Fine-grained Czech News Article Dataset: An Interdisciplinary Approach to Trustworthiness Analysis

Matyáš Boháček[1,2], Michal Bravanský[1,3], Filip Trhlík[1,3] and Václav Moravec[1]

[1]*Faculty of Social Sciences, Charles University, Prague, Czech Republic*
[2]*Gymnasium of Johannes Kepler, Prague, Czech Republic*
[3]*University College London, United Kingdom*

Abstract

We present the Verifee Dataset: a novel dataset of news articles with fine-grained trustworthiness annotations. We develop a detailed methodology that assesses the texts based on their parameters encompassing editorial transparency, journalist conventions, and objective reporting while penalizing manipulative techniques. We bring aboard a diverse set of researchers from social, media, and computer sciences to overcome barriers and limited framing of this interdisciplinary problem. We collect over $10,000$ unique articles from 60 Czech online news sources. These are categorized into one of the 4 classes across the credibility spectrum we propose, raging from entirely trustworthy articles all the way to the manipulative ones. We produce detailed statistics and study trends emerging throughout the set. Lastly, we fine-tune multiple popular sequence-to-sequence language models using our dataset on the trustworthiness classification task and report the best testing F-1 score of 0.52. We open-source the dataset, annotation methodology, and annotators' instructions in full length at https://verifee.ai/research to enable easy build-up work. We believe similar methods can help prevent disinformation and educate in the realm of media literacy.

Keywords
disinformation detection, low-resourced language, dataset

1. Introduction

Donald Trump has called journalists and news outlets "fake news" nearly $2,000$ times since the beginning of his presidency, averaging more than one daily broadside against the press between 2016 and 2020 [1]. Because of Trump, the term fake news underwent a fundamental change in its meaning. At first, it referred to a satirical and ironic genre of fictional news designed to entertain the audience. The original "fake news" have appeared on TV shows such as Saturday Night Live on NBC or in print, such as The Onion. However, during Trump's campaign for

the US presidential election in 2016 and his presidency, the concept of fake news became an integral part of his political communication. It aimed to discredit critical journalistic content or the whole news media as "fake media." The successful stigmatization strategy of "fake news" has led to a fascination with this phenomenon in the public discourse and science.

Fake news has become a label for false news and a synonym for both disinformation and misinformation. This has strengthened the binary perception of the credibility of information in a true-false dichotomous perspective. However, this reductionist approach has become a barrier to understanding the more profound meaning that the buzzword "fake news" covers. If we want to examine the credibility of the news content seriously, it is not possible to adopt the binary approach of either truth or lie. By creating the Verifee Dataset, we try to overcome the interdisciplinary barrier between social sciences (especially journalism and media studies) and computer science. This barrier prevents specialists in automated or robotic journalism from adopting a more analytical approach to various types of information disorders that we have become used to labelling with the general term "fake news".

2. Related Work

Herein, we first review the current literature focusing on disinformation and misinformation in the journalistic ambit. We later provide an overview of existing methods treating these phenomena within the Artificial intelligence (AI) and Natural language processing (NLP) research communities. We first list some of the already-available datasets and then focus on the architectures solving the tasks of fake news detection and automatic fact-checking.

The task of fake news detection resides in classifying whether a given news article (or occasionally another medium such as a Tweet) is considered fake (disinformative) or truthful (credible). There is no consensus in the literature on what specific parameters derive from these states, but truthfulness is usually considered the primary one. Some approaches recognize more fine-grained scales with specific classes, such as tabloid news, mixed reliability news, whereas others only recognize fake and credible news. Either way, this task extracts the class from the text on its own.

The task of automatic fact-checking, on the other hand, requires a source of truth to which the news article is compared. The task then lies in determining whether the article is supported by facts therein. Hence, one can consider this task a specific abbreviation of stance detection focusing on news media and large-scale ground-truth databases.

We review datasets and approaches in both of these tasks, as our dataset lies somewhere in between.

2.1. Disinformation, Misinformation

With the advent and development of digital network media at the beginning of the 21st century, there has been a dynamic spread of unverified, inaccurate, or false information (raging from textual to audiovisual), which is referred to as information disorders. Information disorders as part of information pollution are thus in direct contrast to trustworthy content that is accurate, factually correct, verified, reliable, and up-to-date. According to the media and journalism theorist [2], it is misleading to label information disorders with the umbrella term "fake news."
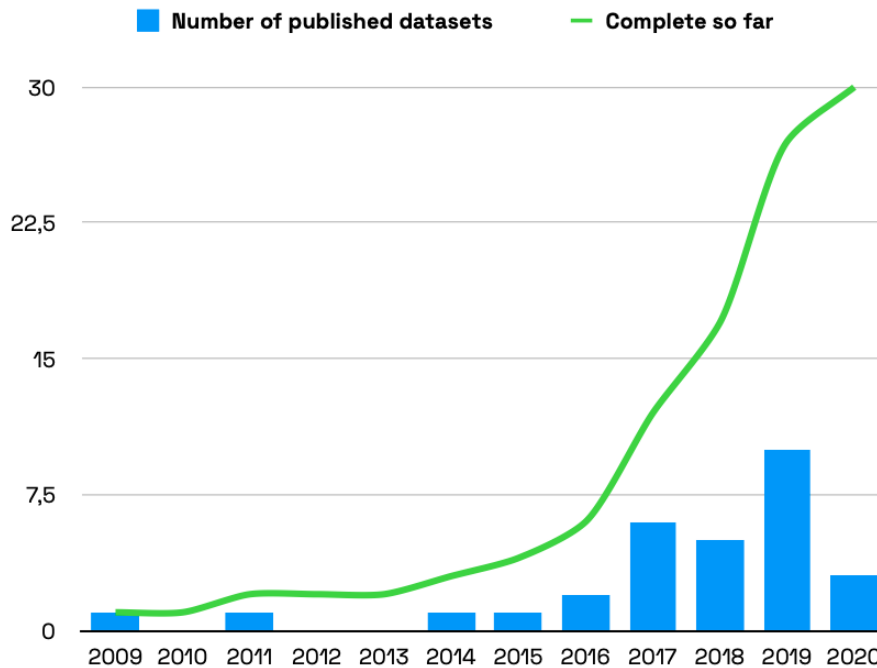
Figure 1: Continual statistics on disinformation classification datasets publishing throughout the years 2009-2020. The bar charts denote the number of new datasets published in the respective year, while the overlay line captures the cumulative number of datasets published until that year. Years are plotted on the x-axis, and the numbers of published datasets are represented on the y-axis.

