Benchmark Dataset for Propaganda Detection in Czech Newspaper Texts

Vít Baisa and **Ondřej Herman** and **Aleš Horák**
Natural Language Processing Centre
Masaryk University, Faculty of Informatics
Botanická 68a, Brno
{xbaisa,xherman1,hales}@fi.muni.cz

Abstract

Propaganda of various pressure groups ranging from big economies to ideological blocks is often presented in a form of objective newspaper texts. However, the real objectivity is here shaded with the support of imbalanced views and distorted attitudes by means of various manipulative stylistic techniques.

In the project of Manipulative Propaganda Techniques in the Age of Internet, a new resource for automatic analysis of stylistic mechanisms for influencing the readers' opinion is developed. In its current version, the resource consists of 7,494 newspaper articles from four selected Czech digital news servers annotated for the presence of specific manipulative techniques.

In this paper, we present the current state of the annotations and describe the structure of the dataset in detail. We also offer an evaluation of bag-of-words classification algorithms for the annotated manipulative techniques.

1 Introduction

State and pressure groups propaganda is a very well studied phenomenon from the sociological point of view (Herman and Chomsky, 2012; Zhang, 2013; Paul and Matthews, 2016). With the spread of digital media, the influence of propaganda news grows rapidly (Helmus et al., 2018) and the consequences of public opinion manipulation reach new levels (Woolley and Howard, 2017).

The main way of self-protection against such propaganda influence lies in careful verification of the presented information sources. Nevertheless, psycholinguistic evidence (Fazio et al., 2015) shows that a prevailing opinion often outweighs even direct knowledge. Computational tools that could warn against possible manipulation in the text can thus offer an invaluable help even to an informed reader.

In the following text, we are presenting the first results of a research project aimed at automatic analysis of the *style* of a newspaper text to identify a presence of specific manipulative techniques. In the first phase, a specific tool for expert annotations of selected news from 4 Czech internet media sites was developed (Baisa et al., 2017). This tool has now been used to obtain 7,494 annotated articles with detailed manipulative techniques annotations of texts expressing e.g. blaming, demonizing, relativizing, labelling, or fear mongering. The following Section 2 provides detailed information about the dataset characteristics and content. In Section 3, an evaluation of 10 classification techniques and their results with the benchmark dataset is presented.

2 The Benchmark Dataset

The Propaganda benchmark dataset currently contains data from two successive years. The first part is based on two sets of articles from 2016. The newspaper texts were extracted from four newspaper media domains[1] which were previously scrutinized by annotators as possible sources of pro-Russian propaganda. The downloaded cleaned data were merged with the annotation data stored separately in a SPSS[2] format (converted with the GNU PSPP tool[3]) which is used widely in Social science research. The result is a corpus with metadata (structure attributes) available for full-text

[1] sputnik.cz, parlamentnilisty.cz, ac24.cz and www.svetkolemnas.info.

[2] https://www.ibm.com/products/spss-statistics

[3] https://www.gnu.org/software/pspp/

Proceedings of Recent Advances in Natural Language Processing, pages 77–83,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_010

