Machine Learning Approach to Fact-Checking in West Slavic Languages

Pavel Přibáň[1,2], **Tomáš Hercig**[2], and **Josef Steinberger**[1]

[1]Department of Computer Science and Engineering,
Faculty of Applied Sciences, University of West Bohemia, Czech Republic
[2]NTIS – New Technologies for the Information Society,
Faculty of Applied Sciences, University of West Bohemia, Czech Republic
{pribanp,tigi,jstein}@kiv.zcu.cz
http://nlp.kiv.zcu.cz

Abstract

Fake news detection and closely-related fact-checking have recently attracted a lot of attention. Automatization of these tasks has been already studied for English. For other languages, only a few studies can be found (e.g. (Baly et al., 2018)), and to the best of our knowledge, no research has been conducted for West Slavic languages. In this paper, we present datasets for Czech, Polish, and Slovak. We also ran initial experiments which set a baseline for further research into this area.

1 Introduction & Motivation

Fake news is designed to incite agitation against an individual or a group of people. Its aim is to influence and manipulate public opinion on targeted topics. Fake news detection, including fact-checking, which can be used as the first step of a detection system, are currently receiving a lot of attention in the research community and journalism.

This attention is apparent from the rise of fact-checking websites that verify mainly political claims (see the list of signatories of the code of principles of the International Fact-Checking Network[1]). Research related to these tasks is on the rise in a variety of fields, including natural language processing, machine learning, knowledge representation, databases, and journalism (Thorne and Vlachos, 2018).

The automation of these tasks or their parts would greatly benefit journalism and perhaps help the public to verify the credibility of various media. It is evident that fact-checking needs external knowledge or detailed context. However, in order to achieve the goal of a robust automatic fact-checking system, we must first find a way how to evaluate such a system. For English, there are publicly available datasets that researchers can use to evaluate their systems. However, no systematic research has been conducted in West Slavic languages yet; thus we establish a common ground for further research by providing large datasets for fact-checking in Czech, Polish, and Slovak languages including initial experiments which reveal the complexity of the task. We set a baseline which uses standard machine learning approach, and set an upper bound which uses manually created external knowledge.

2 Related Work

This section presents a brief overview of related work, for a more detailed survey, please refer, for example to Thorne and Vlachos (2018).

For the development of the first fact-checking systems, Vlachos and Riedel (2014) manually labeled a dataset and defined fact-checking as the assignment of a truth Boolean value to a claim made in a particular context. They also discussed baseline approaches to fact-checking.

Wang (2017) presented a dataset of 12.8K manually labeled statements from the Politifact[2] website. He experimented with logistic regression, support vector machines, Long Short-Term Memory neural networks (LSTM), and convolutional neural networks (CNN). He then introduced a modified neural network architecture integrating text with other meta-data. He performed similar experiments to our work on English dataset with six labels achieving 27.7% accuracy as the best result.

Tacchini et al. (2017) showed that fake news

[1]https://ifcncodeofprinciples.poynter.org/signatories

[2]https://www.politifact.com/

Proceedings of Recent Advances in Natural Language Processing, pages 973–979,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_113

Language	MISLEADING	UNVERIFIABLE	FALSE	TRUE	ALL
Czech	848 (9.3%)	1343 (14.8%)	1222 (13.5%)	5669 (62.4%)	9082 (100%)
Polish	313 (11.0%)	113 (4.0%)	648 (22.9%)	1761 (62.1%)	2835 (100%)
Slovak	1146 (9.1%)	1751 (13.9%)	1670 (13.3%)	7987 (63.6%)	12554 (100%)

Table 1: Dataset label statistics.

Lang.	M.	U.	F.	T.	ALL
CS	39 / 44	38 / 44	33 / 39	33 / 38	34 / 39
PL	28 / 32	19 / 25	19 / 24	22 / 26	22 / 26
SK	36 / 40	36 / 40	29 / 33	32 / 36	32 / 37

Table 2: Dataset size in words (median/average).

could be detected based on user likes. Using an adaptation of a Boolean label crowdsourcing algorithm, they were able to detect hoaxes with 99% accuracy. Their dataset consists of 15.5K posts (58% fake news, 42% real news) with over 2,300K likes from 900K users.

Jin et al. (2017) focused on detecting fake news on Twitter related to the U.S. presidential elections. They labeled the data according to the Snopes[3] website. They analysed tweets of followers of the presidential candidates.