Although the definition of fake news is complicated, it is possible to define at least seven criteria that contribute to the contamination of information to such an extent that the use of the term information disorder is appropriate.

Satire/parody as the least problematic form of information pollution and, therefore, a factor reducing the credibility of news content is on the one end of the seven-scale spectrum. In contrast, fictional content that was created for the intentional dissemination of false information lies at the other end. Wardle introduces a typology of the three main information disorders based on the seven criteria. The typology is established on the degree of truth/falsity and the intention to cause harm. Erroneous, inaccurate, or untrue content that is not intended to harm recipients because it reflects, for example, ignorance of the disseminator is referred to as misinformation. This term includes satire, parody, or misleading texts, images, or quotes. False or untrue content that is distributed to deceive or manipulate its recipients, whether for financial, ideological, political, social, or psychological reasons, is referred to as disinformation. This term includes malicious lies, fabricated information, disinformation campaigns, etc. Finally, true information disseminated with the intention to cause harm (for example, by revealing a person's religion, sexual orientation, etc.) is referred to as malinformation.

The conceptual framework of individual information disorders in the professional literature is relatively inconsistent. Thus, part of the scientific community [3] considers disinformation
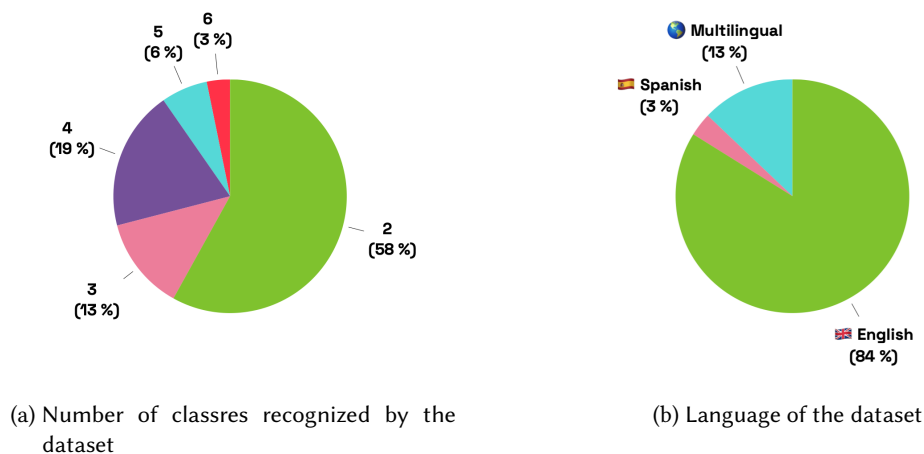
(a) Number of classres recognized by the dataset

(b) Language of the dataset

Figure 2: Proportional statistics of the available disinformation classification datasets.

"misinformation with an attitude," while attitude is the aforementioned deliberate deception of recipients. According to another approach [4, 5], disinformation is part of misinformation because it is difficult to demonstrate the intention (not) to spread it. In both cases, the notion of misinformation encompasses the term disinformation. However, one can also encounter a more subtle division of individual forms of information disorders [6]. In addition to the terms disinformation and misinformation, the authors also distinguish autonomous terms such as rumor, conspiracy, hoax, propaganda, opinion spam, false news (i.e., fake news), clickbait, satire, etc. Within the classification of information disorders, we can perceive disinformation and misinformation as overarching concepts because disinformation can take the form of clickbait, rumor, hoax, opinion spam, or conspiracy theory. Similarly, misinformation can be based on rumors or satire.

2.2. Disinformation-related datasets

D'Ulizia et al. [7] have conducted a thorough study on fake news detection datasets. We highlight three of these based on the traction within the research community and direct the reader to this review for more detail.

Wang [8] created the LIAR dataset with 12, 836 text excerpts of 6 classes. Later, Nørregaard et al. [9] published NELA-GT dataset containing 713,000 news articles belonging to 2 classes. Lastly, Slovikovskaya and Attardi [10] presented the FNC-1 dataset with 49, 972 news articles classified into 4 labels. All these datasets are in English.

Guo et al. [11] have presented a survey of the current fact-checking datasets. Once again, we mention some of these below and refer the reader to the study for more detail.

