence of pre-defined emojis in the presented document.

Table 1: Overview of proposed stylometric text features

Feature Type	Feature Subtype	# features	Language Independent
Word Length	naive	30	✓
	improved	77	✓
	n -grams	30	✓
Sentence Length	naive	25	✓
	improved	127	✓
	n -gram	25	✓
Word Repetition	avg. repetition per sent.	1	✓
	avg. repetition per doc.	1	✓
	word class repetition	13	
	prob. word class repetition	13	
	word repetition distance	12	✓
	bag of words repetition	100	✓
Word Class N-Grams	1 to 4-grams	514	
Morphological Tags N-Grams	full	10,000	
	simplified tags	200	
Letter Casing	1 to 3-grams	77	✓
	indexed 1 to 3-grams	417	✓
Word Suffixes	stemmed	100	✓
	parametrized n -grams	325	✓
Word Richness	richness metrics	6	✓
Stopwords	for lemmas	300	✓
	for tokens	300	✓
Punctuation	frequency	11	✓
	position frequency	60	✓
	n -gram frequency	76	✓
Typography	fixed rules	11	✓
	dynamic	100	✓
Character N-Gram Distribution	1 to 5-grams	6,550	✓
Emoticons Presence	n -grams	28	✓
Total		19,529	

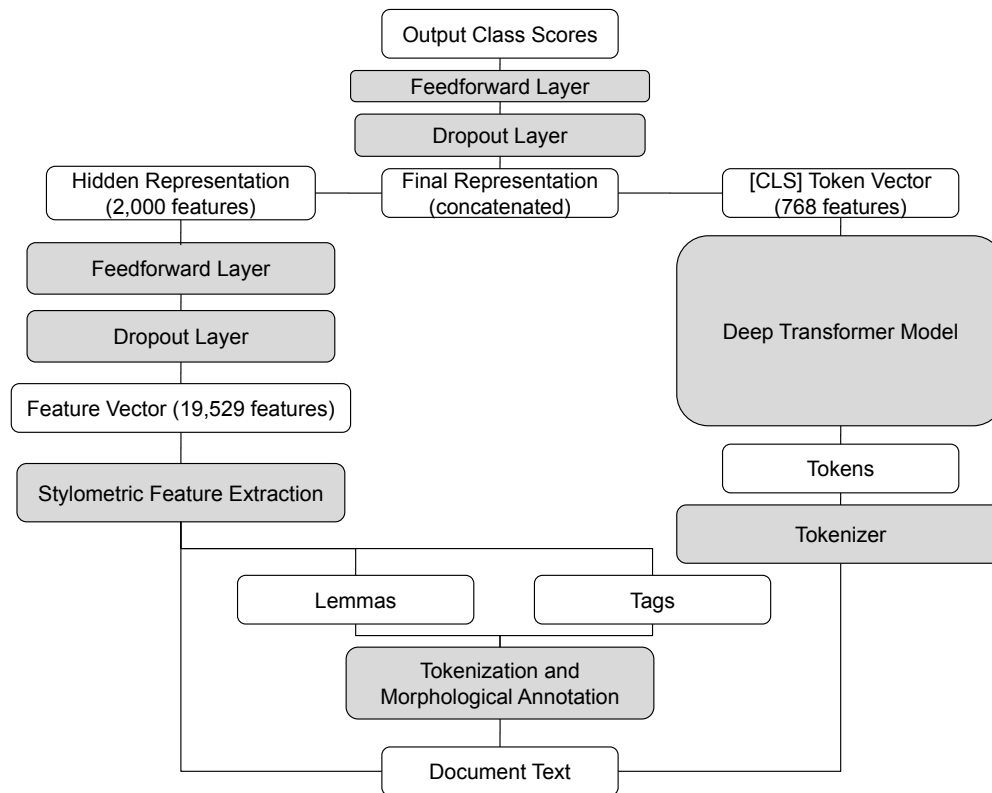


Fig. 1: Deep neural architecture for manipulative style detection

4 Detection Approach

The following section proposes an approach for manipulative style recognition using both stylometric text analysis and deep neural transformer models. A visual schema of this architecture is depicted in Figure 1.