Yang et al. (2018) used a dataset of 20K news (12K fake news, 8K real news) for fake news detection. They used a modified convolutional neural network trained using the title, images and text of the news articles, making use of both explicit and latent features to detect fake news. They achieved F_1-measure of 92% overcoming a baseline LSTM text-based model by 3%. They presented a thorough analysis of the dataset, including text style and image resolution.

In this paper we present the following novel contributions:

1. The availability of multi-lingual data for non-English languages is lacking. Our paper addresses this need.

2. The dataset also contains reasoning for labeling each claim - this can be used in future research, e.g. argumentation mining.

3. The claims are also labeled by Political party affiliations - this may facilitate fine-grained analysis.

[3] https://www.snopes.com/

3 Dataset

We provide three datasets for fact-checking - one for each language downloaded from the following fact-checking websites.

- Czech (https://demagog.cz/)
- Polish (http://demagog.org.pl/)
- Slovak (http://www.demagog.sk/)

Each dataset contains claims of politicians[4] annotated with one of four classes: FALSE, TRUE, UNVERIFIABLE, and MISLEADING. The labels have the following meaning:

- FALSE These statements are not in line with publicly available numbers or information. It may also be a situation where the calculation method of the indicator differs, but none of these sources confirms the number or claim in question.

- TRUE Statement using the right information in the right context.

- UNVERIFIABLE If it is not possible to find the source of the claim, or it is not possible to confirm or refute it based on the available information.

- MISLEADING These are statements that use correct facts, but in a wrong or incomplete context, or are being torn out or otherwise distorted from the original context. These are inappropriate or disproportionate comparisons.

The labels are manually annotated by the authors of the corresponding language websites. The dataset also contains information about the speaker and his or her political affiliation. The reasoning[5] for the given label is also included in the dataset. The data were downloaded from the respective websites in April 2018. The following example has been translated into English.

[4] Other publicly active people such as journalist are included in the dataset as well.

[5] Including external knowledge.