First, Mitra and Gilbert [12] created the CredBank dataset with over 1, 000 English Tweets classified into 5 labels. Multiple works followed, including the much larger Suspicious dataset [13] containing over 130, 000 English Tweets with 2 assigned classes. Lastly, Nakov et al. [14] presented the CheckThat21-T1A dataset with over 17, 000 Tweets of 2 classes. These Tweets

come from multiple languages.

To capture the trends in the publishing intensity of these datasets, we include a plot of all the fake news detection datasets from D'Ulizia et al. [7] by their year of origin in Figure 1. This shows that the popularity of this task in the AI and NLP community is very much a recent phenomenon, corresponding to the general focus on the topic of disinformation in public. However, the sizeable collective excitement goes hand-in-hand with the inconsistency of the problem's framing and methodologies. This can be easily demonstrated with Figure 2a, which captures the distribution of these datasets by the pure number of labels into which they classify the articles. Furthermore, we see significant inconsistencies in the methodologies leading to these classifications if we go deeper. Some works [9] derive the class based on the high-level credibility assessment of its source (i.e., whenever they deem a source problematic, they treat all its articles in this manner, leaving no room for exceptions). Others [8, 10] treat the articles on an individual basis. Alongside, all of these differ in the specific features deducing the classification. Some consider the context of the article and editorial proprieties, while others only use the texts and its attributes.

Moreover, other major problematic characteristics of the dataset population emerge. Despite disinformation being a global threat, the vast majority of these datasets are in English only, as can be seen in Figure 2b. Alarmingly, most of the datasets did not include professionals or academics from the relevant fields, such as the media sciences and humanities in general. We believe that this calls for establishing a robust and uniform methodology for approaching the problem of disinformation holistically and an emphasis on developing datasets for non-English speaking regions with the oversight of relevant experts across domains and industries. We are addressing all of these in this in our work.

2.3. Automated fake news detection

The task of automated fake news detection has been usually approached by fine-tuning general-purpose language models, such as BERT [15], ELECTRA [16], or RoBERTa [17]. Specific architectures for this task have been studied in the literature, too. [18], for instance, provide additional parameters such as political bias, the domain from the article's originating URL, and prior information about the domain as inputs to their model. [19] create the first multi-modal architecture for this task as they combine the texts at the input with images included in the article.

2.4. Automated fact-checking

Architectures for automated fact-checking usually consist of an evidence retrieval module and a verification module [20]. Recent dense retrievers with learned representations and fast dot-product indexing [21, 22] have shown strong performance, too. There have also been approaches considering multiple texts with potential evidence for the claims as a single evidence piece by concatenating them [23, 24]. Later, an entailment model is employed to determine whether the article's text is supported or refuted by the evidence. We refer the reader to [11] for a concise overview of such methods.

3. Trustworthiness Assessment Methodology

Having familiarized ourselves with the current state of research, we concluded that the best way forward is to build upon the previous work and introduce a new language-agnostic methodology for classifying news articles. The primary motivation for this was the inability of prior approaches to fully reflect the complexity of the problem in terms of media studies and fully appreciate each article uniquely and independently of its source. We hope to provide better data for AI-based tools concerned with handling dubious news articles with this methodology. Below, we introduce the basic framework of our methodology. Its complete overview is available in Appendix A.

3.1. Trustworthiness

To strengthen clear division between the fake news detection and fact-checking tasks, our methodology focuses solely on the content aspects of the article. We hence do not reflect the truthfulness or context of the news, as we believe such practices fall under the latter task. These parameters on their own serve as robust evidence of an article being disinformative [25].

Despite not determining the factualness, we can fully assess how deceptive it is and hence deem its trustworthiness. By omitting context and focusing solely on trustworthiness, we aim to improve the annotation process since there is no requirement for outside information and the class is final (i.e., unlike with methods employing truthfulness, no later information can reverse it).

3.2. Classes

Our methodology strays from the binary classification of true or fake news and allows for more granularity of class definitions because its sole focus is the assessment of trustworthiness.

To quantify trustworthiness, we propose 15 negative linguistic attributes of an article (e.g., hate speech, clickbait title, logical fallacies) and 6 positive ones (e.g., real author, references, objective profiling). With these, we define the following classes of trustworthiness:

1. **Trustworthy**: These articles are credible and make no effort to deceive the reader. When relevant, they cite their sources of information and present the opinions of all involved parties. In terms of our framework, this means that they do not contain any negative attributes while having at least five positive ones.
2. **Partially Trustworthy**: These articles still do not deceive their reader, but they often attempt to exaggerate the topic while making less effort to uphold journalistic norms. They often contain clickbait headlines and appeal to the readers' feelings. When translated into our framework, these include 2 to 5 negative attributes.
3. **Misleading**: These articles contain deception outside the boundaries of pure conspiracies. These articles often alter the framings of the news to fit their agenda. Any article containing 6 to 8 negative attributes belongs to this class.
4. **Manipulative**: These articles strive to manipulate their reader. Hence, their arguments often use conspiracy narratives. They contain over 8 negative attributes or 3, especially problematic ones, such as using conspiracies or hate speech.

4. Verifee Dataset

Using our newly introduced methodology, we collected a dataset of over $10,000$ Czech news articles classified into the just-described categories. Apart from the class, each entry in the dataset consists of the article's text, HTML source, title, description, authors, keywords, source name, URL, covered controversial topics, and included images. We open-source the dataset at https://verifee.ai/research under a custom license[1]. We provide pre-defined train (80 %), validation (10 %), and (10 %) testing splits that have been assigned randomly. Below, we describe the process of the dataset's collection.

