Misinformation Detection in the Wild: News Source Classification as a Proxy for Non-article Texts

Matyáš Boháček

Gymnasium of Johannes Kepler,
Prague, Czech Republic

matyas.bohacek@matsworld.io

Abstract

Creating classifiers of disinformation is time-consuming, expensive, and requires vast effort from experts spanning different fields. Even when these efforts succeed, their roll-out to publicly available applications stagnates. While these models struggle to find their consumer-accessible use, disinformation behavior online evolves at a pressing speed. The hoaxes get shared in various abbreviations on social networks, often in user-restricted areas, making external monitoring and intervention virtually impossible. To re-purpose existing NLP methods for the new paradigm of sharing misinformation, we propose leveraging information about given texts' originating news sources to proxy the respective text's trustworthiness. We first present a methodology for determining the sources' overall credibility. We demonstrate our pipeline construction in a specific language and introduce CNSC: a novel dataset for Czech articles' news source and source credibility classification. We constitute initial benchmarks on multiple architectures. Lastly, we create in-the-wild wrapper applications of the trained models: a chatbot, a browser extension, and a standalone web application.

1 Introduction

With the never-ending growth of the internet user base and its impact on our day-to-day lives, a significant portion of our work and leisure nowadays happens online. For many internet users, a substantial part of their time online, if not most, takes place on social media platforms (Paliszkiewicz et al., 2017; Riehm et al., 2019). Herein, most are constantly exposed to the information overload phenomenon. This means that the users are met with an unprecedented mass of posts, articles, images, and comments, which makes orienting within this space strenuous. Constantly verifying truthfulness of each presented information becomes virtually incompatible with the quick scrolling through timelines of new posts.

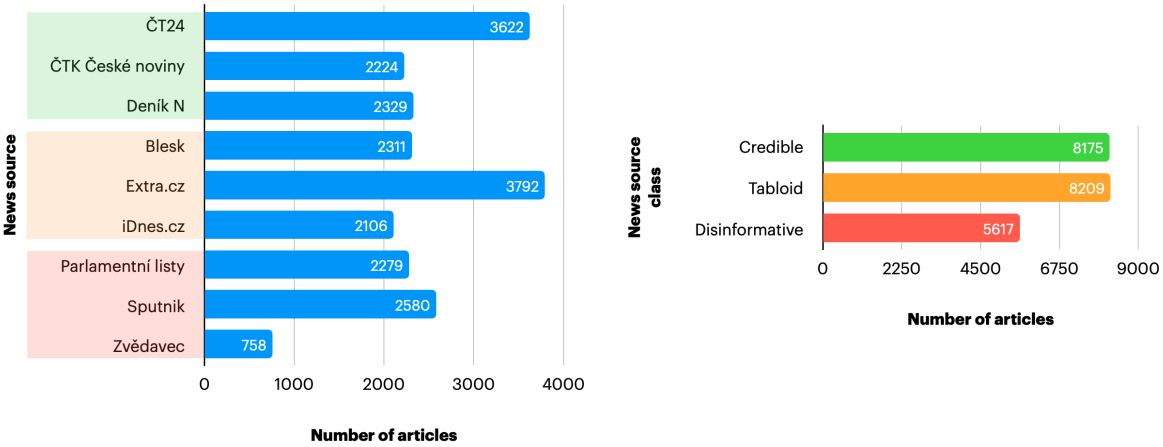
At the same time, the assessment of online media's trustworthiness is becoming more critical than ever. We could already see disinformation (i.e., deliberately constructed false information with the intention of someone's manipulation) being employed during critical social events, such as but not limited to elections, refugee crises, and controversial trials.

As a result, we observed an immense interest in methods that could automatically assess various aspects of credibility online in the literature throughout recent years. These include stance detection, automated fact-checking, or specific disinformation detection (also referred to as fake news detection). Tasks analyzing constituent attributes of disinformation, such as hate speech, stereotypization, or logical fallacy detection, have also been studied. Research in this field is not only restrained to text analysis: studies of visual disinformation detection, namely deepfake classification, are also often visited.