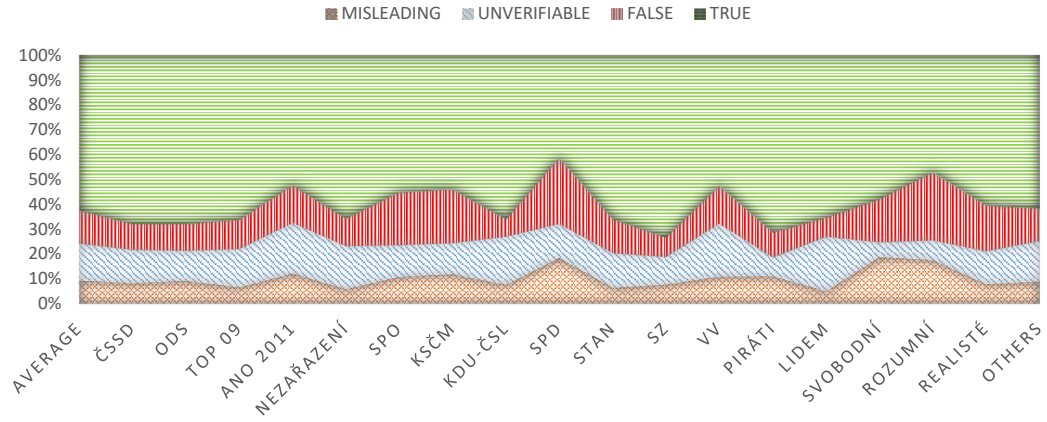

Figure 1: Czech Political Parties Statistics

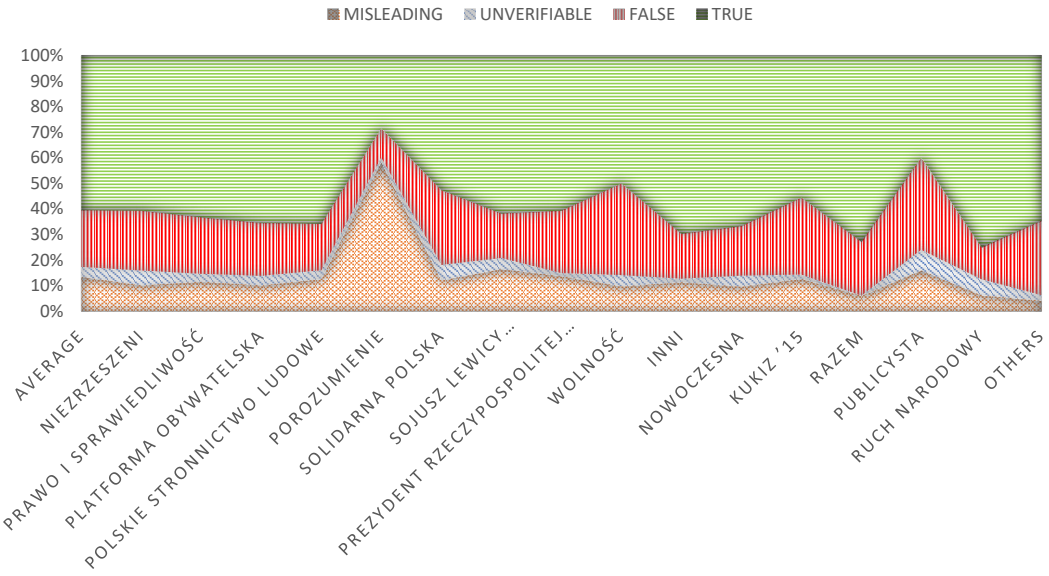

Figure 2: Polish Political Parties Statistics

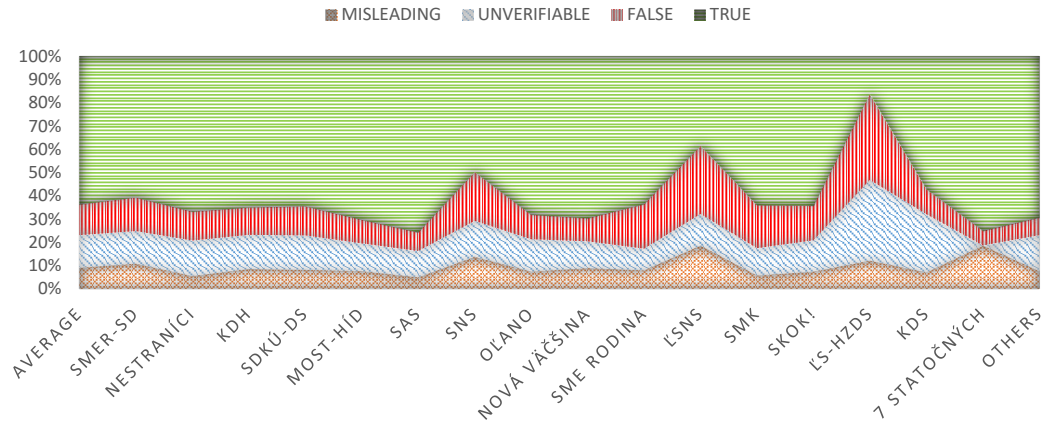

Figure 3: Slovak Political Parties Statistics

Miloš Zeman (SPO) → FALSE
CLAIM: "The Swedes have seven million inhabitants."
REASONING: Sweden has according to the latest official data from November 2017 10,113,000 inhabitants.

The data distribution, according to the labels, is shown in Table 1. Table 2 shows the median and the average number of words in a claim.

We compare the label distribution among selected political parties with the most claims. Figures 1, 2, and 3 show the average label distribution and the distribution for the selected political parties sorted by a number of claims for Czech, Polish, and Slovak languages. Note that the labels *Nezařazení* for Czech, *Niezrzeszeni* for Polish, and *Nestraníci* for Slovak represent claims of people without any political party affiliation. The `Other` label is the average of the rest[6] of the political parties present in the dataset [7].

It is clear that the claims of some political parties often tend to be truth compared to other parties. This phenomenon can be observed for all three languages. The opposite applies to the Czech parties *SPD*, *Rozumní*, Polish party *Porozumienie* and Slovak party *L'S-HZDS*. However, the inconsistency of `UNVERIFIABLE` label across languages was more surprising. We believe that it is caused by differences in labeling i.e. that in the Polish dataset the `UNVERIFIABLE` label is used only under stringent rules in comparison with the other two languages.

4 Experiments

We performed identical classification experiments for each language to allow a comparison for future research. The main goal of these experiments is to illustrate the complexity of the task and to set a baseline for these datasets.

We use 10-fold cross-validation for the evaluation of both balanced and imbalanced datasets. We also perform binary experiments only with `FALSE` and `TRUE` classes. The input for the classifier is either the text of a claim or a text of a claim supplemented by the reasoning text. Experiments using the reasoning text set up an upper bound of performance that can be achieved with an automatic approach. Our evaluation metrics are macro-average F_1 score and accuracy.