4.1. Scraping and Pre-processing

Initially, we assembled a collection of almost $94,000$ articles by scraping URLs of 60 Czech news sources obtained from Common Crawl[2]. These sources included mainstream journalistic websites, tabloids, independent news outlets, and websites that are part of the disinformation ecosystem [26], capturing the full scope of journalistic content in the Czech Republic. Their complete list can be found in Appendix B.

4.1.1. Enrichment

Next, we determined the category (opinion, interview, general) and the topic (general, sport, economics, hobby, tabloid) of each article through pattern matching. We detected mentions of any controversial topics relevant to the Czech media context (Russia, Covid-19, EU, NATO, USA, and migration) similarly. Additionally, we ascertained whether the article disposes of a real author via an out-of-the-box Named Entity Recognition model [27] for the Czech language.

4.1.2. Filtering

We applied multiple filters and balancing mechanisms to mitigate deficiencies caused by inherent flaws in Common Crawl, which reduced the dataset's size from $94,000$ to $10,000$ items. This way, we also ensured that the data is as representative of the Czech news ecosystem and as diverse as possible. The factors used for filtering were:

- **Length of the text**: Only articles with a length between 400 and $10,000$ characters were included.
- **Category**: For mainstream media, we filtered opinion pieces. However, we kept these for alternative news sources, as the line between reporting and conveying opinion is often blurred here. Interviews were excluded in both cases.
- **Source**: We selected articles in such a way that all sources are as balanced as possible, no matter their actual distribution in the media ecosystem.

---

[1]Our license — building on top of Creative Commons BY-NC-SA (https://creativecommons.org/licenses/by-nc-sa/2.0/) — is available at https://www.verifee.ai/files/license.pdf.

[2]https://commoncrawl.org

- **Topic**: Articles concerning hobbies and sports each form only 5% of the dataset, as they typically are not connected to disinformation. The rest of the topics (general, economic, and tabloid) each form 30% of the dataset.
- **Controversial topics**: We made sure to balance the coverage of controversial topics by including the same number of such articles from mainstream and alternative or extremely opinionated news sources.

4.2. Annotations Organization

We conducted two rounds of annotation. The first round involved $7,347$ unique articles, where just the class was denoted to each article. The second round included $2,655$ unique articles. This time, annotators were asked to provide both the class and flag any problematic attributes of each article defined in our methodology. This enabled us to examine the importance of the various metrics in the methodology. Every annotator was assigned 40 articles in both rounds.

4.2.1. Annotators

All the raters were students of journalism who are native speakers of the Czech language. They thus had a more advanced understanding of the topic of news credibility than the general population. We have to note that due to their age [28] and education [29], their possible bias toward more progressive/liberal schools of thought may have influenced the rating of topics in corresponding areas. We briefed all the annotators on an extensive seminar, provided all of them with detailed materials, and encouraged them to come forward with any problems.

4.2.2. Platform

We adapted the open-source tool Doccano[3] for our task and used it in the collection. Inside the application, annotators were presented with one article at a time in its HTML form with all images included. The platform allowed the user to add necessary tags and comments to each piece.

4.2.3. Source Identity Masking

We masked any elements in the article that would enable the annotators to identify the source or author of the text. Specifically, we replaced their mentions with placeholders. As a result, annotators' media and author preferences could not influence their evaluation.

4.2.4. Monitoring

We tracked the raters' activity on the platform during the annotation process. This includes the time they spent annotating each item and the exact timeline of their activity. In the second wave, 10% of each annotator's assigned set of articles was the same for later evaluation of the inter-annotator agreement. We selected these before the collection process and gave them our

---

[3]https://doccano.github.io/doccano/

Class	Number of articles
Trustworthy	3534
Partially trustworthy	2577
Misleading	1526
Manipulative	1859
Unclassifiable	733

Table 1
Distribution of article items per annotated credibility class

ground-truth annotations to ensure that the class distribution within this subset is balanced. We present the results of speed and inter-annotator agreement analyses later in this section.

4.3. Data Analysis

In this section, we present statistics of the newly-collected dataset. Overall, it contains 10229 articles spanning 60 Czech sources. A detailed list of all sources, their respective item counts, and class distributions can be found in Appendix B. Plotted in Figure 3 is the distribution of their times spent on individual articles. By average, annotators spent 2, 95 minutes (177 seconds) on a single article, which indicates reasonable time allocation.

We pay close attention to the per-source class distributions and ensure that the general tendencies in annotations match the Czech media space analyses studying the high-level credibility of news outlets. State-owned media (ČTK, ČT24, and iROZHLAS) and local newspapers (Jihlavské listy and Mostecké listy) have a majority of their stories classified as 'Trustworthy.' Articles from private media outlets (Seznam Zprávy, iDnes, Deník) are also most often classified as 'Trustworthy.' This time, however, other classes are more prominent. Openly left-wing (A2larm) or right-wing (Echo 24 and Forum24) sources have more items classified as misleading or manipulative in comparison to their counterparts without distinctive political tendencies. The 'Partially trustworthy' news stories occur the most by tabloid news sites (Blesk, Aha!, Extra.cz).

We can see the disinformative news sites (Aeronet, Protiproud, Skrytá pravda) on the other side of the spectrum, as their articles get exceedingly labeled as 'misleading' and 'manipulative.'

Overall, we can see that the high-level patterns in the annotations match the news sources' characteristics, as described in media science literature [26].

