Czech-ing the News: Article Trustworthiness Dataset for Czech

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

matyas.bohacek@matsworld.io, michal@bravansky.com,
me@trhlikfilip.com, vaclav.moravec@fsv.cuni.cz

Abstract

We present the Verifee dataset: a multimodal dataset of news articles with fine-grained trustworthiness annotations. We bring a diverse set of researchers from social, media, and computer sciences aboard to study this interdisciplinary problem holistically and develop a detailed methodology that assesses the texts through the lens of editorial transparency, journalist conventions, and objective reporting while penalizing manipulative techniques. We collect over $10,000$ annotated articles from 60 Czech online news sources. Each item is categorized into one of the 4 proposed classes on the credibility spectrum – ranging from entirely trustworthy articles to deceptive ones – and annotated of its manipulative attributes. We fine-tune prominent sequence-to-sequence language models for the trustworthiness classification task on our dataset and report the best F-1 score of 0.53. We open-source the dataset, annotation methodology, and annotators' instructions in full length at https://www.verifee.ai/research/ to enable easy build-up work.

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 (Woodward, 2020). 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 another medium, such as a Tweet) is considered fake (disinformative) or truthful (credible). There is no consensus in the literature on what specific parameters constitute this distinction, but truthfulness is usually considered the primary one. Some approaches recognize more fine-grained scales with specific classes (e.g., tabloid news, mixed reliability news),

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 96–109

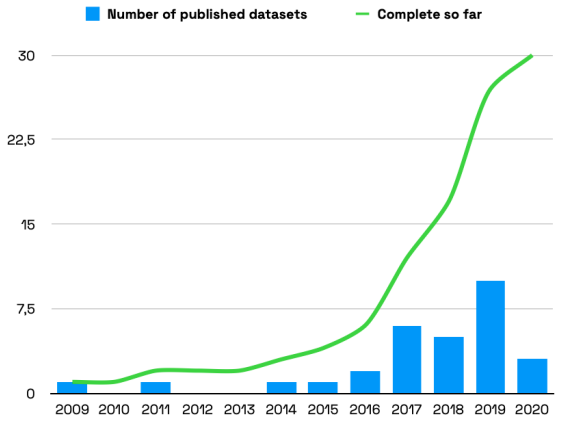


Figure 1: Continual statistics on disinformation classification datasets publishing throughout the years 2009-2020. The bar charts denote the number of new datasets (y-axis) published in the respective year (x-axis), while the overlay line captures the cumulative number of datasets published until that year.

whereas others are binary (including fake and credible classes). Either way, the sole text is considered.

Meanwhile, the task of automatic fact-checking 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 variant 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

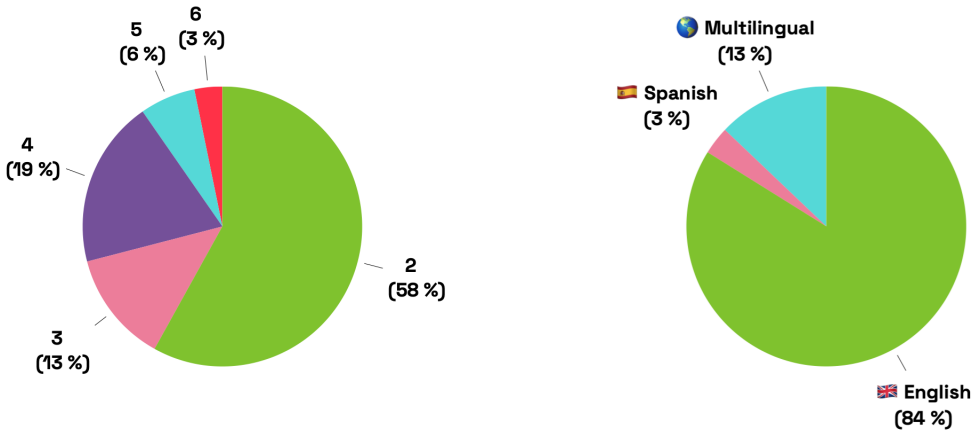
With the advent and development of digital network media at the beginning of the 21st century, there has been a dynamic spread of unverified, inaccurate, or false information (ranging from textual to audiovisual), which is referred to as information disorders. Information disorders as part of information 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 (Wardle, 2018), it is misleading to label information disorders with the umbrella term "fake news." 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 created to disseminate false information intentionally 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 (Fetzer, 2004) considers disinformation "misinformation with an attitude," while attitude is the aforementioned deliberate deception of recipients. According to another approach (Swire-Thompson et al., 2020; Wang et al., 2019), 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 (Meel and Vishwakarma, 2020). 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. (2021) have conducted a thorough study on fake news detection datasets. We high-



