POLygraph: Polish Fake News Dataset

Daniel Dzienisiewicz, Filip Graliński, Piotr Jabłoński
Marek Kubis, Paweł Skórzewski, Piotr Wierzchoń
Adam Mickiewicz University, Poznań
ul. Wieniawskiego 1, 61-712 Poznań, Poland
{dzienis,filip.gralinski,piotr.jablonski,
marek.kubis,pawel.skorzewski,piotr.wierzchon}@amu.edu.pl

Abstract

This paper presents the POLygraph dataset, a unique resource for fake news detection in Polish. The dataset, created by an interdisciplinary team, is composed of two parts: the "fake-or-not" dataset with 11,360 pairs of news articles (identified by their URLs) and corresponding labels, and the "fake-they-say" dataset with 5,082 news articles (identified by their URLs) and tweets commenting on them. Unlike existing datasets, POLygraph encompasses a variety of approaches from source literature, providing a comprehensive resource for fake news detection. The data was collected through manual annotation by expert and non-expert annotators. The project also developed a software tool that uses advanced machine learning techniques to analyze the data and determine content authenticity. The tool and dataset are expected to benefit various entities, from public sector institutions to publishers and fact-checking organizations. Further dataset exploration will foster fake news detection and potentially stimulate the implementation of similar models in other languages. The paper focuses on the creation and composition of the dataset, so it does not include a detailed evaluation of the software tool for content authenticity analysis, which is planned at a later stage of the project.

1 Introduction

This paper describes a dataset created for a project aimed at detecting and analyzing fake news on the Polish web. Fake news poses a significant threat in real-world situations, eroding trust in institutions, manipulating public opinion, and fueling societal tensions. To address this challenge, our project employs a unique hybrid research approach, merging narratological, comparative, and sociological techniques with natural language processing and big data analytics. An interdisciplinary team of experts in various fields, including mathematics, computer science, philology, media studies, law, philosophy, folklore, and IT, collaborates on this endeavor. The project aims to develop a fake news detection software tool that uses a comprehensive database of sources, authors, and content, as well as advanced machine learning techniques and implicit trust ranking analyses to determine the authenticity of the content.

The dataset described in this paper consists of two parts. The first part, referred to as the "fake-or-not" dataset, contains 11,360 pairs of news articles (identified by URLs) and labels indicating whether the news is fake or not. The second part, known as the "fake-they-say" dataset, comprises 5,082 news articles (identified by URLs) and tweets commenting on them. Each tweet is accompanied by a label expressing the commentator's opinion about the article's truthfulness.

Our software tool and its underlying dataset are intended to serve various beneficiaries, including public sector entities like the Ministry of Internal Affairs and Administration, the Ministry of Defense, the Police, the Internal Security Agency, and the Internal Security Service for public safety purposes. It could also be helpful for publishers, the Warsaw Stock Exchange, the Financial Supervision Commission (to monitor potential manipulations affecting company valuations or the country's macroeconomic status), fact-checking organizations, and analytical firms.

2 Related Work

2.1 Tasks and Datasets

In today's digital age, the rapid dissemination of information has led to an intertwined web of factual narratives and misinformation. The challenge of distinguishing between the two has spurred extensive research in various domains. Tasks such as fact verification (Schuster et al., 2019; Lewis et al., 2020), fact-checking (Wang, 2017; Bhattarai et al., 2022), fact-based text editing (Iso et al.,

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 250–263
August 15, 2024 ©2024 Association for Computational Linguistics

2020), and table-based fact verification (Chen et al., 2020; Eisenschlos et al., 2020) are crucial in this endeavor. The complexity is further heightened by the introduction of counterfactual elements, which encompass counterfactual detection (Yang et al., 2020), inference (Pawlowski et al., 2020; Poulos and Zeng, 2021), and explanation (Plumb et al., 2020; Ramon et al., 2020). Moreover, the classification of comments (Bornheim et al., 2021) based on their toxicity, engagement, and fact-claiming nature is an emerging area of interest.

The broader challenge of misinformation (Thorne and Vlachos, 2021; Bhattarai et al., 2022) encapsulates various facets, including fake news detection (Shu et al., 2017; Wang, 2017), deepfake detection (Rossler et al., 2019; Li et al., 2020b), and fake image detection (Afchar et al., 2018; Rossler et al., 2019). The political sphere, as evidenced by stance detection tasks (Hanselowski et al., 2018; Borges et al., 2019) related to the US 2020 Election (Kawintiranon and Singh, 2021), is particularly susceptible to these challenges. Complementary research areas such as hate speech detection (Davidson et al., 2017; Mathew et al., 2021), propaganda technique identification (Blaschke et al., 2020), aggression identification (Orăsan, 2018; Risch and Krestel, 2018), satire detection (Li et al., 2020a; Ionescu and Chifu, 2021), humor detection (Castro et al., 2016; Weller and Seppi, 2019), rumor detection (Kochkina et al., 2017; Zubiaga et al., 2018; Gorrell et al., 2019), and deception detection (Guo et al., 2023) further underscore the multifaceted nature of this challenge.