Despite being around for years, the rollout of such methods to real-world applications stagnates (Nicas, 2020; Achimescu and Chachev, 2020). Moreover, as disinformation gets scrutinized by extensive public interest and threatened by improved education about the problem, it evolves rapidly, making itself harder to spot and monitor. Many hoax campaigns have moved to access-restricted parts of the internet, such as closed Facebook groups, Telegram channels, and e-mail, making external monitoring and tailored debunking campaigns virtually impossible. Consequently, designing new tasks for machine learning models to combat these phenomena is challenging, as the relevant input and desired predictions become less definite. In the Czech Republic, for instance, many hoax websites started sharing augmented versions of their articles on these exact channels.

We thus set out to investigate whether existing tools and data resources from the domain of Nat-

Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI), pages 79 - 88



(a) per individual news sources

(b) per news source labels

Figure 1: Statistics of the article counts in the CSNS dataset. In Subfigure (a), the individual news sources are highlighted with the colors of their respective trustworthiness label.

ural language processing could be modified into generic tools, which would help assess the trustworthiness of various texts online. We chose the field of Czech media space as our proof-of-concept field. The problem framing seemed ambiguous at first – we wanted to analyze whether a given news text observed on social networks, blogs, or general websites seems trustworthy or not. We know that many of the subjected texts will likely be abbreviations of standard news articles, but the model should be resilient to arbitrary texts, too.

As we show in the review of related work, we later realized that many existing fake news classifiers' methodologies use only a single news source per class as their training reference. Such models are, hence, trained to classify originating news sources. While this may seem like a design flaw or a result of minimizing annotation complexity, we use it to our advantage. By training classifiers of a given article's originating source, we can utilize the trustworthiness associated with that medium's brand as a proxy for the reliability of the analyzed text. This way, eventual users interacting with the models' predictions will be able to use their existing experience and quickly recall their trust for the respective medium.

Additionally, we hypothesize that when the users are exposed to familiar labeling (in the form of likely originating news sources), getting accustomed to the application's framework and terminology may become more effortless than learning a completely new assessment system. We believe that internet users could benefit from accessing these models' predictions in-the-wild and thus assess their possible wrapper applications. We develop a chatbot, a browser extension, and a standalone web application.

To demonstrate the feasibility of this in-the-wild application of disinformation classifiers, we follow the process of developing such a tool from scratch. All of the artifacts produced in our study are open-source so that entities wishing to create similar projects can reproduce our results for their regional context effortlessly. Our contributions can be summarized as follows:

- We collect a novel Czech news article dataset with more than 22,000 articles from 9 sources, along with a methodology for their credibility categorization.
- We fine-tune multiple language models for the restated tasks of news source classification and source credibility label classification, whose results are reported in Section 4.1.
- We create three example in-the-wild wrappers (applications) of the newly trained models: a Messenger chatbot, a browser extension, and a standalone web application.
- We open-source the dataset, the training code, model weights, and code for all three wrapper applications under the Creative Commons CC BY-NC 4.0 license [1] at https:

[1] https://creativecommons.org/licenses/by-nc/4.0/

//github.com/matyasbohacek/misinfo-detection-wild-emnlp22.

2 Related work

In this section, we review other works which presented datasets for the task of disinformation classification. For a survey of the disinformation classification or automated-fact checking methods as such, we refer the reader to Oshikawa et al. (2020) and Guo et al. (2022) respectively. Apart from classification methods that use articles' full text at the input, numerous works have studied utilizing granular manipulative techniques (Zhang et al., 2018) or associated metadata instead (e.g., authors, hyperlinks) (Sitaula et al., 2020).

Most of the disinformation classification datasets in the public domain have emerged after 2017 (D'Ulizia et al., 2021). The most prominent and intensively studied ones have become the LIAR, FEVER, r/Fakeddit, and FakeNewsNet datasets.

Wang (2017) have proposed the LIAR dataset consisting of shorter excerpts of political speeches and quotes across six trustworthiness classes. The dataset includes over $10,000$ instances in total. Similarly, Shu et al. (2020) have introduced the FakeNewsNet, which holds over $20,000$ instances and distinguishes two basal classes (fake or real). These texts are primarily political quotes and speech excerpts, too.

Next, Thorne et al. (2018) have presented the FEVER dataset, which includes nearly $200,000$ instances of concise texts with respective links to Wikipedia. The annotations include whether the statements dispute or not, and thus this dataset has a larger basis in the task of stance detection. Lastly, we mention the r/Fakeddit dataset by Nakamura et al. (2020), which contains Reddit posts automatically annotated with a trustworthiness label derived from the overall credibility of the originating subreddit.