(a) Number of classes recognized by the dataset

(b) Language of the dataset

Figure 2: Proportional statistics of the existing datasets for disinformation classification.

light three of these based on the traction within the research community and direct the reader to this review for more detail. Wang (2017) created the LIAR dataset with 12,836 text excerpts of 6 classes. Later, Nørregaard et al. (2019) published NELA-GT dataset containing 713,000 news articles belonging to 2 classes. Lastly, Slovikovskaya and Attardi (2020) presented the FNC-1 dataset with 49,972 news articles classified into 4 labels. All these datasets are in English.

Guo et al. (2022) 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 (2015) created the CredBank dataset with over 1,000 English Tweets classified into 5 labels. Multiple works followed, including the much larger Suspicious dataset (Volkova et al., 2017) containing over 130,000 English Tweets with 2 assigned classes. Nakov et al. (2021) presented the CheckThat21-T1A dataset with over 17,000 Tweets of 2 classes. These Tweets come from multiple languages. Recently, domain-specific datasets have also emerged, prominently centered around COVID-19 facts (Saakyan et al., 2021).

Shown in Figure 1 is a visualization of datasets publishing (D'Ulizia et al., 2021) over time, showing that the popularity of this task in the AI and NLP community is a recent phenomenon, corresponding to the general focus on disinformation in the public discourse. However, the sizeable collective excitement goes hand-in-hand with the inconsistency of the problem's framing and methodologies. This can be demonstrated with Figure 2a, which captures the distribution of these datasets by the pure number of labels they recognize. Furthermore, we see significant inconsistencies in the methodologies leading to these classifications. Some works (Nørregaard et al., 2019) derive the class based on the high-level credibility assessment of its source (i.e., they assess a source and treat all its articles in the same manner, leaving no room for exceptions). Others (Wang, 2017; Slovikovskaya and Attardi, 2020) 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 their 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. 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.

2.3 Automated fake news detection

The task of automated fake news detection has usually been approached by fine-tuning general-purpose language models, such as BERT (Devlin et al., 2019), ELECTRA (Clark et al., 2020), or

RoBERTa (Liu et al., 2019). Specific architectures for this task have been studied in the literature, too. (Reis et al., 2019), 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. (Singhal et al., 2019) create the first multi-modal architecture for this task as they combine the texts at the input with images included in the article. Some recent works also formulate the task as graph classification (Vaibhav et al., 2019).

2.4 Automated fact-checking

Architectures for automated fact-checking usually consist of an evidence retrieval module and a verification module (Thorne et al., 2018). Recent dense retrievers with learned representations and fast dot-product indexing (Lewis et al., 2020; Maillard et al., 2021) 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 (Luken et al., 2018; Nie et al., 2019). 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 (Guo et al., 2022) for a concise overview of such methods. Recent works also use sequence-to-sequence models to generate natural logic-based inferences later used as proofs (Krishna et al., 2022).

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 establish a clear division between the fake news detection and fact-checking tasks, our methodology focuses solely on the content aspects of the article. We 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 (Damstra et al., 2021).

In our framing of the problem, trustworthiness is assessed by the presence of transparent and deceptive attributes. Focusing solely on trustworthiness allows us to streamline the annotation process since there is no requirement for outside context, and the given class is thus final (i.e., unlike with fact-checking methods employing truthfulness, no later information can reverse the assessment).

3.2 Classes

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**: Such an article is entirely credible, cites its sources, and presents the opinions of all involved parties. In our framework, it does not contain any negative attributes while having at least five positive ones.

2. **Partially Trustworthy**: While not deceiving its readers, such an article attempts to exaggerate the topic while neglecting to uphold journalistic norms. In our framework, it includes 2 to 5 negative attributes.