Several datasets and competitions, such as those hosted on Kaggle[1] and the ISOT Fake News Dataset (Ahmed et al., 2017, 2018), have been developed to foster advancements in this domain. RumourEval competition (Gorrell et al., 2019) provided a dataset of dubious posts and ensuing conversations in social media, annotated both for stance and veracity. The competition received many submissions that used state-of-the-art methodology to tackle the challenges involved in rumor verification. Another example is the FEVER (Fact Extraction and VERification) dataset (Thorne et al., 2018), consisting of 185,445 claims generated by altering sentences from Wikipedia and subsequently classified without knowledge of the sentence they were derived from as “supported”, “refuted”, or “not enough info”.

For a comprehensive approach, it is imperative to integrate diverse sources, including fact-checking websites, encyclopedias, urban legends, conspiracy theories, and Wikipedia entries on fake news. Archival resources, such as the urban legend archive curated by Graliński (2012), offer unique insights. Furthermore, domain-specific datasets, focusing on works of sci-fi authors like Lem, Pratchett, and Sapkowski, or niche forums like Wykop.pl[2] and Hyperreal[3], provide a rich tapestry of data for analysis. An example of such a dataset is BAN-PL (Kolos et al., 2024), collecting content from the Wykop.pl web service that contains offensive language, which makes an essential contribution to the automated detection of such language online, including hate speech and cyberbullying.

Our methodology for categorizing fake news and non-fake news is anchored in established guidelines, as outlined by resources like EUfactcheck[4]. Additionally, the emergence of fake news detectors, evident in browser plugins and extensions such as SurfSafe[5], Reality Defender[6], or Fake News Chrome Extension[7], presents promising avenues for real-time misinformation mitigation.

This research aims to introduce a comprehensive Polish fake news dataset to lay a robust foundation for future endeavors in the realm of misinformation detection and analysis within the Polish context.

2.2 Annotation Methodologies

The current fake news detection techniques can be classified into several groups. For instance, according to Wang et al. (2021), there are three categories of methods: propagation structure-based, user information-based, and news content-based. Propagation structure-based methods involve extracting features related to news dissemination in social media. User information-based methods focus on the users involved in the circulation of news, covering aspects such as users’ gender, social media friends, followers, and location. On the other hand, news content-based methods concentrate solely on analyzing the content of the news rather than information about users and news dissemination.

[1] https://www.kaggle.com/c/fake-news

[2] https://wykop.pl

[3] https://hyperreal.info

[4] https://eufactcheck.eu/wp-content/uploads/2020/02/EUfactcheck-manual-DEF2.pdf

[5] https://www.getsurfsafe.com/

[6] https://realitydefender.com

[7] https://tlkh.github.io/fake-news-chrome-extension

A mixed approach to fake news detection was proposed by Zhang and Ghorbani (2020), who identified four components considered particularly important in characterizing fake news: creator/disseminator, target, news content, and social context. Zhou and Zafarani (2020), on the other hand, divide fake news detection models into methods based on the analysis of the annotator's knowledge (knowledge-based fake news detection), the style in which the news is written (style-based fake news detection), the method of disseminating the news (propagation-based fake news detection), and assessing the credibility of news sources (source-based fake news detection).

3 Data Collection

The POLygraph: Polish Fake News Dataset was collected entirely from the Internet. The research team designed a mechanism using two methods: API data access and web scraping. For Twitter (nowadays X), we utilized the Twitter API[8], which provided a powerful set of tools for Academic Researchers[9] at the time. This allowed us to access archived data without putting additional strain on web services. The functions and methods provided in the API allowed us to search and filter the entire available content of Twitter freely, going all the way back to the first published tweet in 2006[10]. We downloaded tweets from 2021-01-01 to 2022-04-30 to match the timeframe of other data sources. Twitter API provided the ability to search the entire archive and download up to 10 million tweets. For websites, a custom scraper was employed to extract and save only the relevant content.

3.1 Sources, Contents, and Authors

The database of 5,000 sources was prepared by scraping a list of 1,300 starter websites. The scraper then visited at least 25 documents from each page and extracted subsequent links to external documents. Then, it repeated the process of searching and archiving documents. The XPath expression used to extract links from documents[11] provided the ability to retrieve all links whose `href` attribute does not start with `mailto:` or `tel:` and then return them as a list. In the next step, this list was iterated, and each address was passed to the parser, which added the address to the internal queue. The scraped pages were archived as HTML files with linked materials in a structure consistent with the command `wget -H -k -r -l 1 url`. The downloaded HTML files were automatically anonymized and then compressed into a ZIP archive, taking as the name documents a 128-bit hash function calculated based on the URL of the archived document.

3.2 Tagged Press Articles

The aim of this stage of data collection was to create a database of about 3,000 tagged press articles. For this purpose, we queried Twitter to search for tweets whose content would be related to commenting on the truthfulness of the information, particularly expressing the opinion that some content constitutes fake news. We expected that entries of this type would contain references to newspaper articles and other sites that would be interesting to annotate for potential false information. To obtain the URLs we were interested in, we used access to the Twitter API. We performed two variants of this search, differing in the query used and the time frame, resulting in two sets of entries:

- V2 dataset – a query focused on finding tweets where the author directly expresses their opinion on whether something is fake or not; uses phrases like "it wasn't fake" and "it was fake" in Polish and English[12] (1–29 April 2022; 574,545 obtained entries).
- V3 dataset – a query like in V2, but extended with terms for debunking or verifying information, e.g., "verified", "correction", "where is this info from"[13] (1 January 2010–31 July 2022; 3,580,901 obtained entries).

In total, we collected 4,155,446 tweets. Using a script to extract URLs from text, we obtained 339,259 URLs from this set.