4.3.1. Inter-annotator Agreement

We gathered data for determining the inter-annotator agreement by assigning 4 duplicated articles (10 %) to each annotator in the second annotation wave. We computed Randolph's Kappa [30] and arrived on a value of 0.615, which corresponds to a moderate agreement [31]. This indicates a general understanding of the task and rather. At the same time, there is room for the latter filtering of problematic annotators, who can be spotted by largely deviating in categorizing these duplicated articles.
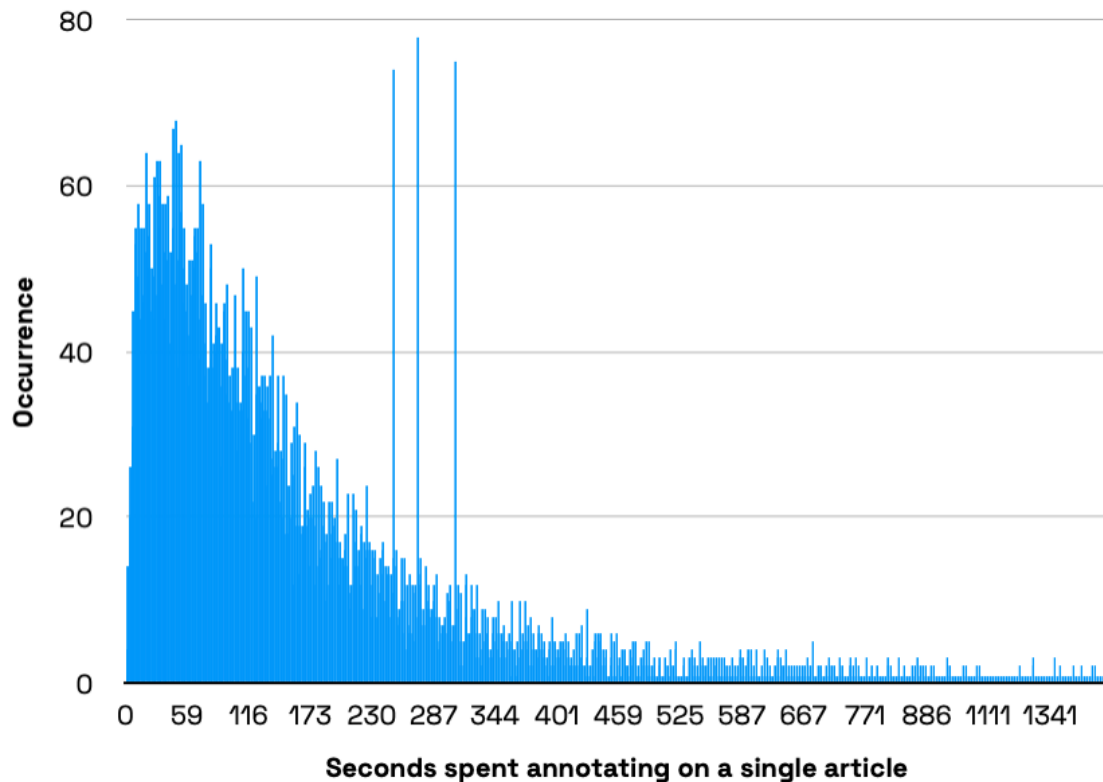
Figure 3: Distribution of the times spent on single articles by annotators. The x-axis denotes the number of seconds and the y-axis the count of respective occurrences.

5. Experimental Results

To establish initial benchmarks, we trained models of 3 distinct architectures for the task of news trustworthiness classification on our dataset. Herein, we present the experimental setup and the results.

Model	F-1 score				Overall average macro F-1 score
	Trustworthy	Partially credible	Misleading	Manipulative	
TF-IDF SVM	0.52	0.40	0.35	**0.68**	0.49
FastText LG	**0.58**	0.28	0.14	0.60	0.40
Czert (BERT)	0.55	**0.47**	**0.44**	0.61	**0.52**

Table 2
Verifee dataset benchmarks using three classification architectures: TF-IDF, FastText, and Czert (a BERT model for the Czech language). We report the F-1 score on the testing split.

5.1. Experimental setting

We experimented with three model architectures: Term frequency-inverse document frequency (TF-IDF)-based Support Vector Machines (SVM) classifier [32, 33], FastText classifier [34], and BERT [35]. We first describe the data preparation procedure and later review the configuration of each employed model.

5.2. Data Preparation

We follow the pre-defined configuration of train, test, and validation sets described in Section 4. To balance the dataset, we removed all duplicates used to measure kappa and randomly selected a sample of 1400 articles from each credibility class. We do not employ the 733 items categorized as 'Unclassifiable'. We insert the article's title and body concatenated with a period as the input to each evaluated model.

5.2.1. TF-IDF

We first trained a TF-IDF-based Support Vector Machines model. Therein, the text of an article is first vectorized using TF-IDF. Then, the SVM model classifies the content as one of the 4 credibility classes. We used the scikit-learn library [36] for its implementation.

We kept the model's vocabulary unfiltered by setting its `min_df` and `max_df` parameters to 1. For the SVM, we used Radial basis function kernel and Regularization parameter set to 1.

5.2.2. FastText

Another architecture we included in the benchmarking is a FastText classifier. In this pipeline, each article text is first tokenized using nltk. An article vector is then obtained by averaging FastText word embeddings of these tokens. We use one of FastText library's pre-trained models [37], which was trained on the Czech corpus from Common Crawl and Wikipedia. Lastly, a one-vs-rest logistic regression [38] classifies the article.