The reasoning text often contains words or phrases which are strictly related to the assigned label, for example, *Výrok je pravdivý* (The statement is true) for the `TRUE` label or *Výrok nelze ověřit* (The statement is unverifiable) for the `UNVERIFIABLE` label. We call these words *give-away words* as they alone will be a sufficient source of information for the classifier. In other words, the reasoning text in a large number of cases de facto contains the label.

We removed these words from the reasoning text and repeated the experiments with the modified reasoning text. The list of removed give-away words was manually selected from the words with highest label occurrence ratio[8]. All words were selected only if they occurred at least 20 times in the corresponding label class. Finally, we manually chose words and removed them from the reasoning text, see Table 3 that contains examples of the removed give-away words. For Czech, we removed $9,601$ words out of $1,552,878$, for Slovak, we removed $9,147$ words out of $2,146,465$ and for Polish, we removed 573 words out of $367,435$. The complete list of the removed words is available at `http://nlp.kiv.zcu.cz/projects/fact-checking`.

Czech	Polish	Slovak
nepravdivý	fałszywą	nepravdivý
pravdivíy	prawdziwą	pravdivý
neověřitelný	nieweryfikowalną	neoveriteľný
neodpovídá	manipulację	nevieme

Table 3: Examples of give-away words.

4.1 Models Settings

The preprocessing includes tokenization using NLTK *TreebankWordTokenizer* (Bird et al., 2009), text lowercasing, removing HTML tags and entities. No other preprocessing steps are employed. We use Logistic Regression classifier from the LIBLINEAR library (Fan et al., 2008) with penalty parameter $C = 1$ and L2 regularization (see Fan et al. (2008) for detailed description), along with

[6]The rest of the parties that had fewer claims than the selected parties. In the Czech dataset, this includes the `null` value used for people who changed parties over time.

[7]The dataset is available for research purposes at `http://nlp.kiv.zcu.cz/projects/fact-checking`

[8]The number of occurrences of words for a given label divided by the total frequency. We selected words with a ratio ≥ 0.8 for the `TRUE` label, and words with a ratio ≥ 0.6 for the other three labels.

		Czech		Polish		Slovak	
Dataset	Labels	Macro F_1	Accuracy	Macro F_1	Accuracy	Macro F_1	Accuracy
Imbalanced	4	0.21 / 0.19	0.25 / 0.62	0.21 / 0.19	0.25 / 0.62	0.21 / 0.19	0.25 / 0.64
Balanced	4	0.25 / 0.25	0.25 / 0.25	0.25 / 0.25	0.25 / 0.25	0.25 / 0.25	0.25 / 0.25
Imbalanced	2	0.35 / 0.45	0.35 / 0.82	0.34 / 0.42	0.34 / 0.73	0.33 / 0.45	0.33 / 0.83
Balanced	2	0.50 / 0.50	0.50 / 0.50	0.50 / 0.50	0.50 / 0.50	0.50 / 0.50	0.50 / 0.50

Results of *random / majority* class classifiers.

Table 4: Results of a random and majority class (separated by slash *random / majority*) classification. For example, the accuracy for Czech imbalanced dataset for all four labels is 0.25 for the random classifier, 0.62 for the majority class classifier.

		Czech		Polish		Slovak	
Dataset	Labels	Macro F_1	Accuracy	Macro F_1	Accuracy	Macro F_1	Accuracy
Imbalanced*	4	0.26	0.61	0.25	0.60	0.27	0.62
Balanced*	4	0.31	0.31	0.26	0.26	0.35	0.35
Imbalanced*	2	0.48	0.81	0.49	0.72	0.51	0.82
Balanced*	2	0.57	0.57	0.54	0.55	0.58	0.58
Imbalanced†	4	0.87	0.91	0.45	0.69	0.79	0.86
Balanced†	4	0.85	0.85	0.46	0.47	0.78	0.78
Imbalanced†	2	0.88	0.94	0.63	0.77	0.86	0.93
Balanced†	2	0.86	0.87	0.64	0.64	0.85	0.85
Imbalanced‡	4	0.51	0.72	0.36	0.64	0.53	0.72
Balanced‡	4	0.54	0.54	0.41	0.43	0.56	0.56
Imbalanced‡	2	0.65	0.85	0.59	0.74	0.67	0.85
Balanced‡	2	0.71	0.71	0.61	0.61	0.68	0.68

* dataset only with claim

† dataset with both claim and reasoning.

‡ dataset with both claim and reasoning without give-away words.