[8] https://developer.twitter.com/en/docs/twitter-api

[9] http://web.archive.org/web/20230212021429/https://developer.twitter.com/en/products/twitter-api/academic-research

[10] https://twitter.com/jack/status/20

[11] `response.xpath(”//body//a[not(starts-with(@href,’mailto:’))][not(starts-with (@href,’tel:’))]/@href”).getall()`

[12] `(lang:pl (fejk OR fake OR fakenews OR ”to nie był fake” OR ”to był fake” OR ”to nie był fejk” OR ”to był fejk”)) OR (lang:en (fejk OR ”to nie był fake” OR ”to był fake” OR ”to nie był fejk” OR ”to był fejk”))`

[13] `(lang:pl (fake OR fakenews OR ”fake news” OR factcheck)) OR (”to byl fejk” OR ”to byl fake” OR ”to nie byl fejk” OR ”to nie byl fake” OR fejk OR ”fejk-njus” OR dementi OR zweryfikowane OR ”zrodlo potwierdzone” OR sprostowanie OR sprostowane OR ”skad to info” OR ”skad ta informacja” OR ”przepraszam za podanie”)`

The list of URLs was processed with another script, which uses Mercury Parser[14], `html2text`[15], and BeautifulSoup[16] to extract text from the website located at the given URL. During the script execution, the following are rejected:

- pages for which Mercury Parser found no text,
- pages for which the HTML returned by Mercury Parser was empty,
- pages that failed to convert HTML to text with either `html2text` or BeautifulSoup,
- pages whose language, detected based on the text using the `langdetect5` library, was other than Polish,
- pages for which `langdetect5` was unable to detect the language,
- repeated pages.

As a result, we received 63,776 examples in the JSON format supported by the Doccano (Nakayama et al., 2018) annotation tool.

To give annotators access to a website preview, we created a spider (web crawler) that takes screenshots of the pages referenced by the URLs in the list and saves them to PNG files. The script uses the Scrapy[17] framework and the `splash`[18] library. Then, using another script, we filtered the obtained examples in JSON format, discarding those for which it was impossible to take a screenshot of the page. Ultimately, we received 7,242 examples in JSON format (for Doccano), divided into 19 packages of 400 examples each (the last package was incomplete). In this way, a collection of articles was prepared for detailed tagging. Each example in the collection was designated for annotation by at least three independent annotators. The annotation was carried out using the Doccano platform, as described in Section 4.1.

3.3 Tweets Expressing Opinions about Press Articles

The starting point for obtaining a database of tweets expressing opinions about press articles was the dataset of 4,155,466 tweets described in Section 3.2. The subsequent processing stage was to extract external URLs of websites in Polish from this set of tweets. We wanted the resulting list of URLs to be both representative and diverse. To achieve this, we only considered one entry from each author and discarded URLs obtained through URL shorteners because they were likely redirects to other URLs in the set. Of the 4,155,446 tweets we rejected:

- 3,249,033 tweets that did not refer to any external URL,
- 466,002 tweets in a language other than Polish,
- 197,208 tweets whose author was repeated,
- 63,885 tweets that contained more than one link to an external URL, and it was not possible to clearly indicate which of them they directly referred to,
- 46,665 tweets containing a URL that was most likely obtained using a shortener,
- 38,720 tweets containing the URL of a fact-checking website,
- 18,999 tweets containing an invalid URL.

74,934 examples left.

We wrote a Python spider called `tsv2pngs` using the Scrapy framework and the `splash` library. For each example from the source `data.tsv` file, the spider takes a screenshot of the tweet and a screenshot of the page the tweet refers to, combines them and saves the result as a PNG file. To access tweet content more easily, we used the Nitter service, which is a free, open-source front-end Twitter mirror. Before combining the screenshots, we scale them as needed to ensure the resulting PNG file is readable for annotators. Screenshots with aspect ratios (picture height to width ratio) greater than 8:1 are rejected. As a result, we obtained 22,206 PNG images of page screenshots. A script that transforms data from TSV to JSONL files allowed us to obtain 74,934 examples in JSONL format. An additional script utilizes the `urllib` library to filter out specific examples from the input file, including those without corresponding PNG screenshots, those with repeated website domains, and those that are part of a user-provided list. In our case, we supplied a list of examples annotated as part of

[14] https://hub.docker.com/r/wangqiru/mercury-parser-api

[15] https://github.com/Alir3z4/html2text

[16] https://www.crummy.com/software/BeautifulSoup

[17] https://scrapy.org

[18] https://splash.readthedocs.io

the pilot annotation. Ultimately, we ended up with 8,108 examples divided into three packages, which constituted data for three "fake-they-say" annotation tasks on the Doccano platform, described in Section 4.4.1.

4 Data Annotation

4.1 *Fake-or-not* Annotation Methodology

The starting point for creating a set of questions for annotators in the discussed POLygraph dataset was the annotation scheme used in research on fake news in Japanese media by Murayama et al. (2022). The cited researchers proposed an annotation scheme that includes seven types of information: 1) the factuality of the news, 2) the disseminator's intention, 3) the target of the news, 4) the sender's attitude towards the recipient, 5) purpose of the news, 6) degree of social harmfulness of the news, 7) the type of harm that the news can cause.