4.1 Architecture Description

The input document is tokenized using `unitok` and morphologically annotated using `majka` and `desamb` tools. The resulting tokens, lemmas, and morphological tags are used to extract a feature vector using the stylometric analysis described in Section 3. The representation is then passed to a single feedforward layer. The resulting, more dense representation highlights the essential features for classification and represents the writing style used in the input document.

In tandem with the previous paragraph, the input text document is processed using **XLM Roberta Large** [3] deep transformer model that was pre-trained on 100 languages, the Czech language included. The model was pre-trained on sequences with a maximum token length of 512, so parts of the input document that exceed this limit are removed. The CLS token vector is extracted from the first item in the resulting sequence and it is then concatenated to the

hidden stylometric representation. In the final step, the concatenated representation is passed through the final feedforward layer, which models the predictions for each class in the attribute. The final prediction is selected using the *argmax* function.

4.2 Training Description

The proposed model is trained using the *HuggingFace* framework on 20 epochs. Hyperparameter values for the training were empirically determined. We use the learning rate of 3×10^{-6} and the linear warmup ratio of 0.1 for the AdamW optimizer. Due to the lack of training examples and label unbalance in some classes, more aggressive values for generalization were chosen. We use the dropout probability $p = 0.5$ for each presented feedforward layer, and a weight decay of 0.01 is used. The training computations were accelerated using GPU with a batch size of 32 and gradient accumulation to fit the GPU memory adequately.

5 Evaluation

In this section, we compare the proposed approach with two other approaches. The *dummy* baseline approach described in [1] always predicts the majority class. The second approach uses **XLM Roberta Large** with a standard HuggingFace classification head, where all the stylometric features are omitted.

The dataset is not split identically to Baisa et al. [1] because they use a different version of the *Propaganda* dataset. New train/test split was defined for the final version of the dataset, where 1,000 test examples are reserved for evaluation purposes. The testing set is identical throughout all evaluated attributes. Also, 500 examples were split as a development set for early stopping.

The experiments are performed for all the attributes mentioned in Section 2, where each setup is trained three times. The average **weighted F1** metric is computed from the three performed runs to factor in the label imbalance.

5.1 Results

Table 2 summarizes the performance of the proposed techniques for all attributes. The results showed that the dummy baseline was outperformed by a large margin in most categories. Notable exceptions can be seen in *Demonization* and *Relativization*, where the binary label imbalance is considerably higher than in other attributes.

The average weighted F1 score of the stylometric approach is lower than the text-only classification in cases of *Argumentation*, *Topic*, and *Labelling*. *Argumentation* is considered a complex and noisy attribute with a relatively low inter-annotator agreement. The current definition of *Argumentation* allows for anything from simple reasoning to a solid argument, along with some logical

Table 2: Summary of weighted F1 scores in % for the presented techniques. XLMR refers to XLM Roberta. Diff refers to the difference between the stylometric and non-stylometric XLM Roberta approach.

Attribute	Dummy	XLMR Large	XLMR Large with Stylometry	Diff
Argumentation	42.46	70.69	70.64	-0.05
Blaming	60.67	74.55	74.92	0.37
Demonization	95.67	96.13	96.19	0.06
Emotions	77.82	81.81	82.63	0.82
Fabulation	74.87	80.57	80.92	0.35
Fear Mongering	88.89	91.71	91.85	0.14
Labelling	76.7	83.37	83.09	-0.27
Relativizing	92.27	92.75	92.84	0.09
Genre	85.99	96.46	96.8	0.34
Topic	10.22	71.93	71.12	-0.81
Scope	41.03	89.36	90.15	0.79
Location	20.45	82.95	83.77	0.82
Sentiment	74.59	83.14	83.06	-0.08
Expert	39.03	76.1	77.42	1.32
Source	44.39	52.06	55.46	3.4
Opinion	80.52	87.61	88.35	0.74
Russia	53.12	82.88	83.63	0.75

fallacies, to be included in this class. Due to such high variation, the *Argumentation* may be challenging to grasp using automated machine learning methods. The difference between approaches is **0.05%** in favor of the non-stylometric one, which is considered a non-significant difference.

Topic results report **0.81%** difference in the favor of non-stylometric approach. The reason behind this difference may be that the attribute dwells in the semantics and the content of the document, but the proposed features specialize in non-content features. Thus in the learning process here, stylometric features may present overabundant information that degrades the overall performance.