Table 5: Results of logistic regression classification.

unigram and bigram features. Experiments with the reasoning are performed on a combination of the claim text and the reasoning text. First, the reasoning text and the claim text are concatenated, and then we extract the unigram and bigram features. These features are used as an input to the classifier.

4.2 Results

We report results for the experiments for all three languages, including results of a random and majority class classification in Table 4.

In Table 5 we show the results for the Logistic Regression classifier on the balanced and imbalanced datasets for the following text combinations:

- claim
- claim & reasoning
- claim & reasoning without give-away words

On the balanced dataset we can see that using only unigrams and bigrams as features is not enough for the classifier as the results are only slightly better than the majority baseline; thus more sophisticated methods are needed to extract the information contained in the reasoning part of the dataset.

We can see that the results achieved on both claim and reasoning are very high (F_1 0.87, accuracy 0.91 for Czech) confirming our hypothesis that in ideal conditions this task could be solved

by a machine learning algorithm. In the case of using only the claim on the balanced binary dataset, accuracy drops to 0.57, 0.55, and 0.58 for Czech, Polish, and Slovak, respectively. Polish appears as the most challenging language as the results are lower compared to the other two languages. One reason could be a smaller size of the Polish dataset (see Table 2).

In the case of experiments with the *give-away words*, results are still much higher than in experiments where only the claim was used (F_1 0.53, accuracy 0.72 for Slovak). We observed the highest performance drop for experiments with all four labels, especially for Czech, in comparison to the experiments with the original reasoning text. The performance of the Polish model was least affected. This was caused by the low number of removed words (573 words out of 367,435 in total) for Polish. Thus the assumption that the reasoning contains only give-away words is false; leading us to believe that some information about the validity of the claim is contained in the reasoning.

5 Conclusion

This paper represents the initial research of fact-checking in Czech, Polish, and Slovak languages.

- We presented datasets for fact-checking in three West Slavic languages and provided them to the research community.
- We ran initial experiments which revealed baseline results for further research.

It is clear that this task is very challenging. However, we showed that when a machine learning approach uses label reasoning in addition to the claim, it can perform very well. Although such human-written reasoning rather sets a performance upper bound, the way to go forward might include generating such reasoning automatically using external data.

Disclaimer

Any views, findings, and conclusions expressed in this article are based on a thorough analysis of very limited and dated data. They do not necessarily reflect the views, opinions, or official positions of the authors or the University of West Bohemia.

Acknowledgments

This work has been supported by Grant No. SGS-2019-018 Processing of heterogeneous data and its specialized applications, by the project LO1506 of the Czech Ministry of Education, Youth and Sports under the program NPU I and from ERDF "Research and Development of Intelligent Components of Advanced Technologies for the Pilsen Metropolitan Area (InteCom)".

References

Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018. Integrating Stance Detection and Fact Checking in a Unified Corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, pages 21–27. https://doi.org/10.18653/v1/N18-2004.

S Bird, E Loper, and E Klein. 2009. Natural language processing with Python:“O’Reilly Media Inc.“ .

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research* 9:1871–1874.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, Yu Wang, and Jiebo Luo. 2017. Detection and Analysis of 2016 US Presidential Election Related Rumors on Twitter. In Dongwon Lee, Yu-Ru Lin, Nathaniel Osgood, and Robert Thomson, editors, *Social, Cultural, and Behavioral Modeling*. Springer International Publishing, Cham, pages 14–24.

Eugenio Tacchini, Gabriele Ballarin, Marco L Della Vedova, Stefano Moret, and Luca de Alfaro. 2017. Some like it hoax: Automated fake news detection in social networks. *CoRR* abs/1704.07506. http://arxiv.org/abs/1704.07506.

James Thorne and Andreas Vlachos. 2018. Automated Fact Checking: Task Formulations, Methods and Future Directions. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, pages 3346–3359. http://aclweb.org/anthology/C18-1283.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. Association for Computational Linguistics, Baltimore, MD, USA, pages 18–22. http://www.aclweb.org/anthology/W14-2508.

William Yang Wang. 2017. “Liar, Liar Pants on Fire“”: A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, pages 422–426. https://doi.org/10.18653/v1/P17-2067.

Yang Yang, Lei Zheng, Jiawei Zhang, Qingcai Cui, Zhoujun Li, and Philip S. Yu. 2018. TI-CNN: Convolutional Neural Networks for Fake News Detection. *CoRR* abs/1806.00749. http://arxiv.org/abs/1806.00749.