The above set of questions is multidimensional, as it allows for considering a more comprehensive range of information than just the factuality aspect of the news. However, our catalog of questions expands beyond the above data. Although it is dominated by a text-centric approach, the questions are also aimed, among others, at determining the annotator's attitude towards the content, which helps recognize their bias and emotions evoked by the text. The detailed list of all 19 questions used in this annotation and related statistics are presented in Appendix A.

The annotation was performed on the Doccano platform by a total of 161 annotators. The annotators in this task were experts and students of political sciences and journalism (see Section 9). All annotators underwent detailed training, including special case analysis. The total number of annotated news articles was 7,006, including 6,339 articles annotated by at least two independent annotators. The level of agreement between annotators was estimated by calculating Fleiss' kappa and varied depending on the question.

It is worth noting that our questionnaire contained many subjective and ambiguous questions because we wanted to investigate fake news in depth. Therefore, we do not expect perfect agreement among human annotators, especially when dealing with ambiguous or controversial cases. The nuanced nature of fake news detection further contributes to this expectation. The agreement scores reported by other studies on similar tasks take values around $0.3 \sim 0.4$. For instance, the RumourEval 2019 shared task achieved a Fleiss' kappa of 0.39 for veracity annotation and 0.35 for stance annotation (Gorrell et al., 2019). Thus, we believe that kappa scores within these limits would confirm the dataset's usefulness for the purpose for which it was built.

4.2 Gonito.net Platform

We used the Gonito.net (Graliński et al., 2016) platform with the GEval (Graliński et al., 2019) evaluation tool to store and manage training, validation and testing data and evaluate the models used in the project. Gonito.net is an open-source platform for comparing and evaluating machine learning models, enabling reproducibility of experiments. On the Gonito.net platform, individual machine-learning tasks are organized as so-called challenges. A challenge is a set of training, validation and test data stored in a Git repository, associated with a set of evaluation metrics. Solutions to individual challenges can be put on the platform (in the form of model prediction results on a test set), which are automatically assessed using the GEval tool according to metrics related to the challenge. We have prepared two challenges for the project: *fake-or-not* and *fake-they-say*.

4.3 *Fake-or-not* Challenge

The *fake-or-not* challenge is to create a model that will determine whether the article underneath it is fake news or not, based on the URL. The data for the challenge comes from three sources, detailed descriptions of which are provided below in the appropriate subsections: pilot annotation (Section 4.3.1), annotation tasks on the Doccano annotation platform (Section 4.3.2), and fact-checking websites (Section 4.3.3). Based on these three sources, a dataset (set A) was created containing 10,191 records – pairs: URL, label `1` (fake news) or `0` (not fake news). Set A was split in the proportions 9:2:5 into a training set (4,482 records), validation set (1,256) and test set (3,202). The split was made deterministically – based on the last hexadecimal digit of the MD5 hash function value for the URL. Additionally, set B was obtained from annotation tasks on the Doccano platform, containing 2,420 analogous records (pairs: URL, label `1`/`0`). Set B has been fully included in the training set. To sum up, we have a total of 6,902 records in the training set, 1,256 records in the validation set and 3,202 records in the test set. Out of all 11,360 records,

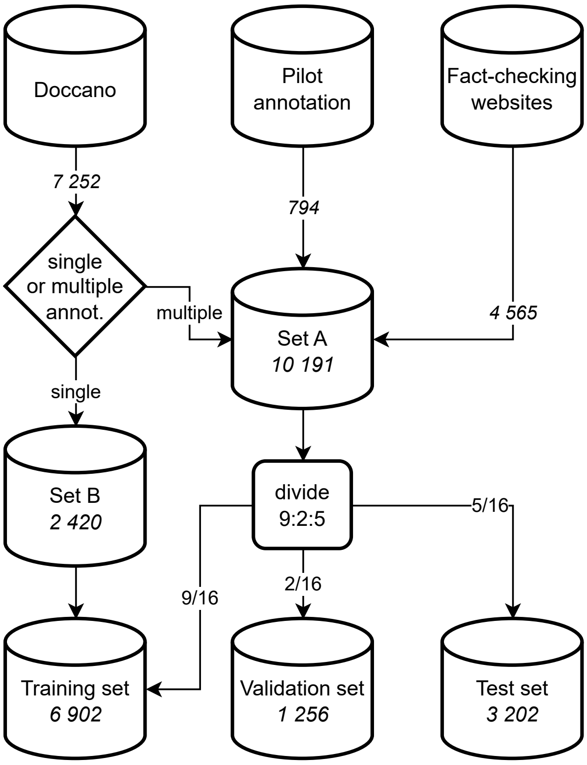

Figure 1: Data acquisition and processing workflow for *fake-or-not* challenge.

there were 4,350 records marked with label 1 and 7,116 records marked with label 0. The diagram in Figure 1 summarizes the whole process.

4.3.1 Pilot Annotation Task

As part of the annotation pilot, we prepared a set of 998 URLs of press articles. The method of collecting data is described in Section 3. Each article was annotated by two independent annotators with one of three labels: "fake news", "truth", and "unknown". The inter-annotator agreement measured by Cohen's kappa was 0.421. Then, URLs marked as "fake news" or "truth" by at least one annotator and for which both annotators' annotations did not conflict were labeled 1 and 0, respectively. This way, we obtained 794 records (97 with label 1 and 697 with label 0), which were added to set A.