We also wish to highlight that many recent works focus on languages other than English. Resources for disinformation detection have been introduced for Arabic (Khalil et al., 2022; Bsoul et al., 2022), Danish (Derczynski et al., 2019), French (Meddeb et al., 2022), and others. For a detailed survey of other datasets with less traction in the literature, we refer the reader to D'Ulizia et al. (2021).

While the listed datasets are usually referred to as the best training and evaluation resources for disinformation (or fake news) classification, none actually hold news articles' data. In fact, most contain just shorter texts or excerpts. Moreover, all of these infer the individual items' class based on the overall source credibility while providing little or no methodology that would support their approach in terms of media sciences.

3 CNSC Dataset

Herein we present the Czech news source classification dataset (CNSC). In the latter subsections, we review the technical details of the data acquisition, the methodology for news source credibility categorization, and lastly, present statistics of the data.

3.1 Technical details

We have selected 9 Czech news domains for the collection of our dataset. To first obtain URLs of sites with individual articles from those domains, we used the Commoncrawl API [2]. We specified for the API to include only articles discovered between January 2019 and September 2021. Once these were obtained, we manually reviewed a random set of the data to find any undesired data points that also inhabit the respective domains (such as discussion forums or pages about the authors) and set up general flags to filter for these. We then scraped structured data of these articles using the Newsplease library (Fhamborg). After looking at the lengths of the texts, we noticed outliers that had as many as $25,000$ characters in length. These often included articles, for which the scraping library incorrectly yielded user discussions as parts of the text. We hence filtered any articles that would have more than $10,000$ characters.

This process resulted in a dataset of $22,001$ articles with the following textual attributes for each article item: title, text, URL, source name, author, and metadata description.

3.2 Methodology

To provide additional information about the news sources contained in the dataset, we created a methodology for their overall credibility categorization. Note that one cannot straightforwardly derive the truthfulness of all articles from any given source solely by the respective credibility class. It

[2]Commoncrawl library, https://commoncrawl.org/

Source	Source label	CNSC article examples
ČT24	Credible	*(Original Czech version:)* **Title:** Přibývá žen s rakovinou plic. Hlavní příčinou jsou cigarety **Text:** „Nejvýznamnějším rizikovým faktorem bezpochyby je aktivní kouření, které podle střízlivých odhadů je odpovědné za 30 až 40 procent všech úmrtí na rakovinu. V případě rakoviny plic je podíl na vzniku onemocnění dokonce až devadesátiprocentní," upozornil primář kliniky pneumologie nemocnice Na Bulovce Norbert Pauk. ... *(Translated into English:)* **Title:** More women are getting lung cancer, cigarettes being the main cause **Text:** "The most significant risk factor is undoubtedly active smoking, which is responsible for 30 to 40 per cent of all cancer deaths, according to sober estimates. In the case of lung cancer, the contribution to the disease is as high as 90 per cent," said Norbert Pauk, head of the pneumology clinic at Na Bulovce Hospital. ...
Blesk	Tabloid	*(Original Czech version:)* **Title:** Kadeřávková o návratu do Ulice: Takovou smršť nelidskosti nečekala! **Text:** Vážné zdravotní problémy donutily herečku Annu Kadeřávkovou (21), aby zpomalila a některé věci ve svém životě přehodnotila. Dokonce i svůj konec v nekonečném seriálu Ulice. Do něj se teď vrací po dlouhých dvou letech. Jak svůj krok vysvětlila fanouškům? Když se minulý týden objevila zpráva, že v Ulici budeme moci opět přivítat Rozinu v podání Kadeřávkové, strhla se na herečku lavina různorodých reakcí. (...) HALÓ! ... *(Translated into English:)* **Title:** Kadeřávková on her return to Ulice: She didn't expect such a storm of inhumanity! **Text:** Serious health problems forced actress Anna Kadeřávková (21) to slow down and rethink some things in her life. Even her ending in the endless series Ulice. She is now returning to it after two long years. How did she explain her move to her fans? When the news broke last week that we will be able to see Rozina again in Ulica, played by Kadeřávková, an avalanche of different reactions came to the actress.
Zvědavec	Disinformative	*(Original Czech version:)* **Title:** Kdo ovládá Ameriku? III. **Text:** Dva ze čtyřech největších mediálních konglomerátů (Disney a Viacom) jsou v židovských rukou. Židovští manažeři řídí mediální podnik NBC Universal. Židé tvoří velké procento na vedoucích postech v Time Warner. Je nepravděpodobné, že by tak velká míra židovského vlivu v této oblasti nastala bez cílené, záměrné snahy ze židovské strany. ... *(Translated into English:)* **Title:** Who controls America? III. **Text:** Two of the four largest media conglomerates (Disney and Viacom) are in Jewish hands. Jewish executives run NBC Universal's media business. Jews make up a large percentage of the top positions at Time Warner. It is unlikely that such a large degree of Jewish influence in this area would occur without a focused, deliberate effort on the Jewish side. ...