The *Labelling* manipulative technique usually refers to a **small segment** of the text containing short, powerful phrase. Stylometric features, on the other hand, summarize the writing style of the **entire document**. The reason behind the **0.27%** decrease in the performance metric may be that the stylometric features could not properly capture this attribute’s characteristics.

The most notable performance **increase** of **3.4%** can be seen with the *Source* attribute. The most important features responsible for the improvement relate to the presence and position of **capital letters**, as cited sources tend to be capitalized. Similar reasoning can apply to the *Expert* attribute, where a notable improvement is also present.

Another significant improvement of **0.82%** can be noticed with the *Emotions* attribute. The emotive writing style differs considerably from regular news, allowing for improved detection capabilities using the proposed features.

6 Conclusions and Future Directions

In the paper, we have evaluated a deep neural architecture that combines transformer models and stylometric text features to solve the manipulative style recognition task. We have introduced 13 feature categories from various levels of linguistic analysis, resulting in a feature vector of almost 20,000 dimensions. The results show that the proposed approach increases the performance for most Propaganda benchmark dataset attributes. The most notable increase was observed in the *Source* and *Expert* attributes, followed by the *Emotions* manipulative technique. It was also discovered that for some attributes (mainly *Topic*), the extracted non-content features do not relate to the specifics of the attribute, introducing noise and subsequently decreasing performance.

In the future development of the approach, we aim to increase the set of explored style attributes in both language independent and dependent features and evaluate the approach with other datasets and languages.

Acknowledgements This work has been partly supported by the Ministry of Education of CR within the LINDAT-CLARIAH-CZ project LM2018101. Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum provided under the programme “Projects of Large Research, Development, and Innovations Infrastructures” (CESNET LM2015042), is greatly appreciated.

References

1. Baisa, V., Herman, O., Horak, A.: Benchmark dataset for propaganda detection in Czech newspaper texts. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). pp. 77–83. INCOMA Ltd., Varna, Bulgaria (Sep 2019)
2. Baisa, V., Heřman, O., Horak, A.: Manipulative propaganda techniques. In: RASLAN (2017)
3. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. CoRR [abs/1911.02116](https://arxiv.org/abs/1911.02116) (2019), <http://arxiv.org/abs/1911.02116>
4. Da San Martino, G., Barron-Cedeno, A., Nakov, P.: 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. pp. 162–170 (2019)
5. Da San Martino, G., Yu, S., Barrón-Cedeno, A., Petrov, R., Nakov, P.: 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). pp. 5636–5646 (2019)
6. Lagutina, K., Lagutina, N., Boychuk, E., Vorontsova, I., Shliakhtina, E., Belyaeva, O., Paramonov, I., Demidov, P.: A survey on stylometric text features. In: 2019 25th Conference of Open Innovations Association (FRUCT). pp. 184–195 (2019). <https://doi.org/10.23919/FRUCT48121.2019.8981504>

7. Martino, G.D.S., Cresci, S., Barrón-Cedeño, A., Yu, S., Di Pietro, R., Nakov, P.: A survey on computational propaganda detection. arXiv preprint arXiv:2007.08024 (2020)
8. Michelfeit, J., Pomikálek, J., Suchomel, V.: Text tokenisation using unitok. In: Horák, A., Rychlý, P. (eds.) RASLAN 2014. pp. 71–75. Tribun EU, Brno, Czech Republic (2014)
9. Rygl, J.: Automatic adaptation of author's stylometric features to document types. In: International Conference on Text, Speech, and Dialogue. pp. 53–61. Springer (2014)
10. Simpson, E.H.: Measurement of diversity. *Nature* **163**(4148), 688–688 (1949)
11. Šmerk, P.: Fast morphological analysis of Czech. Proceedings of recent advances in slavonic natural language processing, RASLAN **2007**, 13–16 (2009)
12. Yule, C.U.: The statistical study of literary vocabulary. Cambridge University Press (2014)
13. Šmerk, P.: K počítačové morfologické analýze češtiny (Towards computer morphological analysis of Czech). Ph.D. thesis, Masaryk University, Faculty of Informatics, Brno (2010), <https://theses.cz/id/28r7vj/>