4.3.2 Massive Annotation Task

Annotation of the tasks described in Section 4.1 consisted of answering 19 detailed questions about the text. We only used the answers to question 12 to prepare data for the fake-or-not challenge ("In your opinion, does the text contain false information?"). Annotated examples for which the annotator chose the answer "yes" or "no" were selected (the answer "not subject to assessment" was omitted). Replies have been grouped by the related URLs. If the majority of the annotations for a given URL were "yes", then a record consisting of the URL and label 1 was added to the dataset, whereas if the majority of the annotations for the given URL were "no", a record consisting of the URL address and label 1. The URL was omitted in case of an equal number of "yes" and "no" annotations.

Additionally, the obtained records were divided into two sets, depending on how many majority annotations there were for a given URL. If only one annotator indicated the majority answer (this also means that no annotator indicated the minority answer), the record was put in set B. Otherwise, i.e., if at least two annotators indicated the majority answer, the record was put in set A. This way, we obtained 7,161 records, with 4,751 records in set A and 2,410 in set B. The label distribution is shown in Table 1.

Set	Label 1	Label 0	Total
A	354	4,397	4,751
B	1,179	1,231	2,410
Total	1,533	5,628	7,161

Table 1: Label distribution

4.3.3 Data from Fact-checking Websites

Opinions from fact-checking websites (476 opinions from fakehunter.pap.pl, 2,125 opinions from demagog.org.pl, and 2,637 reviews from afp.com) were used as another source of data. If the opinion was expressed as "fake news", "false", "manipulation", etc., a record consisting of the appropriate URL address and label 1 was added to the dataset. If the opinion was expressed as "true", the appropriate record was tagged with 0. This way, we obtained 4,924 records (3,784 with label 1 and 1,140 with label 0), which were added to set A.

4.4 *Fake-they-say* Annotation

4.4.1 Annotation Methodology

The "fake-they-say" annotation task was developed to assess the degree of the tweet author's belief in the (un)truthfulness of the information they commented on. The annotators received access to the content of 1) the tweet being rated, 2) the entire discussion regarding the news, and 3) the news itself. The task was to read the content of the comment on a specific piece of news and/or the entire accompanying discussion and then select one of the following six labels defining the tweet author's attitude towards the content of the article:

- *hard-claim-fake* (the author of the tweet claims that the news they are commenting on is false),
- *hard-claim-not-fake* (the author of the tweet claims that the news they are commenting on is true),
- *no-claim* (it is impossible to determine what the author of the tweet thinks, or the comment does not refer to the issue of (un)truthfulness of the news),
- *sarcasm* (the author of the tweet is ironizing, expressing themselves sarcastically),
- *soft-claim-fake* (the author of the tweet probably believes that the news they are commenting on is false),
- *soft-claim-not-fake* (the author of the tweet probably does not think the news they are commenting on is false).

The annotators in this task were experts and students of political sciences and journalism. All annotators underwent detailed training. There were 48 annotators, and they annotated 4,356 press articles in total, including 3,235 articles annotated by at least two independent annotators. The level of agreement between annotators was estimated by calculating Fleiss' kappa as $\kappa = 0.4343$.

4.4.2 Challenge Description

The *fake-they-say* challenge is to create a model that, based on the tweet's text and the URL, will determine what the tweet's author thinks about the article located at the given URL. The data for the challenge comes from two sources (detailed descriptions provided below in the relevant subsections): pilot annotation and annotation tasks on the Doccano annotation platform. These two sources created a dataset containing 5,082 records, consisting of the following fields:

- label: one of the 6 labels described in Section 3.3 (*hard-claim-fake*, *hard-claim-not-fake*, *no-claim*, *sarcasm*, *soft-claim-fake*, *soft-claim-not-fake*),
- tweet text,
- tweet URL,
- URL address of the commented article,
- PNG image consisting of a screenshot of the tweet and a screenshot of the commented article.

The dataset was split in the proportions 13:1:2 into the training set (4,040 records), validation set (316 records) and test set (726 records). The split was made deterministically – based on the last hexadecimal digit of the MD5 hash function value for the URL. In total, we obtained 806 *hard-claim-fake* records, 102 *hard-claim-not-fake* records, 1,254 *no-claim* records, 44 *sarcasm* records, 421 *soft-claim-fake* records and 166 *soft-claim-not-fake* records.

4.4.3 Pilot Annotation Data

As part of the annotation pilot, we prepared a collection of 1,000 tweets referring to various URL addresses. The method of collecting data is described in Section 3. Each tweet was annotated by 4 independent annotators with one of the 6 labels described in Section 3.3 (*hard-claim-fake*, *hard-claim-not-fake*, *no-claim*, *sarcasm*, *soft-claim-fake*, *soft-claim-not-fake*). Then, the annotations for each tweet were aggregated according to the following algorithm:

1. If all annotators have chosen the same label, assign that label.
2. Otherwise:
   - if any annotators have chosen the label **-claim-fake* and no annotators have chosen the label **-claim-not-fake*, assign the label *soft-claim-fake*,
   - if any annotators have chosen the label **-claim-not-fake* and no annotators have chosen the label **-claim-fake*, assign the label *soft-claim-not-fake*.
3. In other cases, assign the label *no-claim*.

This way, we obtained 1,000 records.

4.4.4 Data from Annotation Tasks on the Doccano Platform

The method of collecting data for annotation tasks is described in Section 3. Annotation in these tasks consisted of selecting one of the 6 labels described in Section 3.3 (*hard-claim-fake*, *hard-claim-not-fake*, *no-claim*, *sarcasm*, *soft-claim-fake*, *soft-claim-not-fake*) based on the text of the tweet and the content of the website to which the tweet concerned. Then, the annotations for each tweet were aggregated according to the same algorithm as in the case of the pilot annotation. This way, we obtained 4,082 records.