Table 1: Example items (articles) from the CNSC dataset spanning all three credibility source labels, which were assigned according to our methodology (described in Subsection 3.2).

Model	Architecture	Classification task	F-1 score	Precision	Recall
Czert	BERT		0.94	0.95	0.94
Small-E-Czech	ELECTRA	NSC (source)	0.87	0.88	0.86
RobeCzech	ROBERTA		**0.95**	0.**96**	**0.95**
Czert	BERT		0.96	**0.97**	0.96
Small-E-Czech	ELECTRA	SCLC (source label)	0.93	0.94	0.93
RobeCzech	ROBERTA		**0.97**	**0.97**	**0.97**

Table 2: Top-1 macro F-1 score, precision, and recall of the individual fine-tuned models on the NSC and SCLC tasks, as further described in Subsection 4.2.

should serve as a general, indicative flag of the prevailing trend and with which level of caution the author should read its articles.

For most languages and regions, open-source studies on the state of credibility of the individual media houses, newspapers, and news sites are available. These are often published by journalism activists, social scientists, and other involved figures. As one of the primary motivations of this work is to make the process less financially and organizationally demanding, we propose to re-use one of these works. When choosing the determinative one, we suggest preferring those of more diverse stakeholders and authors and whose methodology quantifies the overall assessments. This way, dividing individual credibility groups (labels) will be more exact.

We built upon the metrics and rankings of the Czech Endowment Fund for Independent Journalism (EFIJ) [3], but reduced the complexity of their final scale. The authors study various parameters of each source on a sample counting 100 of its articles and score them with detailed grades to maximize the objectiveness of the study. The parameters determining the source rating include:

- **Publication attributes:** Presence of authors by each article, transparent structure, and potential ownership conflicts (such as the owner being a politician);
- **Individual article attributes:** Usage of clickbait, stereotypization, hyperlinks, and more.
- **Editorial attributes:** Clear distinction between news reporting and commentaries, flagging of advertisement.

Each attribute is weighted and disposes of a specific prevalence reference. For instance, if less than 15 % of articles in a given source's sample contain a clickbait headline, the medium still receives a full score in the 'relevant headline' category. It receives half the score for a prevalence between 15 % and 30 % and no points for a clickbait rate above 30 %. Finally, the total of scores received across attributes determines the source's class. The category ranges are delineated as even portions of the scale for the given number of classes. In our case, these are three portions of the range between 0 and the maximal potential score. Each encompasses 33 % of the scale.

We arrived at three general classes of credibility. We provide their list with general descriptions below (for detailed description and analyses for each respective news source, we refer the reader to the EFIJ's website[4]):

- **Credible news sources:** Established and reliable news sources that are generally honest and truthful. Their articles contain hyperlinks to further sources of information, present arguments of all involved parties, distinguish between facts, speculations, and commentaries. (e.g., public media, objective press)
- **Tabloid news sources:** News sources one cannot rely on as generally honest and truthful. These sources often present speculations as facts or do not present arguments of all involved parties. (e.g., gutter media, press owned by political figures, press with strong political bias)
- **Disinformative / non-credible news sources:** News sources whose texts generally have no basis in fact but present themselves as being factually accurate. Such sources are often linked to (e.g., owned or funded by) entities intending to influence general political

[3] The Endowment Fund for Independent journalism, https://www.nfnz.cz/en/

[4] https://www.nfnz.cz/rating-medii/

views. (e.g., fake news media, state propaganda press)

Representative examples of the articles from our dataset are located in Table 1. These articles originate from 3 distinct news sources spanning all our source credibility labels. We can observe apparent differences in their topics and narrative styles: while the credible article deals with a factual description of a political event, the tabloid one presents news about a celebrity in a very sensation-seeking manner. Lastly, the disinformative report covers a conspiracy theory and disposes of a very constrained argumentative style.