3. **Misleading**: Such an article contains deception but does not yet fall under conspiracies. In our framework, it includes 6 to 8 negative attributes.

4. **Manipulative**: Such an article strives to manipulate its reader by employing conspiratory narratives. In our framework, it contains over 8 negative attributes or one of 3 highly problematic ones (e.g., conspiracies, hate speech).

4 Dataset

We collected a dataset of 10,197 Czech news articles. Each entry in the dataset consists of the article's text, HTML source, title, description, authors, source name, URL, main image, and the

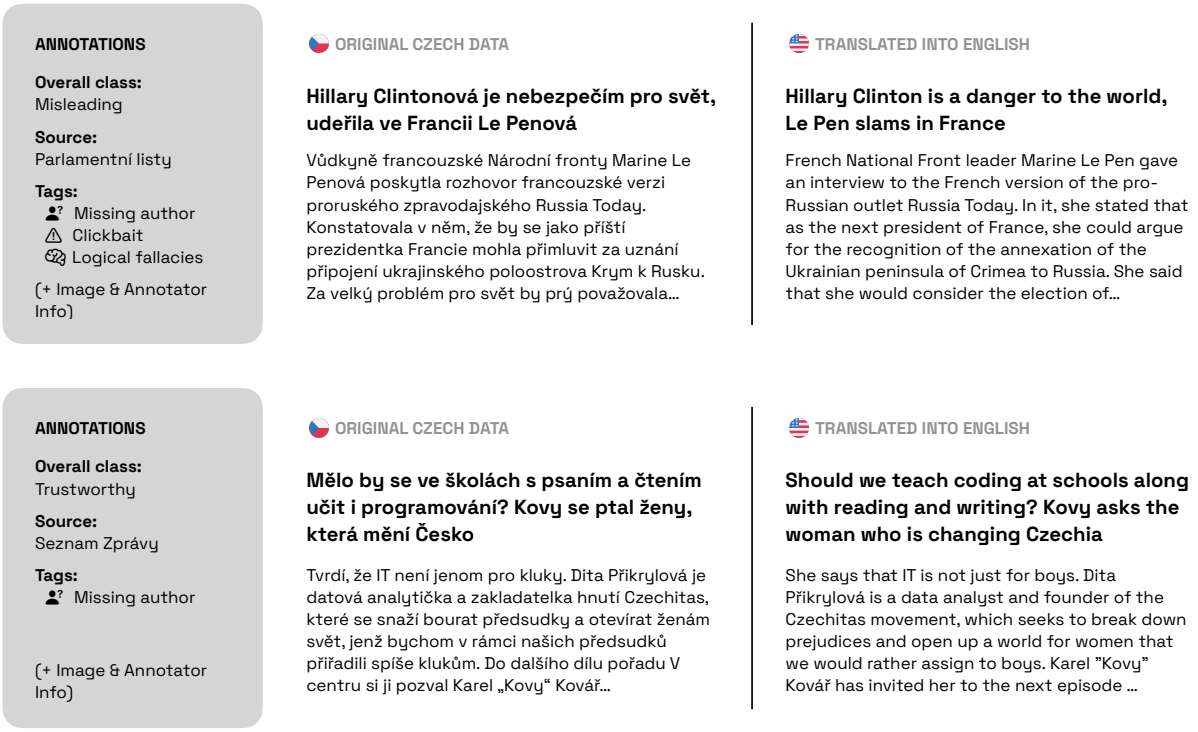


Figure 3: Representative examples of two article items in the dataset including their annotations and metadata. The original texts are translated into English for demonstrative purposes.

annotated class. A subset of the dataset also contains the linguistic attribute annotations, which led up to the classification. Representative examples of two article items are shown in Figure 3.

We open-source the dataset at https://www.verifee.ai/research/ under a custom license[1]. We provide pre-defined train (80 %), validation (10 %), and testing splits (10 %) that have been assigned randomly.

4.1 Scraping and Pre-processing

Initially, we assembled nearly 94,000 articles by scraping URLs of 60 Czech news sources[2] obtained from Common Crawl[3]. These sources included mainstream journalistic websites, tabloids, independent news outlets, and websites that are part of the disinformation ecosystem (Štětka et al., 2021), capturing the full scope of journalistic content in the Czech Republic. Their complete list can be found in Appendix C.