5 Anonymization/Privatization

Privatization is an important step in the process of constructing any language resource that combines news and social media text. It requires thoughtful planning with regard to the categories of personal identifiable data that should or should not be anonymized. On the one hand, the names of public figures and coarse-grained descriptions of geographical locations of events are not considered private. Thus, they should not be anonymized in the corpus. On the other hand, the names of private citizens, their home addresses or any other personal identifiable information should be removed. To solve this problem, we developed a privatization tool that consists of three modules: 1) named entity recognizer, 2) alphanumeric expression classifier, and 3) privacy checker.

The named entity recognizer follows Transformer architecture (Vaswani et al., 2017) and utilizes a pre-trained language model (Devlin et al., 2019). It is based on the HerBERT model[19](Mroczkowski et al., 2021) with a token classification head attached. The alphanumeric expression classifier is responsible for detecting potentially private phrases with strict definitions that can be described using regular expressions. The categories of expressions identified by this module are summarised in Table 2. The privacy checker considers all expressions detected by the named entity recognizer and the alphanumeric expression classifier to be private by default. It makes public only the names that appear in an index of public figures built on the basis of DBpedia (Lehmann et al., 2015) entries that belong to the <https://dbpedia.org/ontology/Person> class in the DBpedia ontology, denoted by the Polish or English language code.

Category	**Description**
url	uniform resource locator
email	e-mail address
cardnumber	credit/debit card number
zipcode	postal code
username	username in social media
nip	tax ID
passport	passport number
idcard	identity card number
crypto	crypto wallet address
macaddr	MAC address
accountnumber	bank account number
address	physical address
phone	phone number

Table 2: Categories of data detected by the alphanumeric expression classifier.

Set	**fake-or-not**	**fake-they-say**
Training set	6,902	4,040
Validation set	1,256	316
Test set	3,202	726
Total	11,360	5,082

Table 3: The POLygraph dataset summary

6 Dataset Summary and Discussion

The POLygraph Polish fake news dataset consists of two parts: *fake-or-not* and *fake-they-say*, which are detailed in Sections 4.3 and 4.4.Together, they form a new dataset for detecting fake news in Polish. Unlike existing datasets, this dataset is not solely or predominantly based on a binary true-false classification but draws on various approaches proposed in source literature. The overview of the dataset is shown in Table 3.

This approach results in collecting a range of data typically utilized in news-content-based, knowledge-based, and user-information-based fake news detection methods. Although the POLygraph dataset has not yet been used in real-world scenarios, it was developed for a project aimed at verifying information sources and detecting fake news. Further exploration of the collected data by an interdisciplinary team of researchers will foster fake news detection and provide institutions and scholars with a more comprehensive range of data than previous fake news datasets. The envisioned use case involves building tools that detect false information and mark such information in search engines, potentially tested by monitoring social media messages over some time.

Additionally, adapting the POLygraph dataset for other languages should not pose a significant problem. The dataset itself is based on solutions proposed for other languages, often very different from one another, such as English and Japanese. This universality strengthens the argument that the

[19] https://huggingface.co/allegro/herbert-base-cased

core concept can be applied across various languages and cultural settings. Some proposed solutions might require modifications depending on the specific language, but the core strength remains – the applicability across diverse contexts. The presented annotation scheme will hopefully serve as a stimulus for implementing an analogous detection model for other languages.

7 Acknowledgments

This research was conducted as part of a project "From Urban Legend to Fake News. A Global Detector of Contemporary Falsehoods", funded by the Polish National Centre for Research and Development, grant number INFOSTRATEG-I/0045/2021.

8 Limitations

This study acknowledges the inherent challenges in building a comprehensive fake news detection system. The dataset, while extensive, might not capture every form of misinformation online, limiting the generalizability of the findings. Additionally, the use of human annotation introduces subjectivity, as annotators may have differing definitions of what "fake news" is. Including subjective and ambiguous questions to explore fake news in depth can lead to disagreements, especially in borderline cases. However, perfect agreement is not expected in such nuanced tasks – similar projects report agreement scores around $0.3 \sim 0.4$ (Gorrell et al., 2019), which is deemed acceptable here.

The complexity of fake news detection is reflected in the multidimensional annotation scheme employed. This paper focuses on data collection and annotation, with the evaluation of the dataset's efficacy in machine learning tasks planned for a future stage. Similarly, the description and evaluation of a potential fake news recognition tool using this dataset are beyond the scope of this article.

Furthermore, the study primarily focuses on the Polish language, limiting its direct applicability to other languages and cultures. The ever-evolving nature of fake news tactics also necessitates continuous updates to the dataset and any future detection tool to maintain effectiveness.

Despite these limitations, this study offers valuable insights into fake news detection and lays a robust foundation for future research in this area.

9 Ethics Statement

The human annotators were recruited from a pool of student volunteers who expressed interest in participating in the project. They were informed about the project's purpose, methods, and expected outcomes, and they gave their consent before starting the annotation task. They were given clear instructions and guidelines for the annotation task and received feedback and support whenever needed. They were free to withdraw from the project at any time without any negative consequences. The annotators did not receive payment for participating in the project, as they agreed to volunteer their time and effort for scientific research. The authors have the right to use the data presented in the paper, and they ensured that the data was anonymized and privatized to protect the privacy and confidentiality of the individuals and entities involved.