3.3 Statistics

Herein, we present the statistics of the dataset. The complete set contains $22,001$ articles. We have created training, validation, and testing splits counting $17,600$ (80%), $2,200$ (10%), $2,201$ (10%) articles respectively. To constitute the splits, we sorted the articles by their publishing date and found two dates that would partition them into three temporally exclusive time windows of desired proportions. The distribution of articles by their source of origin is depicted in Figure 1a. As you can see, all of the sources except for ČT24, Extra.cz, and Zvědavec have a comparatively similar number of instances. The outliers result from our effort to preserve the overall trends in the volume of articles published by these sources every day and yet not develop significant margins. We have hence reduced the number of articles in most sources to compensate for the meager per-day publication rate of Zvědavec. As this source falls into the category of Disinformative / non-credible news sources, it can provide insight into how frequently such media publish instead of the conventional ones. The dataset class distribution for when the articles are grouped by their credibility label is shown in Figure 1b.

We have also evaluated the text lengths of the articles in the dataset. We used the NLTK library [5] for tokenization and filtering of punctuation. The articles from credible sources are, on average, 291 words long, while the tabloid and disinformative media have a mean of 379 and 551 words per article, respectively. The large margin between these counts for the credible and disinformative sources (almost double the value) caught our attention. We later reviewed the data manually and confirmed that this was not a mistake in scraping.

Overall, we can observe that while the disinformative sources tend to publish less frequently, their articles are, on average, recognizably longer. During the manual analysis of these articles, we also observed a trend of mentioning many seemingly unrelated topics from different areas at once. We hypothesize that this may be caused by the conspiratory nature of such sources, in which they draw false links and causations between uncorrelated events. Nevertheless, this calls for a thorough analysis of its own. We believe our dataset can serve as the first reference for further studies on such news patterns in the central European regional context.

4 Baseline experiments

In the following section, we present the baseline results for the two newly formulated tasks on the CNSC dataset:

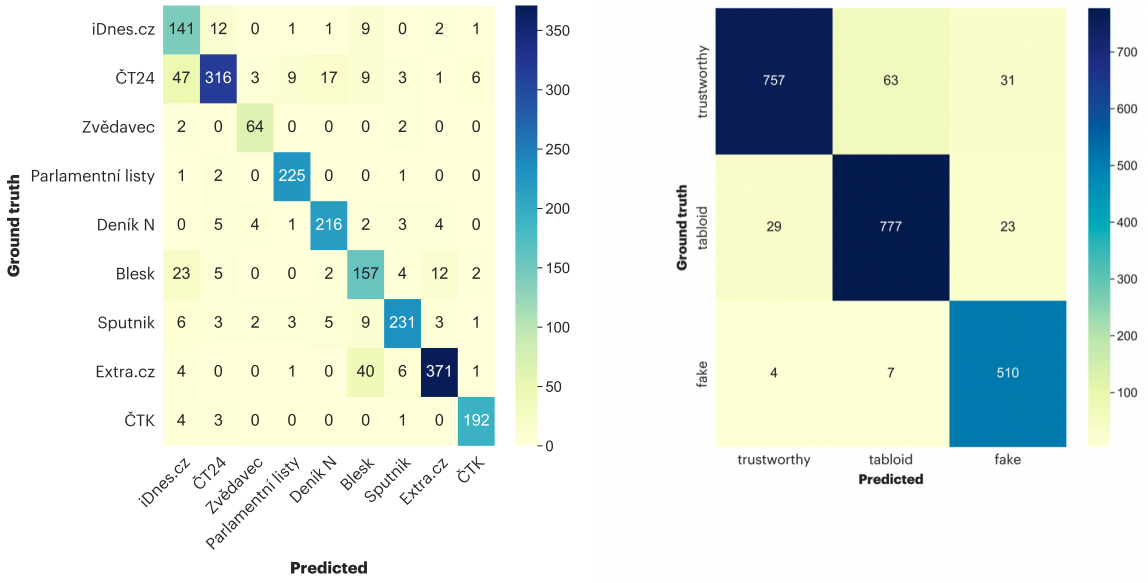
- **News source classification (NSC):** the task is to classify the originating news source of an article based on its title and body texts from a pre-defined set of media,
- **Source credibility label classification (SCLC):** the task is to classify the news source credibility label to which the article's originating news source belongs based on its title and body texts from a pre-defined set of media.