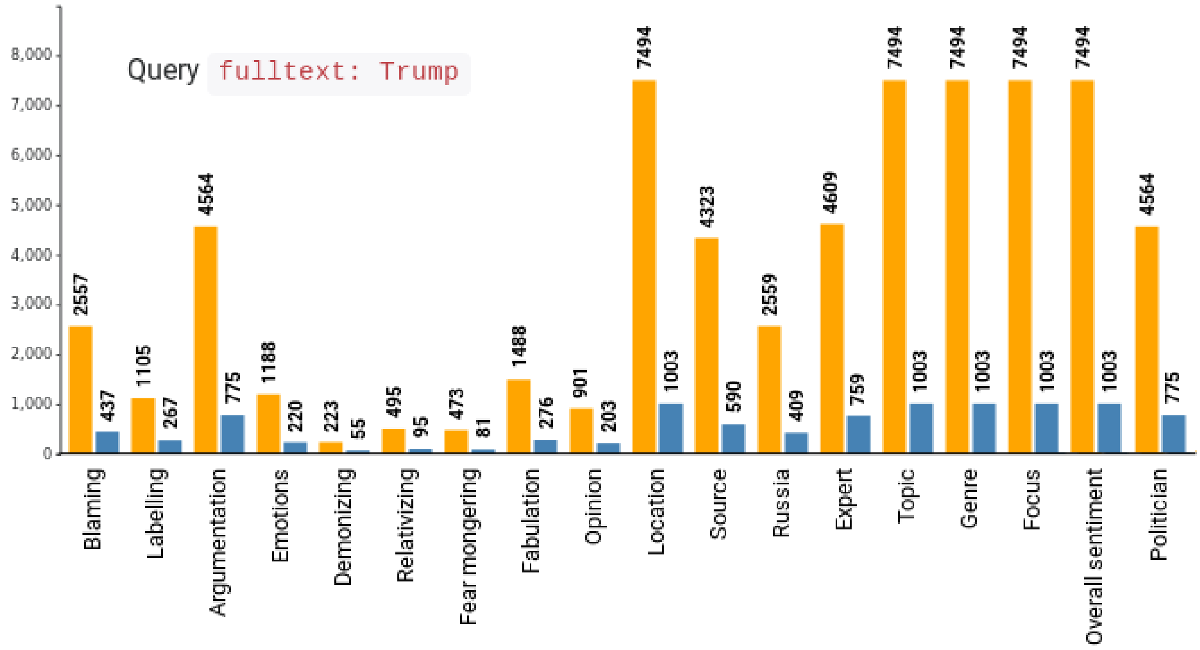

Figure 1: Numbers of articles with significant attribute values (not null, neutral or missing) in the whole collection of 7,494 documents. The first (yellow) columns show numbers for the whole collection and the second (blue) columns show an example of a filtered subset of articles containing the word "`Trump`".

search in the Sketch Engine corpus manager (Kilgarriff et al., 2014). As far as we know, this is the first corpus of propaganda text annotated for detailed ensemble of manipulative techniques. The full document texts were thus supplemented with the following attributes (see Figure 1 for representation of particular attributes in the dataset):

a) **Blaming**: does the text accuse someone of something?

b) **Labelling**: the text uses specific labels – short and impactful phrases or words – to describe a person or a group.

c) **Argumentation**: does the text present facts or arguments (logical, emotional, etc.) to support the main claim?

d) **Emotions**: What is the main emotion the text is trying to evoke in the reader? Anger, hate, fear.

e) **Demonizing**: is the “enemy” and/or his/her goals or interests presented in the text as being evil?

f) **Relativizing**: are the presented actions of a person, group or party being relativized?

g) **Fear mongering**: is the text trying to appeal to fear, uncertainty or other threat?

h) **Fabulation**: does the text contain unsubstantiated, overstated or otherwise incorrect claims?

i) **Opinion**: does the author of the text present his or hers personal opinion?

j) **Location**: what is the main location the text talks about?

k) **Source**: is the text presented as being based on a specific source?

l) **Russia**: is the topic related to Russia?

m) **Expert**: is the text or opinion in the text presented as being supported by an expert?

n) **Attitude to a politician**: neutral, negative, positive for up to 3 mentioned politicians.

o) **Topic**: migrant crisis, domestic politics, etc.

p) **Genre**: report, interview, or commentary.

q) **Focus**: foreign, domestic, can’t be determined.