References

Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7.

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2017. Detection of online fake news using n-gram analysis and machine learning techniques. In *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, pages 127–138, Cham. Springer International Publishing.

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2018. Detecting opinion spams and fake news using text classification. *Security and Privacy*, 1(1):e9.

Bimal Bhattarai, Ole-Christoffer Granmo, and Lei Jiao. 2022. Explainable tsetlin machine framework for fake news detection with credibility score assessment. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4894–4903, Marseille, France. European Language Resources Association.

Verena Blaschke, Maxim Korniyenko, and Sam Tureski. 2020. CyberWallE at SemEval-2020 task 11: An analysis of feature engineering for ensemble models for propaganda detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1469–1480, Barcelona (online). International Committee for Computational Linguistics.

Luís Borges, Bruno Martins, and Pável Calado. 2019. Combining similarity features and deep representation learning for stance detection in the context of checking fake news. *J. Data and Information Quality*, 11(3).

Tobias Bornheim, Niklas Grieger, and Stephan Bialonski. 2021. FHAC at GermEval 2021: Identifying German toxic, engaging, and fact-claiming comments with ensemble learning. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 105–111, Duesseldorf, Germany. Association for Computational Linguistics.

Santiago Castro, Matías Cubero, Diego Garat, and Guillermo Moncecchi. 2016. Is this a joke? detecting humor in spanish tweets. In *Advances in Artificial Intelligence - IBERAMIA 2016*, pages 139–150, Cham. Springer International Publishing.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

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.

Julian Eisenschlos, Syrine Krichene, and Thomas Müller. 2020. Understanding tables with intermediate pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 281–296, Online. Association for Computational Linguistics.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Filip Graliński. 2012. *Znikająca nerka: mały leksykon współczesnych legend miejskich*. Media Rodzina.

Filip Graliński, Rafał Jaworski, Łukasz Borchmann, and Piotr Wierzchoń. 2016. Gonito.net – open platform for research competition, cooperation and reproducibility. In António Branco, Nicoletta Calzolari, and Khalid Choukri, editors, *Proceedings of the 4REAL Workshop: Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language*, pages 13–20. European Language Resources Association (ELRA).

Filip Graliński, Anna Wróblewska, Tomasz Stanisławek, Kamil Grabowski, and Tomasz Górecki. 2019. GEval: Tool for debugging NLP datasets and models. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 254–262, Florence, Italy. Association for Computational Linguistics.

Xiaobao Guo, Nithish Muthuchamy Selvaraj, Zitong Yu, Adams Wai-Kin Kong, Bingquan Shen, and Alex Kot. 2023. Audio-visual deception detection: Dolos dataset and parameter-efficient crossmodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22135–22145.

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Radu Tudor Ionescu and Adrian Gabriel Chifu. 2021. Fresada: A french satire data set for cross-domain satire detection. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Hayate Iso, Chao Qiao, and Hang Li. 2020. Fact-based Text Editing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 171–182, Online. Association for Computational Linguistics.

Kornraphop Kawintiranon and Lisa Singh. 2021. Knowledge enhanced masked language model for stance detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4725–4735, Online. Association for Computational Linguistics.

Elena Kochkina, Maria Liakata, and Isabelle Augenstein. 2017. Turing at SemEval-2017 task 8: Sequential approach to rumour stance classification with branch-LSTM. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 475–480, Vancouver, Canada. Association for Computational Linguistics.

Anna Kolos, Inez Okulska, Kinga Głąbińska, Agnieszka Karlinska, Emilia Wisnios, Paweł Ellerik, and Andrzej Prałat. 2024. BAN-PL: A Polish dataset of banned harmful and offensive content from wykop.pl web service. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2107–2118, Torino, Italia. ELRA and ICCL.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. DBpedia - A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Lily Li, Or Levi, Pedram Hosseini, and David Broniatowski. 2020a. A multi-modal method for satire detection using textual and visual cues. In *Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 33–38, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2020b. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. 2021. HerBERT: Efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10, Kiyv, Ukraine. Association for Computational Linguistics.

Taichi Murayama, Shohei Hisada, Makoto Uehara, Shoko Wakamiya, and Eiji Aramaki. 2022. Annotation-scheme reconstruction for "fake news" and Japanese fake news dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7226–7234, Marseille, France. European Language Resources Association.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Constantin Orăsan. 2018. Aggressive language identification using word embeddings and sentiment features. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 113–119, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Nick Pawlowski, Daniel C. Castro, and Ben Glocker. 2020. Deep structural causal models for tractable counterfactual inference. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Gregory Plumb, Jonathan Terhorst, Sriram Sankararaman, and Ameet Talwalkar. 2020. Explaining groups of points in low-dimensional representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7762–7771. PMLR.

Jason Poulos and Shuxi Zeng. 2021. RNN-Based Counterfactual Prediction, With an Application to Homestead Policy and Public Schooling. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 70(4):1124–1139.

Yanou Ramon, David Martens, Foster Provost, and Theodoros Evgeniou. 2020. A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: Sedc, lime-c and shap-c. *Advances in Data Analysis and Classification*, 14(4):801–819.

Julian Risch and Ralf Krestel. 2018. Aggression identification using deep learning and data augmentation. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 150–158, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3419–3425, Hong Kong, China. Association for Computational Linguistics.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36.