In this particular case, the number of classes for NSC corresponds to the number of news sources present in the dataset (9). The number of classes for the SCLC task corresponds to the number of credibility labels (3), as outlined in Subsection 3.2.

4.1 Experimental setting

We fine-tune three language model architectures for this purpose: Czert (Sido et al., 2021) (based on BERT (Devlin et al., 2019)), Small-E-Czech (Kocián et al., 2021) (based on ELECTRA (Clark et al., 2020)), and RobeCzech (Straka et al., 2021) (based on ROBERTA (Liu et al., 2019)). We use the HuggingFace Transformers library for implementation and train the models using a learning rate of $2e-5$ for 4 epochs. When obtaining the embeddings for all the examined models, we concatenate the article's title with its text as if it were the first sentence of the body. To

[5] NLTK library, https://www.nltk.org/



(a) per individual news sources

(b) per news source labels

Figure 2: Confusion matrices of the fine-tuned Small-E-Czech's predictions on our CNSC dataset test split. The architecture and training configuration details can be found in Subsection 4.1.

prevent the models from learning other undesired correlated artifacts, which may be left in the article (such as names of the source occurring at the beginning), we delete any occurrences of the article's originating news source name in the body. We also remove common rubric identifiers from the title (e.g., 'Commentary:', 'Interview'). We open-source our training scripts at https://github.com/matyasbohacek/misinfo-detection-wild-emnlp22.

4.2 Results

We present the results in Table 2. The F-1 score, precision, and recall on the testing set are included for each model and task. We can observe that the tested models managed to learn the individual news sources' characteristics in writing for both tasks and generally achieved reasonable performance, with the F-1 scores around 0.9. We found RobeCzech to be performing best in both tasks by reaching 0.95 and 0.97 F-1 scores on the NSC and SCLC tasks, respectively. On the other hand, Small-E-Czech has performed the worst by resulting in respective F-1 scores of 0.87 and 0.93. We presume this is caused by the model's size, as Small-E-Czech is dramatically smaller than the other two models in parameter counts. Lastly, we also evaluated the fine-tuned Czert, which scored under the best RobeCzech with 0.94 and 0.96 respective F-1 scores on the two tasks.

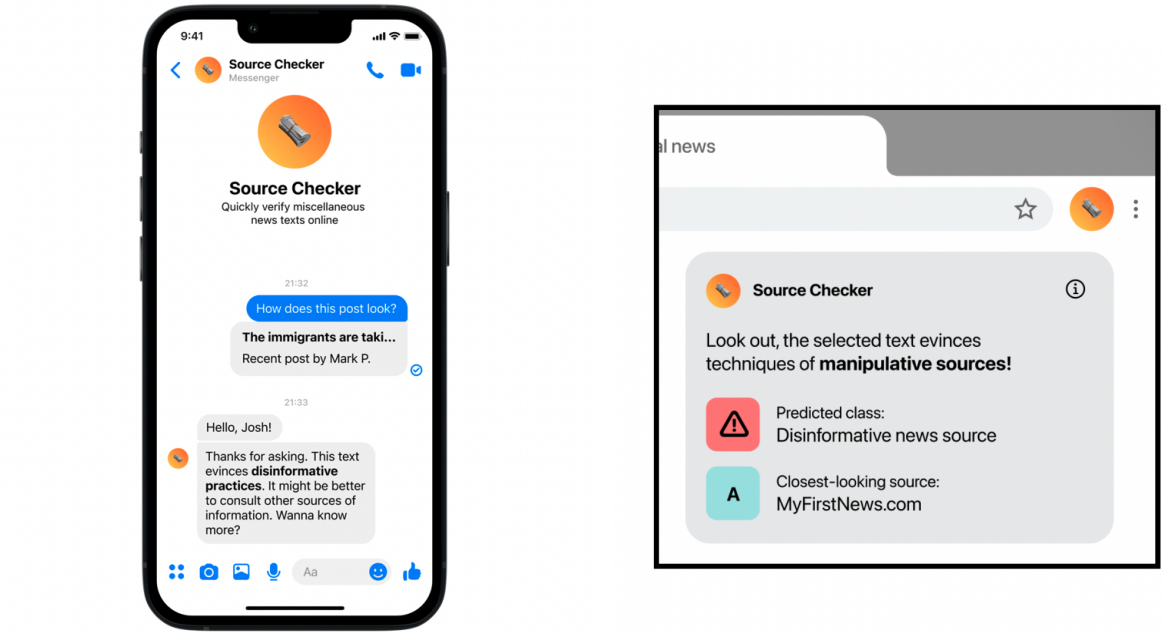
We further depict the confusion matrices of the Small-E-Czech's predictions for both tasks on the test split in Figure 2. As can be observed in Figure 2b, most erroneous predictions mistake the trustworthy and tabloid labels, while there are only a few false positives predictions of the fake label. We argue that this may be caused by the unique and highly distinctive vocabulary used in conspiracies. Trustworthy and tabloid articles, on the other hand, dispose of differences in their narratives that our models can also capture, but often share the topics of general public discourse, and therefore have less distinguishing vocabulary.