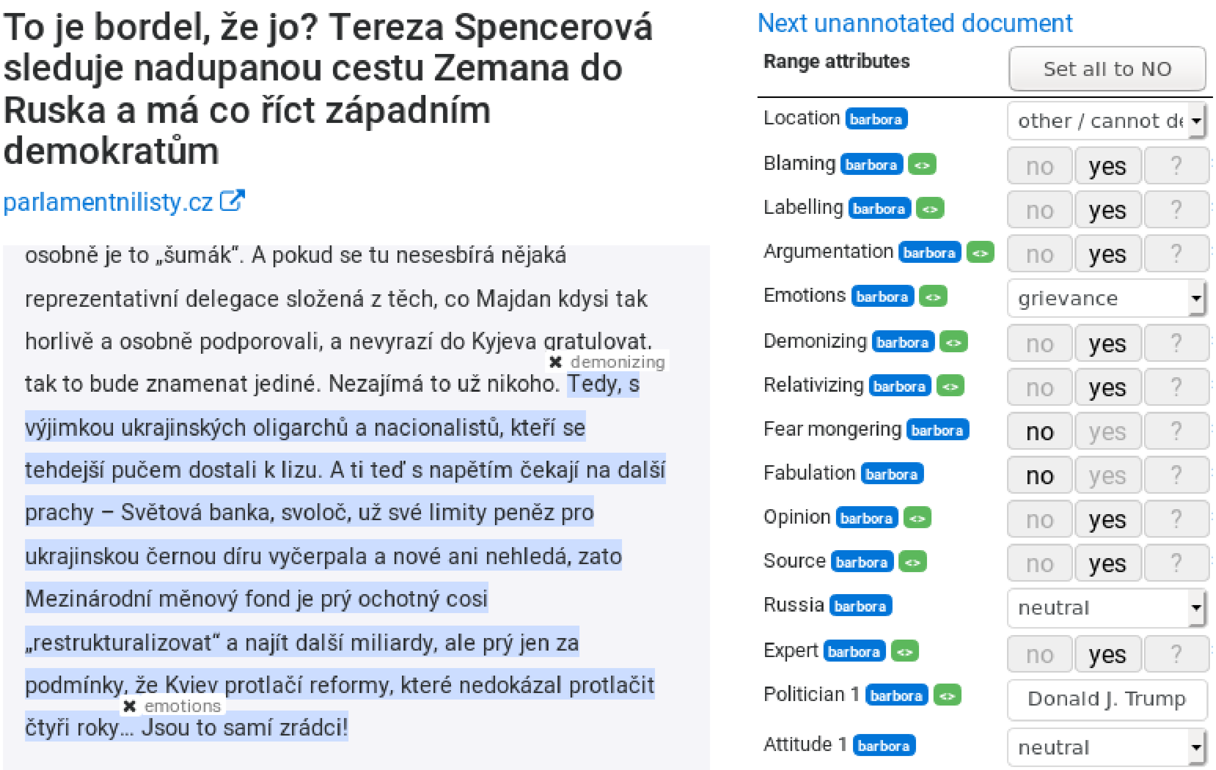

Figure 2: An example of (a part of) an annotated article with ranges showing *demonizing* and *grievance* as a value of the *emotions* attribute.

r) **Overall sentiment**: neutral, negative, or positive.

The second part, articles from the same domains published in 2017, has undergone a fine-grained annotation using a specific data processing and annotating tool (Baisa et al., 2017), which requires the annotators not only to specify the respective attribute values but also enrich them with particular phrase examples. The annotators were asked to amend each significant attribute value (not null, neutral or missing) by marking a particular block (or blocks) of text that offer the evidence of the value. The attributes are split into two groups. The attributes a) to n), denoted as *range attributes*, are bound to a sequence of words from the text, the attributes o) to r), i.e. the document attributes, are related to the article as a whole. An example of annotated range attributes can be seen in Figure 2. Unfortunately, due to the complexity of the annotation process, there was only one annotator per document and the inter-annotator agreement could not be decided.

The text of the articles has been extracted from the media server web pages, then tokenized using *unitok* (Michelfeit et al., 2014) and morphologically annotated using *majka* (Šmerk, 2009) and *desamb* (Šmerk, 2010). The dataset thus allows complicated full-text search in the articles. The size of the data (sub)sets is in Table 1.

Table 1: Text statistics of the two parts of the benchmark dataset.

	2016	2017	Total
Tokens	2,774,178	930,304	3,704,482
Words	2,331,116	781,725	3,112,842
Sentences	144,097	49,140	193,237
Paragraphs	50,554	17,264	67,818
Documents	5,500	1,994	7,494

3 Dataset Evaluation

We have performed the dataset evaluation to express the baseline accuracy of assigning the labels automatically using 10 machine learning classifiers. The classifiers were trained with the 20,000 most frequent lemmata present in the corpus, with the text transformed to a numerical vector format using bag-of-words using TF-IDF weighting.