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. Similarly, we detected mentions of any controversial topics relevant to the Czech media context. Additionally, we ascertained whether the article disposes of a real author via an out-of-the-box Named Entity Recognition model (Sido et al., 2021) for the Czech language.

4.1.2 Filtering

We applied multiple filters and balancing mechanisms based on text length and topics to mitigate deficiencies caused by inherent flaws in Common Crawl, which reduced the dataset's size from 94,000 to 10,197 items. This way, we also ensured that the data is as representative of the Czech news ecosystem and as diverse as possible. The detailed parameters used for filtering are described in Appendix B.

4.2 Annotations Organization

We conducted two rounds of annotation. The first round involved 7,528 articles, where just the class was assigned to each article. The second round included 2,669 articles. This time, annotators were

[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] The complete list of sources can be found in Appendix C.

[3] https://commoncrawl.org

Class	Number of articles
Trustworthy	3520
Partially trustworthy	2574
Misleading	1524
Manipulative	1848
Unclassifiable	731

Table 1: Distribution of article items per annotated trustworthiness class.

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 per round.

4.2.1 Annotators

All the raters were students of journalism who were native speakers of the Czech language. They thus had a more advanced understanding of news credibility than the general population. Due to their age (Peltzman, 2019) and education (Scott, 2022), 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 them with detailed materials, and encouraged them to come forward with any problems.

To further mitigate the influence of annotators' media and author preferences on the assessment, 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.

4.2.2 Platform

We used a tailored version of the open-source Doccano[4] tool. 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.

The platform enabled us to track the annotators' activity, including the time spent on each article. In the second wave of annotation, 10 % of articles were shared among all annotators to evaluate the inter-annotator agreement. These were preselected and equipped with our ground-truth annotations.

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

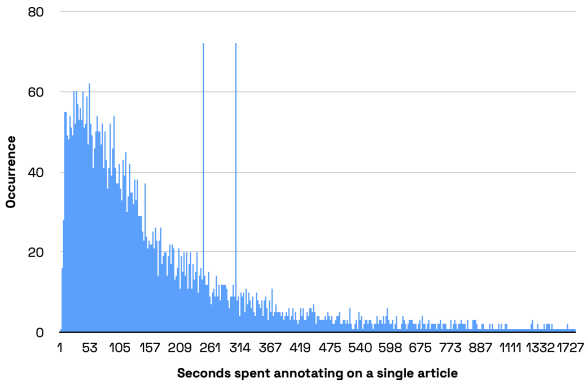


Figure 4: Distribution of single article annotation time. The x-axis denotes the number of seconds, and the y-axis the count of respective occurrences.

4.3 Data Analysis

By average, annotators spent 2.97 minutes (179 seconds) on a single article, which indicates reasonable time allocation. The distribution of the per-item time allocation is shown in Figure 4.

The overall class distribution, reported in Table 1, shows a skew towards the trustworthy end of the spectrum. 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 in 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 (Štětka et al., 2021).

Model	Trustworthy	Partially trustworthy	Misleading	Manipulative	Overall
RoBERTa	0.63	0.44	**0.45**	0.59	**0.53**
BERT	0.55	**0.47**	0.44	0.61	0.52
Electra	**0.66**	0.43	0.39	0.58	0.51
TF-IDF	0.52	0.40	0.35	**0.68**	0.49
FastText	0.58	0.28	0.14	0.60	0.40

Table 2: Dataset benchmarks fine-tuned for the task of trustworthiness classification. We report the testing split F-1 score for each class, as well as an overall average.

4.3.1 Inter-annotator Agreement

The Randolph's Kappa (Randolph, 2010) of the second annotation wave amounts to 0.615, corresponding to a moderate agreement (McHugh, 2012). Atop this innate solid baseline, we filtered problematic annotators, who were identified by largely deviating in categorizing these duplicated articles.

5 Experimental Results

We conduct baseline results by fine-tuning general-purpose architectures.

5.1 Data Preparation

We follow the pre-defined configuration of train, test, and validation splits described in Section 4. To balance the training, we selected a random sample of 1400 articles from each credibility class. We insert the article's title and body (concatenated with a period) as the input to the model.