5 In-the-wild wrapper applications

We construct and open-source three in-the-wild wrappers of the just-described models. We do so to support future studies of such interventions' efficacy and associated user behavior. As the primary motivation of our work lies in enabling internet users to gauge the perceived trustworthiness of various texts online, we want the tools to be easily reachable from different workflows. The applications thus include:

1. **Standalone web application.** Created using Gradio [6], we present a simple website that

[6] Gradio library, https://gradio.app



(a) Messenger chatbot companion

(b) Browser extension

Figure 3: Screenshots of the end-customer model wrappers with mock data, as described in Section 5.

enables users to insert text and quickly see the top 3 predicted classes by both models.

2. **Browser extension.** As depicted in Figure 3b, we build a standard Chromium-based [7] browser extension, letting users infer the models with any highlighted text on the screen. The extension shows the most likely originating source and its respective trustworthiness class.

3. **Chatbot.** To serve mobile users, too, we create a Facebook Messenger chatbot, which wraps the inference of both models in a simple prompt heuristic. Apart from the inference features, the chatbot comes with additional explanatory phrases and links built in. A mock conversation is shown in Figure 3a.

6 Ethical Discussion and Limitations

In this section, we review the limitations of our solution and discuss the ethical aspects of its use. As already mentioned, one must bear in mind that the overall credibility of a given news source does not deduce all of its articles' trustworthiness or factual correctness. Still, different studies (Cone et al., 2019; Pehlivanoglu et al., 2021) found the source trustworthiness to be an effective indicator of its articles' credibility (especially when other coverage or context are limited). The literature on machine learning identification has mainly built classifiers on this premise. We believe this approach offers a reasonable trade-off between the annotation complexity and overall performance. In our solution, the originating news source serves as a proxy of credibility. While writing in a style of a particular outlet does not, once again, conclusively derive the text's eventual trustworthiness, detecting patterns used in fraudulent and hoax outlets can provide a helpful warning flag for potentially deceptive and harmful texts. Any publicly available application of this technology should clearly state this information at the very beginning and provide its users with additional resources about the methodology. Moreover, the users should be aware that the analysis is automatic. We include examples of best practices (with short descriptions easily understandable by the general public) in our wrapper applications.

The technology could be misused by falsely labeling misinformation as trustworthy and manipulating its users according to the agenda of the service provider. Therefore, we believe only trusted, independent institutions (e.g., university-affiliated centers and non-governmental organizations) should assume the role of operators. We advise prospective providers to disclose the source labeling methodology and the samples used fully.

[7] The Chromium Projects, https://www.chromium.org

7 Conclusion

We show that when appropriately adapted and wrapped, the existing methods for disinformation detection can serve as supportive tools for the new form of disinformation contexts online. We present an open-source CNSC dataset with over $22,000$ Czech news articles spanning 9 sources across the credibility spectrum, the first of its kind in such a small language. We build on top of a detailed methodology for news trustworthiness assessment in the Czech Republic and establish 3 credibility classes for the news sources. We train baseline models for the news source and source credibility label classification and achieve F-1 scores of 0.95 and 0.97, respectively. Lastly, we introduce three in-the-wild wrapper applications of our models, whose code we are making public.