Table 2: Classifier Accuracy

	Blaming	Labelling	Argumentation	Emotions	Demonizing	Relativizing	Fear mongering	Fabulation	Opinion	Location	Source	Russia	Expert	Topic	Genre	Focus	Overall sentiment	Server
dummy	.59	.79	.69	.81	.96	.93	.91	.74	.86	.41	.60	.70	.74	.32	.89	.53	.75	.63
bernoulli_nb	.67	.78	.59	.74	.87	.85	.84	.75	.84	.56	.63	.73	.63	.53	.91	.72	.72	.80
multinomial_nb	.67	.79	.70	.81	.96	.93	.91	.74	.86	.52	.60	.71	.74	.54	.89	.86	.75	.72
nearest_centroid	.66	.71	.62	.63	.74	.80	.75	.71	.75	.58	.60	.55	.67	.56	.80	.66	.65	.73
passive_aggressive	.70	.79	.72	.77	.96	.94	**.92**	.78	.84	.74	.67	.79	.80	.69	.95	.85	.73	.92
random_forest	.69	.81	.74	.81	.96	.93	.92	.77	.87	.67	.68	.80	.80	.63	.92	.85	.76	.88
ridge	**.72**	**.82**	**.75**	**.81**	.96	.94	.92	**.79**	.89	.75	**.70**	.80	.81	.71	.96	**.87**	**.78**	.91
sgd_elasticnet	.71	.82	.73	.81	.96	**.94**	.92	.78	.89	.76	.70	.82	.80	.71	.96	.87	.77	.93
sgd_l1	.70	.81	.72	.81	.96	.94	.92	.78	.89	.76	.70	**.82**	.81	.70	.96	.87	.77	**.94**
sgd_l2	.70	.82	.73	.81	**.96**	.94	.92	.78	**.89**	**.76**	.70	.81	.80	.71	.96	.87	.77	.92

3.1 Selected Classifiers

For the evaluation, we have chosen a representative subset of classification techniques, which are often employed in bag-of-words tasks for attribute value estimation. The resulting set of classifiers includes:

- `dummy`: a baseline, classifies every instance as the majority class present in the input data.
- `passive_aggressive`: an efficient Perceptron-like classifier (Crammer et al., 2006).
- Two Naive Bayes variants: `bernoulli_nb` assumes that the data is Bernoulli distributed, while `multinomial_nb` assumes a Multinomial distribution (McCallum et al., 1998).
- Three different Support Vector Machine classifiers trained using stochastic gradient descent: `sgd_l1` with L1 regularization, `sgd_l2` with L2 regularization and `sgd_elasticnet` with Elasticnet regularization (Zhang, 2004).
- `ridge` is a regularized linear regression based classifier (Rifkin and Lippert, 2007).
- `random_forest`: An ensemble of decision tree classifiers is built on samples drawn from the training set. The resulting class during the classification is obtained by taking the most common class as assigned by each of the decision trees (Breiman, 2001).
- `nearest_centroid`: computes a per-class mean of examples during training, the classification then assigns class according to the closest mean (McIntyre and Blashfield, 1980).

Table 3: Examples of word sentiment data used in the experiment.

Czech	English	Positive	Negative
neschopný	incapable	0	0.75
čistý	clean	0.5	0
poměrný	proportional	0.25	0.5
hojný	abundant	0.125	0
přijatelný	acceptable	0.625	0
závadný	harmful	0	0.375
přístupný	accessible	0.625	0
zastrčený	inserted	0.125	0
úslužný	obliging	0.75	0

3.2 Evaluation Strategy

The final accuracy scores have been obtained by stratified 3-fold cross validation to evaluate the performance of the classifiers. In the 3-fold cross validation, documents were first grouped by their classes. Each of these classes was then divided into 3 parts. The training set for the investigated classifier then consists of two parts of all groups and the test set consists of the remaining parts of all groups. There are three different ways to select which of the parts will go into the training and the evaluation sets. Each classifier has been evaluated three times, once with each of these ways or folds. The resulting score was computed as the average of the three scores obtained for each of the folds.