As for the hyperparameters, we used L2 penalty term combined with regulization set to 1.

5.2.3. BERT

Lastly, we also tried the Czert model [27], which utilizes the BERT architecture but is trained on the Czech national corpus [39], Czech Wikipedia, and a scrape of Czech news.

We fine-tuned the model for the purposes of our classification task using the Cross-entropy loss and the learning rate set at $3 * 10^{-5}$. We fine-tuned the model for 4 epochs and batch size set to 32.

5.3. Results

We present the per-class and overall F-1 score results in Table 2. As can be observed, the scores distinctively differ across classes. Upon closer inspection, both TF-IDF SVM and FastText LG models perform better on the classes at either pole of the trustworthiness spectrum (i.e., 'Trustworthy' and 'Manipulative'), but underperform at the middle ones, resulting in overall

testing F-1 scores of 0.49 and 0.40, respectively. We expect that the poor performance of the FastText LG model is caused due to the inability to construct granular representation that is necessary for our task, which cannot be resembled by averaged word embeddings. Despite sligthly losing on the pole classes, the Czert (BERT) model resembles best robustness across the spectrum. It achieves an overall testing F-1 score of 0.52

6. Ethical Discussion and Limitations

Due to the high-impact nature of the solved task, we consider it appropriate to review ethical considerations that have been made during this research project. Additionally, we outline further steps we are making to ensure safety and transparency even beyond publication and recommendations for build-up work.

First, let us focus on the presence of biases in the data. Despite avoiding this statistical phenomenon being practically impossible, we put extensive procedures in place even at the very start of the project. By inviting media researchers into our core team, we wanted to minimize misunderstandings and mistakes that scientists from the field of computational linguistics could easily make when assembling the methodology for the task of trustworthiness assessment due to their limited knowledge of the current literature and theory in the area of journalism. Prior to the data annotation, we invited scholars in media studies and journalists from the industry to a series of workshops, where we asked them to submit feedback and discuss the methodology. Based on the assembled comments, we kept updating it until a general consensus was reached. In terms of the annotation process itself, multiple safeguards have been employed to prevent the annotators' source or author preference.

Second, let us shift towards the ethics of using any technology built around this data in the wild. We want to stress that anyone using this dataset for the purposes of creating a trustworthiness classification system should provide transparent information to the users that this process is automatic and hence faulty to a certain extent. We must note that it is yet unclear how models trained on this data generalize for future articles (i.e., news about topics and events they have not encountered in the training set) and news sources that were not included in the training set. A study into these should be conducted prior to making this technology available unrestrictedly to the public.

Despite bearing these safety questions in mind is crucial, we believe that such systems can eventually be great assistive tools for people reading news stories online. The potential benefits of such technology should support initiatives into safeguarding it first and hence establishing public and academic trust.

7. Conclusion

This work presents a novel methodology for classifying news article trustworthiness and presents a dataset of nearly $10,000$ Czech news articles with respective annotations. Unlike previous methods that classify all texts from a given media outlet with the same class, we treat the articles on an individual level. The high inter-annotator agreement shows that our

methodology constitutes a good feature-based framework, leaving little to no room for personal annotators' inducement.

To the best of our knowledge, we are the first to include media and computer science researchers in the core team when developing a similar dataset. Additionally, all of our annotators were journalism students. As our methodology underwent extensive feedback loops with professionals in the industry, we hope to establish a new interdisciplinary standard for future related works to follow.

We provide benchmark results on our dataset using 3 different classifier architectures and obtain promising results. We open-source both the methodology and data and encourage researchers to undertake similar initiatives in new languages and social contexts, especially low-resourced ones. Since the framework derives all parameters based on the text contents, it is language-agnostic. Hence, minimal additional methodological work is necessary before new annotations.

In future work, we intend to study the generalization abilities of systems trained using this data and the application of task-specific architectures. Moreover, we wish to further explore the potential of multimodality that our dataset offers and analyze the attached images.

Acknowledgements

References

[1] A. Woodward, Fake news": A guide to trump's favourite phrase and the dangers it obscures, Independent. Retreived from https://www. independent. co. uk/news/world/americas/us-election/trump-fake-news-counter-history-b732873. html (2020).

[2] C. Wardle, The need for smarter definitions and practical, timely empirical research on information disorder, Digital Journalism 6 (2018) 951–963. URL: https://doi.org/10.1080/21670811.2018.1502047. doi:10.1080/21670811.2018.1502047.

[3] J. H. Fetzer, Disinformation: The use of false information, Minds and Machines 14 (2004) 231–240.

[4] B. Swire-Thompson, J. DeGutis, D. Lazer, Searching for the backfire effect: Measurement and design considerations, Journal of Applied Research in Memory and Cognition 9 (2020) 286–299.

[5] Y. Wang, M. McKee, A. Torbica, D. Stuckler, Systematic literature review on the spread of health-related misinformation on social media, Social science & medicine 240 (2019) 112552.

[6] P. Meel, D. K. Vishwakarma, Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities, Expert Systems with Applications 153 (2020) 112986.

[7] A. D'Ulizia, M. C. Caschera, F. Ferri, P. Grifoni, Fake news detection: a survey of evaluation datasets, PeerJ Computer Science 7 (2021) e518.

[8] W. Y. Wang, "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), 2017, pp. 422–426.

[9] J. Nørregaard, B. D. Horne, S. Adalı, Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles (2019).

[10] V. Slovikovskaya, G. Attardi, Transfer learning from transformers to fake news challenge stance detection (fnc-1) task, in: Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 1211–1218.