In our future work, we want to conduct focus group studies analyzing the efficacy and user behavior of the intervention tools we introduced. We also intend to propose better metrics and benchmarks for detecting the ever-evolving disinformation.

Acknowledgements

I would like to thank Kateřina Lesch for the incredible initial insights into academic writing and the internship opportunity in her team I was given despite my age. I would also like to express profound gratitude to Tomáš Trnka for multiple rounds of reviews and discussions about this matter and paper.

References

Vlad Achimescu and Pavel Dimitrov Chachev. 2020. Raising the flag: Monitoring user perceived disinformation on reddit. *Information*, 12(1):4.

Mohammad A Bsoul, Abdallah Qusef, and Saleh Abu-Soud. 2022. Building an optimal dataset for arabic fake news detection. *Procedia Computer Science*, 201:665–672.

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*.

Jeremy Cone, Kathryn Flaharty, and Melissa J Ferguson. 2019. Believability of evidence matters for correcting social impressions. *Proceedings of the National Academy of Sciences*, 116(20):9802–9807.

Leon Derczynski, Torben Oskar Albert-Lindqvist, Marius Venø Bendsen, Nanna Inie, Viktor Due Pedersen, and Jens Egholm Pedersen. 2019. Misinformation on twitter during the danish national election: A case study. In *Truth and Trust Online*.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Fhamborg. Fhamborg/news-please: News-please - an integrated web crawler and information extractor for news that just works.

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

Ashwaq Khalil, Moath Jarrah, Monther Aldwairi, and Manar Jaradat. 2022. Afnd: Arabic fake news dataset for the detection and classification of articles credibility. *Data in Brief*, 42:108141.

Matěj Kocián, Jakub Náplava, Daniel Štancl, and Vladimír Kadlec. 2021. Siamese bert-based model for web search relevance ranking evaluated on a new czech dataset. *arXiv e-prints*, pages arXiv–2112.

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.

Paul Meddeb, Stefan Ruseti, Mihai Dascalu, Simina-Maria Terian, and Sebastien Travadel. 2022. Counteracting french fake news on climate change using language models. *Sustainability*, 14(18):11724.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6149–6157.

Jack Nicas. 2020. Why can't the social networks stop fake accounts?

Ray Oshikawa, Jing Qian, and William Yang Wang. 2020. A survey on natural language processing for fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6086–6093.

Joanna Paliszkiewicz, Magdalena Mądra Sawicka, Tadeusz Filipiak, Salome Svanadze, and Mariam Jikia. 2017. Time-spent online as a factor in usage and awareness of drawbacks in social media. *Issues in Information Systems*, 18(4).

Didem Pehlivanoglu, Tian Lin, Farha Deceus, Amber Heemskerk, Natalie C Ebner, and Brian S Cahill. 2021. The role of analytical reasoning and source credibility on the evaluation of real and fake full-length news articles. *Cognitive research: principles and implications*, 6(1):1–12.

Kira E Riehm, Kenneth A Feder, Kayla N Tormohlen, Rosa M Crum, Andrea S Young, Kerry M Green, Lauren R Pacek, Lareina N La Flair, and Ramin Mojtabai. 2019. Associations between time spent using social media and internalizing and externalizing problems among us youth. *JAMA Psychiatry*, 76(12):1266.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.

Jakub Sido, Ondřej Pražák, Pavel Přibáň, Jan Pašek, Michal Seják, and Miloslav Konopík. 2021. Czert–czech bert-like model for language representation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1326–1338.

Niraj Sitaula, Chilukuri K Mohan, Jennifer Grygiel, Xinyi Zhou, and Reza Zafarani. 2020. Credibility-based fake news detection. *Disinformation, Misinformation, and Fake News in Social Media*, page 163.

Milan Straka, Jakub Náplava, Jana Straková, and David Samuel. 2021. Robeczech: Czech roberta, a monolingual contextualized language representation model. In *International Conference on Text, Speech, and Dialogue*, pages 197–209. Springer.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

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, Vancouver, Canada. Association for Computational Linguistics.

Amy X Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, et al. 2018. A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion Proceedings of the The Web Conference 2018*, pages 603–612.