James Thorne and Andreas Vlachos. 2021. Evidence-based factual error correction. 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 3298–3309, Online. Association for Computational Linguistics.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

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.

Yuhang Wang, Li Wang, Yanjie Yang, and Tao Lian. 2021. Semseq4fd: Integrating global semantic relationship and local sequential order to enhance text representation for fake news detection. *Expert Systems with Applications*, 166:1–12.

Orion Weller and Kevin Seppi. 2019. Humor detection: A transformer gets the last laugh. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3621–3625, Hong Kong, China. Association for Computational Linguistics.

Xiaoyu Yang, Stephen Obadinma, Huasha Zhao, Qiong Zhang, Stan Matwin, and Xiaodan Zhu. 2020. SemEval-2020 task 5: Counterfactual recognition. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 322–335, Barcelona (online). International Committee for Computational Linguistics.

Xichen Zhang and Ali A. Ghorbani. 2020. An overview of online fake news: Characterization, detection, and discussion. *Information Processing and Management: an International Journal*, 57(2):1–26.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5):1–40.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv.*, 51(2).

A Appendix: Annotation Questions

Q1: Specify the type of text.

a) article on a news website
b) social media post
c) blog post
d) other

Q2: Define the subject matter of the text.

a) politics
b) society
c) medicine
d) military
e) economy
f) entertainment
g) education
h) science and technology
i) tourism
j) culture
k) sports
l) business
m) crime
n) safety
o) religion
p) other

Q3: What is your attitude to the text?

a) I agree with the text.
b) I do not agree with the text.
c) I have a neutral attitude to the text.

Q4: What emotions does the text evoke in you?

a) positive
b) negative
c) The text does not evoke emotions in me.

Q5: What content dominates in the text?

a) facts
b) opinions
c) both

Q6: Is the text persuasive?

a) yes
b) no
c) I don't know

Q7: What do you think is the purpose of the news?

a) information - the text is purely informative, it reports and describes events
b) disinformation - the author deliberately provides false information in order to obtain some benefits (e.g. political or financial)
c) propaganda - the text is persuasive and affects the emotions, attitudes, opinions and/or actions of the target audience for ideological, religious and other purposes

d) partisan promotion of political views - the text presents information in a biased way from the perspective of a specific political party or political ideology

e) entertainment (satire / parody) - the purpose of the text is to provide the target with entertainment and / or criticism of individuals or groups

f) other

Q8: Who do you think is the potential target of the news?

a) recipient of general news from news websites

b) recipient of entertainment

c) supporter of a specific political party

d) supporter of a specific socio-political ideology

Q9: Does the author/disseminator believe that the news they are writing about is true?

a) yes, the author openly expresses the belief that they agree with what they are disseminating

b) yes; however, the author expresses doubts about the veracity of the news

c) no, the author openly denies the veracity of the news

d) no comments are made by the author

Q10: Does the author refer to the sources of the cited information?

a) yes

b) no

c) sometimes / not always

Q11: What narrative style is the main basis of the news?

a) conflict (often specific to political events, centered around disagreement, division, difference or rivalry)

b) responsibility (assigning responsibility for the cause/effect of the presented problem to specific persons/institutions etc.)

c) morality (related to the moralizing tendencies of the media; it most often refers to condemnation or other forms of moral evaluation of the presented events)

d) human story (personalization which introduces emotional elements, the main character is most often the victim of a tragic event or crisis; greater importance is attached to the individual affected by the event than its global consequences)

e) consequences (related to a broader context and impact on various areas of social life)

Q12: In your opinion, does the text contain false information?

a) yes

b) no

c) not subject to assessment (the text contains only the author's opinion)

Q13: What kind of false information is contained in the text?

a) fake news - false information has been included in the article intentionally and it is possible to verify it (without referring to external sources!)

b) rumor - the author refers to unconfirmed information (e.g. rumors)

c) satire - the author cites false information that is humorous, ironic, mocking; it is not intended to mislead the reader

d) clickbait - the title attracts attention, but does not reflect the content of the news

Q14: Where is the false information located in the text?

a) in the title/headline

b) in one fragment

c) false information is repeated in several fragments of the text

d) in the image

e) the whole text is false

Q15: How much of the text must be read in order to realize that it contains false information?

a) headline / title

b) the title and part of the text

c) the entire text

Q16: If the news contains false information, do you think the author of the text knows that they are disseminating false information?

a) They know it.

b) They probably know it.

c) They don't know it.

d) They definitely don't know it.

Q17: Have you come across the false information contained in the text before?

a) Yes.

b) No.

Q18: How socially harmful is the false information contained in the text?

a) 0 (harmless)

b) 1 (slight harm, e.g. lack of understanding of certain events)

c) 2 (moderately harmless, e.g. causing confusion and anxiety)

d) 3 (moderately harmful, e.g. leading to conspiracy theories)

e) 4 (relatively harmful, e.g. damage to the reputation of people and institutions, prejudice against a nation, race etc.)

f) 5 (very harmful, e.g. health and life hazard)

Q19: What kind of threat may be posed by the false information?

a) lack of understanding of political and social events

b) damage to the reputation of persons and institutions, undermining trust in persons and institutions

c) prejudice against nation, race, state

d) confusion and fear of society

e) the emergence of conspiracy theories

f) risk to health and life

g) none