[11] Z. Guo, M. Schlichtkrull, A. Vlachos, A survey on automated fact-checking, Transactions of the Association for Computational Linguistics 10 (2022) 178–206.

[12] T. Mitra, E. Gilbert, Credbank: A large-scale social media corpus with associated credibility annotations, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 9, 2015, pp. 258–267.

[13] S. Volkova, K. Shaffer, J. Y. Jang, N. Hodas, Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter, in: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers), 2017, pp. 647–653.

[14] P. Nakov, G. Da San Martino, T. Elsayed, A. Barrón-Cedeno, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, N. Babulkov, et al., The clef-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news, in: European Conference on Information Retrieval, Springer, 2021, pp. 639–649.

[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[16] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, ELECTRA: Pre-training text encoders as discriminators rather than generators, in: ICLR, 2020. URL: https://openreview.net/pdf?id=r1xMH1BtvB.

[17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692.

[18] J. C. Reis, A. Correia, F. Murai, A. Veloso, F. Benevenuto, Supervised learning for fake news detection, IEEE Intelligent Systems 34 (2019) 76–81.

[19] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, S. Satoh, Spotfake: A multi-modal framework for fake news detection, in: 2019 IEEE fifth international conference on multimedia big data (BigMM), IEEE, 2019, pp. 39–47.

[20] J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, A. Mittal, The fact extraction and verification (fever) shared task, EMNLP 2018 80 (2018) 1.

[21] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp

tasks, Advances in Neural Information Processing Systems 33 (2020) 9459–9474.

[22] J. Maillard, V. Karpukhin, F. Petroni, W.-t. Yih, B. Oguz, V. Stoyanov, G. Ghosh, Multi-task retrieval for knowledge-intensive tasks, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 1098–1111.

[23] J. Luken, N. Jiang, M.-C. de Marneffe, Qed: A fact verification system for the fever shared task, in: Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), 2018, pp. 156–160.

[24] Y. Nie, H. Chen, M. Bansal, Combining fact extraction and verification with neural semantic matching networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 6859–6866.

[25] A. Damstra, H. G. Boomgaarden, E. Broda, E. Lindgren, J. Strömbäck, Y. Tsfati, R. Vliegenthart, What does fake look like? a review of the literature on intentional deception in the news and on social media, Journalism Studies 22 (2021) 1947–1963. URL: https://doi.org/10.1080/1461670X.2021.1979423. doi:10.1080/1461670X.2021.1979423. arXiv:https://doi.org/10.1080/1461670X.2021.1979423.

[26] V. Štětka, J. Mazák, L. Vochocová, "nobody tells us what to write about": The disinformation media ecosystem and its consumers in the czech republic, Javnost - The Public 28 (2021) 90–109. URL: https://doi.org/10.1080/13183222.2020.1841381. doi:10.1080/13183222.2020.1841381.

[27] J. Sido, O. Pražák, P. Přibáň, J. Pašek, M. Seják, M. Konopík, Czert–czech bert-like model for language representation, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), 2021, pp. 1326–1338.

[28] S. Peltzman, Political ideology over the life course, Public Choice: Analysis of Collective Decision-Making eJournal (2019).

[29] R. Scott, Does university make you more liberal? estimating the within-individual effects of higher education on political values, Electoral Studies 77 (2022) 102471. URL: https://www.sciencedirect.com/science/article/pii/S0261379422000312. doi:https://doi.org/10.1016/j.electstud.2022.102471.

[30] J. Randolph, Free-marginal multirater kappa (multirater free): An alternative to fleiss fixed-marginal multirater kappa, volume 4, 2010.

[31] M. L. McHugh, Interrater reliability: the kappa statistic, Biochemia medica 22 (2012) 276–282.

[32] C. Sammut, G. I. Webb (Eds.), TF–IDF, Springer US, Boston, MA, 2010, pp. 986–987. URL: https://doi.org/10.1007/978-0-387-30164-8_832. doi:10.1007/978-0-387-30164-8_832.

[33] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, B. Scholkopf, Support vector machines, IEEE Intelligent Systems and their applications 13 (1998) 18–28.

[34] A. Joulin, E. Grave, P. B. T. Mikolov, Bag of tricks for efficient text classification, EACL 2017 (2017) 427.

[35] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[37] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, Learning word vectors for 157 languages, in: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[38] C. Sammut, G. I. Webb (Eds.), Logistic Regression, Springer US, Boston, MA, 2010, pp. 631–631. URL: https://doi.org/10.1007/978-0-387-30164-8_493. doi:10.1007/978-0-387-30164-8_493.

[39] M. Křen, V. Cvrček, T. Čapka, A. Čermáková, M. Hnátková, L. Chlumská, T. Jelínek, D. Kováříková, V. Petkevič, P. Procházka, H. Skoumalová, M. Škrabal, P. Truneček, P. Vondřička, A. Zasina, SYN v4: large corpus of written czech, 2016. URL: http://hdl.handle.net/11234/1-1846, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

A. Annotation Methodology and Annotators Instructions

A.1. Annotation instructions

Each class is defined by the positive aspects it contains and the negative aspects it can and cannot contain. When annotating, we start with the most trustworthy class (credible). We then move down a class whenever an article does not meet the requirements of the current class, for example when it contains too many permissible negative aspects or contains a negative aspect that must not occur in that class.