Table 4: Classifier prediction accuracy sorted by the weighted F1-score which takes into account imbalanced attribute classes. The resulting accuracy is compared to the baseline accuracy of the majority class.

	best classifier	weighted F1	accuracy	baseline	difference
Demonizing	sgd_l2	.85	.96	.96	.00
Genre	sgd_elasticnet	.84	.96	.89	.07
Server	sgd_l1	.83	.94	.63	.31
Relativizing	sgd_elasticnet	.82	.94	.93	.01
Fear mongering	passive_ aggressive	.81	.92	.91	.01
Opinion	sgd_l2	.79	.89	.86	.03
Focus	ridge	.77	.87	.53	.34
Labelling	ridge	.73	.82	.79	.03
Expert	ridge	.73	.81	.74	.07
Russia	sgd_l1	.71	.82	.70	.12
Emotions	ridge	.70	.81	.81	.00
Fabulation	ridge	.70	.79	.74	.04
Overall sentiment	ridge	.70	.78	.75	.04
Location	sgd_l2	.68	.76	.41	.36
Argumentation	ridge	.65	.75	.69	.06
Blaming	ridge	.65	.72	.59	.13
Topic	sgd_elasticnet	.64	.71	.32	.39
Source	ridge	.63	.70	.60	.10

3.3 Evaluation Metrics

Each trained classifier predicts the class for a document based on its text. By comparing the results to the dataset gold standard data, each of the classifier was evaluated by means of its attribute-related accuracy, precision, recall, and F1 score. The accuracy results are summarized in Table 2 and compared with the `dummy` baseline accuracy in Table 4.

3.4 Correlations of Attributes and Sentiment Coefficients

The set of article attributes contains several items which express sentiment values, either to the article as a whole or to a mentioned politician. We have evaluated the possibility of using the article sentiment analysis to predict the corresponding attribute values for the texts.

The paragraph sentiment analysis results were explicitly expressed as an average score of positivity and negativity of particular words. A list of 6,261 words was prepared as projections of SentiWordNet (Baccianella et al., 2010) scores via the Czech WordNet (Rambousek et al., 2018; Horák et al., 2008) database, see Table 3 for examples. Each paragraph received an average value of *only positive* words, *only negative* words and of their *average score* computed as a difference between word positivity and negativity. The overall document scores were then computed as a maximum positive paragraph score, maximum negative paragraph score and maximum and minimum of the average word score for each paragraph.
Each of the resulting document sentiment scores were evaluated for a correlation[4] with positive and negative values of the selected attributes annotated in the data. The results are presented in Table 5. None of the attributes has proven really strong correlation, but several attributes partly correlate with the maximum negative sentiment of the document. Interestingly, there is no correlation in case of the *emotions* attribute.

4 Conclusion and Future Directions

We have introduced a new benchmark dataset for propaganda manipulative techniques detection in Czech newspaper texts. The dataset contains 7,494 documents annotated for the presence of eight manipulative techniques and 10 document attributes relevant for propaganda detection. The

[4]Computed as Spearman's correlation coefficient with statistical significance.

Table 5: Correlations of selected attributes and document sentiment analysis scores. The † symbol denotes statistically significant values ($p < 0.05$) of Spearman's correlation coefficient.

Attribute	max positive		max negative		max average		min average	
blaming	0.18	†	**0.23**	†	0.17	†	-0.23	†
demonizing	0.11	†	**0.13**	†	0.11	†	-0.12	†
fear mongering	0.16	†	**0.18**	†	0.16	†	-0.18	†
emotions compassion	0.02		-0.00		0.03		-0.00	
emotions fear	-0.07	†	0.02		-0.07	†	-0.02	
emotions hate	0.06	†	0.04		0.06		-0.04	
emotions grievance	-0.00		-0.05		-0.00		0.05	
overall sentiment	0.16	†	**0.18**	†	0.16	†	-0.18	†
attitude1	0.04	†	**0.04**	†	0.04	†	-0.04	†
attitude2	0.10	†	**0.15**	†	0.09	†	-0.16	†
attitude3	0.13	†	**0.13**	†	0.11	†	-0.13	†
attitude avg	0.13	†	**0.14**	†	0.11	†	-0.15	†

dataset is currently being expanded with the third part of documents from 2018 and it is planned to be released for public access after this expansion.