5.2 Model architectures

We experimented with 5 model architectures: RoBERTa-based RobeCzech (Liu et al., 2019; Straka et al., 2021), BERT-based Czert (Devlin et al., 2019; Sido et al., 2021), Electra-based Small-e-Czech (Clark et al., 2020; Kocián et al., 2022), Term frequency-inverse document frequency (TF-IDF)-based Support Vector Machines (SVM) classifier (Sammut and Webb, 2010b; Hearst et al., 1998), and FastText Embedding-based Logistic Regression classifier (Joulin et al., 2017). Respective training details follow.

5.2.1 RobeCzech (RoBERTa)

RobeCzech is a RoBERTa model trained on the Czech national corpus (Křen et al., 2016). Setting the learning rate to $3 * 10^-5$ and the batch size to 16, we fine-tuned this model using cross-entropy loss over a span of 4 epochs.

5.2.2 Czert (BERT)

Czert is a BERT model trained on the Czech national corpus. Setting the learning rate to $3 * 10^-5$ and the batch size to 32, we fine-tuned this model using cross-entropy loss over a span of 4 epochs.

5.2.3 Small-e-Czech (Electra)

Small-e-Czech (Kocián et al., 2022) is an ELECTRA-small trained on an internal Czech web corpus of Seznam.cz. Setting the learning rate to $3 * 10^-4$ and the batch size to 64, we fine-tuned this model using cross-entropy loss over a span of 3 epochs.

5.2.4 TF-IDF SVM

Our TF-IDF SVM model pipeline vectorizes the text using TF-IDF and later classifies these representation using SVM. We used the scikit-learn library (Pedregosa et al., 2011) and 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.

5.2.5 FastText

Our FastText pipeline vectorizes the tokenized words found by nltk (Bird and Loper, 2004) using FastText (Grave et al., 2018). By averaging these, a single article vector is obtained for each article, and a one-vs-rest logistic regression (Sammut and Webb, 2010a) is used to yield the predicted class. We used L2 penalty term combined with regularization set to 1.

5.3 Results

We present per-class F-1 score results, as well as their average, in Table 2. As can be observed, the scores differ distinctly across classes. Upon closer inspection, both TF-IDF SVM and FastText models perform better on the at the pole classes of the trustworthiness spectrum (i.e., 'Trustworthy' and 'Manipulative'), but under-perform at the middle ones, resulting in overall testing F-1 scores of 0.49 and 0.40, respectively. We hypothesize that the poor performance of the FastText model is caused

due to the inability to capture apt representation of causality and argumentative approaches in the averaged semantic vectors.

The comparative results of the Transformer-based models follow the same order as in other Czech evaluation tasks. We see the RobeCzech (ROBERTa) model come on top with F-1 score of 0.53, followed by Czert (BERT) with F-1 score of 0.52 and lastly Small-e-Czech (ELECTRA) with F-1 score of 0.51.

6 Ethical Discussion and Limitations

Due to the high-impact nature of the solved task, we review the ethical considerations made during this research project. Additionally, we outline further steps to ensure safety and transparency beyond publication, as well as recommendations for build-up work.

First, let us focus on the presence of biases in the data. 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 annotators' bias towards specific sources or authors (that may affect the classification).

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 still needs to be determined 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 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, such systems can eventually be great assistive tools for people reading news stories online. The potential benefits of such technology should support initiatives to safeguard it first and establish public and academic trust.

7 Conclusion

This work presents a novel methodology for classifying news article trustworthiness and presents a multimodal dataset of $10,197$ 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 5 different classifier architectures and obtain promising results – the best-performing RoBERTa model achieves a testing F-1 score of 0.53. We open-source the complete dataset 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

This paper was supported by the Technology Agency of the Czech Republic under grant No. TL05000057 "The Signal and the Noise in the Era of Journalism 5.0 - A Comparative Perspective of Journalistic Genres of Automated Content".

References

SG Bird and Edward Loper. 2004. Nltk: the natural language toolkit. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Alyt Damstra, Hajo G. Boomgaarden, Elena Broda, Elina Lindgren, Jesper Strömbäck, Yariv Tsfati, and Rens Vliegenthart. 2021. What does fake look like? a review of the literature on intentional deception in the news and on social media. *Journalism Studies*, 22(14):1947–1963.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.