A.2. Trustworthiness classes

A.2.1. Trustworthy

Positive aspects contained in the article (min. 5):

- Citation of relevant authorities on the topic, representing credible institutions
- Views of all interested parties
- Facts presented within the context
- Grammatical correctness, without overtly expressive language
- An identifiable author
- Undistorted data

Negative aspects contained in the article (max. 1):

- Missing citations
- Unrepresented views of opposing parties
- Facts presented without a context
- Grammatically incorrect or overtly expressive language
- Unidentifiable author
- Distorted data

Negative aspects that must not appear in the article:

- Clickbait
- Hate speech
- An attack on an opinion opponent without justification
- Manipulating the reader
- Conspiracy theories
- Appeal to emotion
- Logical fallacies
- Tabloid language

A.2.2. Partially Trustworthy

Positive aspects contained in the article:

- Grammatical correctness, without overtly expressive language
- Undistorted data

Negative aspects contained in the article (2-5):

- Missing citations
- Unrepresented views of opposing parties
- Facts presented without a context
- Grammatically incorrect or overtly expressive language
- Unidentifiable author
- Distorted data
- Clickbait
- Appeal to emotion
- Tabloid language

Negative aspects that must not appear in the article:

- Hate speech
- An attack on an opinion opponent without justification
- Manipulating the reader
- Conspiracy theories
- Logical fallacies

A.2.3. Misleading

Positive aspects contained in the article:

None need to be present

Negative aspects contained in the article (6-7):

- Missing citations
- Unrepresented views of opposing parties
- Facts presented without a context
- Grammatically incorrect or overtly expressive language
- Unidentifiable author
- Distorted data
- Clickbait
- Appeal to emotion
- Tabloid language
- Logical fallacies
- An attack on an opinion opponent without justification

Negative aspects that must not appear in the article:

- Hate speech
- Manipulating the reader
- Conspiracy theories

A.2.4. Manipulative

Positive aspects contained in the article:

None need to be present

Negative aspects contained in the article:

It either contains 8+ negative aspects:
- Missing citations
- Unrepresented views of opposing parties
- Facts presented without a context
- Grammatically incorrect or overtly expressive language
- Unidentifiable author
- Distorted data
- Clickbait
- Appeal to emotion
- Tabloid language
- Logical fallacies
- An attack on an opinion opponent without justification
  Or it contains any of these 3 aspects:
- Hate speech
- Manipulating the reader
- Conspiracy theories

Negative aspects that must not appear in the article:

All negative aspects can be present

A.3. Resolving unclassifiable articles and errors

A.3.1. Unclassifiable articles

Articles that, due to their length or structure, cannot be classified according to this methodology (or do not have sufficient content to allow the aspects mentioned to be analysed) must be labeled as unclassifiable. This may include one-sentence flash news announcements, paywall texts and others. This allows them to be filtered out and not corrupt the rest of the annotated data.

A.3.2. Errors

In the case that an error with the platform or an uncertainty with an article is encounter, we fully encourage annotators to report those issues through comment functionality on the Doccano platform. Our team will do their best to resolve any problem and clarify any ambiguity.

B. Detailed news source statistics

Continued on the next page.

News source	Article items per class				
	Trustworthy	Part. trustworthy	Misleading	Manipulative	Unclassifiable
A2larm	158	106	49	13	20
AC24	22	40	45	26	22
Aeronet	6	19	69	347	9
Aha!	20	73	41	7	16
Aktuálně	234	107	33	5	38
Blesk	40	132	35	5	6
Brněnský deník	27	10	2	0	4
CNN Prima News	232	86	13	2	16
CZ24 News	16	28	12	21	2
Czech free press	3	15	12	10	2
Deník	58	14	4	0	2
Deník N	28	8	4	1	9
Deník Referendum	192	51	22	6	5
E-republika	4	7	13	16	2
Echo 24	203	84	21	2	6
Euro	14	8	1	0	0
Euro Zprávy	55	18	4	0	5
Extra.cz	86	229	116	37	25
Forum24	147	64	39	21	13
Globe 24	15	7	2	0	0
Haló noviny	14	14	16	4	3
Hospodářské noviny	34	14	4	6	78
INFO.cz	18	19	4	2	14
Jihlavské listy	26	3	1	0	3
Lidovky.cz	5	19	4	3	30
MediaGuru	24	18	3	0	1
Metro	135	53	5	0	9
Mostecké listy	22	3	1	0	1
NWOO	8	35	37	63	15
Novinky.cz	67	67	13	3	15
Nová republika	5	34	51	56	8
Outsider Media	94	119	162	235	92
Parlamentní Listy	279	269	138	91	33
Peak.cz	122	49	12	1	6
Proti Proud	15	41	102	301	23
Raptor TV	1	4	2	3	1
Reflex	1	1	3	1	11
Rukojmí	21	52	117	290	12
Seznam Zprávy	181	51	13	1	8
Skrytá Pravda	6	17	64	169	11
Sputnik ČR	219	296	124	43	34
Stars 24	27	42	15	3	2
Svobodné noviny	13	23	45	77	8
TN.cz	213	204	38	3	17
Týden	54	14	4	0	4
VOX Populi	4	14	60	72	22
Zvědavec	6	6	6	7	5
iDnes.cz	93	39	11	1	19
iROZHLAS	254	69	13	1	18
ČT24	246	36	5	3	35
ČTK	38	2	2	0	1
Časopis Šifra	11	10	10	8	3
Česko Aktuálně	36	42	34	35	8

Table 3

Distribution of classes per individual news sources