We have evaluated the current data with 10 current classification techniques. Regularized linear regression and Support vector machines are able to classify the data with the best accuracies, even though the manipulative techniques need to employ extra features to significantly improve over the baseline.

In the currently running experiments, we are preparing new evaluation of the dataset using detailed stylometric features and distributed semantic representations of the texts.

Acknowledgments.

This project was partially supported by the Grant Agency of Masaryk University within the project MUNI/G/0872/2016.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of LREC 2010*. pages 2200–2204.

Vít Baisa, Ondřej Herman, and Aleš Horák. 2017. Manipulative Propaganda Techniques. In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2017*. pages 111–118.

Leo Breiman. 2001. Random forests. *Machine learning* 45(1):5–32.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research* 7(Mar):551–585.

Lisa K Fazio, Nadia M Brashier, B Keith Payne, and Elizabeth J Marsh. 2015. Knowledge does not protect against illusory truth. *Journal of Experimental Psychology* 144(5):993–1002.

Todd C Helmus, Elizabeth Bodine-Baron, Andrew Radin, Madeline Magnuson, Joshua Mendelsohn, William Marcellino, Andriy Bega, and Zev Winkelman. 2018. *Russian Social Media Influence: Understanding Russian Propaganda in Eastern Europe*. Rand Corporation.

Edward Herman and Noam Chomsky. 2012. A propaganda model. *Media and cultural studies: Keyworks* pages 204–230. Reproduced from Manufacturing Content, 1988.

Aleš Horák, Karel Pala, and Adam Rambousek. 2008. The Global WordNet Grid Software Design. In *Proceedings of the Fourth Global WordNet Conference, University of Szegéd*. pages 194–199.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography* 1(1):7–36.

Andrew McCallum, Kamal Nigam, et al. 1998. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*. pages 41–48.

Robert M McIntyre and Roger K Blashfield. 1980. A nearest-centroid technique for evaluating the minimum-variance clustering procedure. *Multivariate Behavioral Research* 15(2):225–238.

Jan Michelfeit, Jan Pomikálek, and Vít Suchomel. 2014. Text tokenisation using unitok. In Aleš Horák and Pavel Rychlý, editors, *RASLAN 2014*. Tribun EU, Brno, Czech Republic, pages 71–75.

Christopher Paul and Miriam Matthews. 2016. The Russian ”firehose of falsehood” propaganda model. *Rand Corporation* pages 2–7.

Adam Rambousek, Aleš Horák, and Karel Pala. 2018. Sustainable long-term WordNet development and maintenance: Case study of the Czech WordNet. *Cognitive Studies/Études cognitives* (18).

Ryan M Rifkin and Ross A Lippert. 2007. Notes on regularized least squares. *Computer Science and Artificial Intelligence Laboratory Technical Reports* https://dspace.mit.edu/handle/1721.1/37318.

Pavel Šmerk. 2009. Fast Morphological Analysis of Czech. In Petr Sojka and Aleš Horák, editors, *Third Workshop on Recent Advances in Slavonic Natural Language Processing*. Masaryk University, pages 13–16.

Pavel Šmerk. 2010. *K počítačové morfologické analýze češtiny (in Czech, Towards Computational Morphological Analysis of Czech)*. Ph.D. thesis, Faculty of Informatics, Masaryk University.

Samuel C Woolley and Philip N Howard. 2017. Computational propaganda worldwide: Executive summary. *Working Paper* 2017(11).

Tong Zhang. 2004. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*. ACM, page 116.

Jianqing Zhang. 2013. *The Propaganda Model and the Media System in China*. Dartmouth College.