Arianna D'Ulizia, Maria Chiara Caschera, Fernando Ferri, and Patrizia Grifoni. 2021. Fake news detection: a survey of evaluation datasets. *PeerJ Computer Science*, 7:e518.

James H Fetzer. 2004. Disinformation: The use of false information. *Minds and Machines*, 14(2):231–240.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28.

Armand Joulin, Edouard Grave, and Piotr Bojanowski Tomas Mikolov. 2017. Bag of tricks for efficient text classification. *EACL 2017*, page 427.

Matěj Kocián, Jakub Náplava, Daniel Štancl, and Vladimír Kadlec. 2022. Siamese bert-based model for web search relevance ranking evaluated on a new czech dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12369–12377.

Michal Křen, Václav Cvrček, Tomáš Čapka, Anna Čermáková, Milena Hnátková, Lucie Chlumská, Tomáš Jelínek, Dominika Kováříková, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Michal Škrabal, Petr Truneček, Pavel Vondřička, and Adrian Zasina. 2016. SYN v4: large corpus of written czech. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Amrith Krishna, Sebastian Riedel, and Andreas Vlachos. 2022. ProoFVer: Natural logic theorem proving for fact verification. *Transactions of the Association for Computational Linguistics*, 10:1013–1030.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Jackson Luken, Nanjiang Jiang, and Marie-Catherine de Marneffe. 2018. Qed: A fact verification system for the fever shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 156–160.

Jean Maillard, Vladimir Karpukhin, Fabio Petroni, Wen-tau Yih, Barlas Oguz, Veselin Stoyanov, and Gargi Ghosh. 2021. 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)*, pages 1098–1111.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Priyanka Meel and Dinesh Kumar Vishwakarma. 2020. 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*, 153:112986.

Tanushree Mitra and Eric Gilbert. 2015. 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, pages 258–267.

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeno, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Nikolay Babulkov, et al. 2021. The clef-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *European Conference on Information Retrieval*, pages 639–649. Springer.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6859–6866.

Jeppe Nørregaard, Benjamin D Horne, and Sibel Adalı. 2019. Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles.

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, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Sam Peltzman. 2019. Political ideology over the life course. *Public Choice: Analysis of Collective Decision-Making eJournal*.

Justus Randolph. 2010. Free-marginal multirater kappa (multirater κfree): An alternative to fleiss fixed-marginal multirater kappa. volume 4.

Julio CS Reis, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benevenuto. 2019. Supervised learning for fake news detection. *IEEE Intelligent Systems*, 34(2):76–81.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. 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)*, pages 2116–2129, Online. Association for Computational Linguistics.

Claude Sammut and Geoffrey I. Webb, editors. 2010a. *Logistic Regression*, pages 631–631. Springer US, Boston, MA.

Claude Sammut and Geoffrey I. Webb, editors. 2010b. *TF–IDF*, pages 986–987. Springer US, Boston, MA.

Ralph Scott. 2022. Does university make you more liberal? estimating the within-individual effects of higher education on political values. *Electoral Studies*, 77:102471.

Jakub Sido, Ondřej Pražák, Pavel Priban, Jan Pašek, Michal Seják, and Miloslav Konopík. 2021. Czert – czech bert-like model for language representation.

Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*, pages 39–47. IEEE.

Valeriya Slovikovskaya and Giuseppe Attardi. 2020. Transfer learning from transformers to fake news challenge stance detection (fnc-1) task. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1211–1218.

Milan Straka, Jakub Náplava, Jana Straková, and David Samuel. 2021. Robeczech: Czech roberta, a monolingual contextualized language representation model. In *Text, Speech, and Dialogue: 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6–9, 2021, Proceedings 24*, pages 197–209. Springer.

Briony Swire-Thompson, Joseph DeGutis, and David Lazer. 2020. Searching for the backfire effect: Measurement and design considerations. *Journal of Applied Research in Memory and Cognition*, 9(3):286–299.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The fact extraction and verification (fever) shared task. *EMNLP 2018*, 80(29,775):1.

Vaibhav Vaibhav, Raghuram Mandyam, and Eduard Hovy. 2019. Do sentence interactions matter? leveraging sentence level representations for fake news classification. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 134–139, Hong Kong. Association for Computational Linguistics.

Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. 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)*, pages 647–653.

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)*, pages 422–426.

Yuxi Wang, Martin McKee, Aleksandra Torbica, and David Stuckler. 2019. Systematic literature review on the spread of health-related misinformation on social media. *Social science & medicine*, 240:112552.

Claire Wardle. 2018. The need for smarter definitions and practical, timely empirical research on information disorder. *Digital Journalism*, 6(8):951–963.

A Woodward. 2020. 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*.

Václav Štětka, Jaromír Mazák, and Lenka Vochocová. 2021. "nobody tells us what to write about": The disinformation media ecosystem and its consumers in the czech republic. *Javnost - The Public*, 28(1):90–109.

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 Filtering Procedure

Following filters and balancing mechanisms were employed to select the dataset articles out of the greater pool of scraped articles.

- **Length of the text**: Only articles with a length of 400 to $10,000$ characters were included.
- **Category**: We filtered out opinion pieces for mainstream media. 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 so that all sources are as balanced as possible, no matter their actual distribution in the media ecosystem.
- **Topic**: Articles concerning hobbies and sports each form only 5% of the dataset, as they typically are not connected to disinformation. The remaining topics (general, economic, and tabloid) each form 30 % of the dataset.
- **Controversial topics**: We balanced the coverage of controversial topics by including the same number of such articles from mainstream and alternative or extremely opinionated news sources.

C Detailed news source statistics

Continued on the next page.

News source	Article items per class				
	Trustworthy	**Part. trust.**	**Misleading**	**Manipulative**	**Unclassifiable**
A2larm	145	10	43	100	20
AC24	22	24	42	39	18
Aeronet	6	321	65	19	9
Aha!	20	7	41	73	16
Aktuálně	194	4	29	94	34
Bez politické korektnosti	1	0	2	1	0
Blesk	38	5	33	132	6
Brněnský deník	27	0	2	10	4
CNN Prima News	187	2	13	71	11
CZ24 News	16	21	12	28	2
Czech free press	3	10	12	15	2
Deník	58	0	4	14	2
Deník N	28	1	4	8	9
Deník Referendum	173	6	18	45	5
E-republika	4	16	13	7	2
E15	3	0	0	1	1
Echo 24	188	1	19	75	6
Euro	12	0	1	8	0
Euro Zprávy	52	0	3	18	5
Extra.cz	71	29	100	205	25
Forum24	140	21	35	60	12
Globe 24	15	0	2	7	0
Haló noviny	14	4	16	14	3
Hospodářské noviny	34	4	4	14	77
INFO.cz	16	2	3	18	13
Jihlavské listy	26	0	1	3	3
Lidovky.cz	5	3	4	19	30
MediaGuru	21	0	3	14	1
Metro	125	0	5	48	9
Mostecké listy	22	0	1	3	1
NWOO	8	63	37	35	15
Neviditelný pes	2	3	1	0	2
Novinky.cz	65	1	12	64	15
Nová republika	5	54	51	31	7
Outsider Media	94	234	162	118	91
Parlamentní Listy	255	80	119	222	32
Peak.cz	103	1	12	48	6
Proti Proud	13	273	97	39	22
Raptor TV	1	3	2	4	1
Reflex	1	1	3	1	11
Respekt	2	0	1	3	0
Rukojmí	18	242	101	43	12
Seznam Zprávy	164	1	12	45	8
Skrytá Pravda	6	162	61	17	10
Sputnik Česká republika	199	39	89	264	32
Stars 24	27	3	15	38	2
Svobodné noviny	13	68	44	22	8
Svobodný svět	0	1	3	0	0
TN.cz	201	3	37	200	17
Týden	54	0	4	14	4
Týdeník občanské právo	0	1	0	0	0
VOX Populi	3	69	54	13	21
Zvědavec	6	7	6	6	5
iDnes.cz	90	1	10	39	19
iROZHLAS	219	1	11	57	18
ČT24	226	3	4	36	35
ČTK	32	0	2	0	1
Časopis Šifra	11	8	10	10	3
Česko Aktuálně	36	35	34	42	8

Table C.1: Class distribution of all unique news sources found in the dataset.