GerDISDETECT: A German Multilabel Dataset for Disinformation Detection

Mina Schütz[1,3], **Daniela Pisoiu**[2], **Daria Liakhovets**[1], **Alexander Schindler**[1],
Melanie Siegel[3]
AIT Austrian Institute of Technology GmbH[1], SCENOR[2],
Darmstadt University of Applied Sciences[3]
Vienna Austria[1,2], Darmstadt Germany[3], Vienna Austria
daniela.pisoiu@scenor.at, melanie.siegel@h-da.de,
{mina.schuetz, daria.liakhovets, alexander.schindler}@ait.ac.at

Abstract

Disinformation has become increasingly relevant in recent years both as a political issue and as object of research. Datasets for training machine learning models, especially for other languages than English, are sparse and the creation costly. Annotated datasets often have only binary or multiclass labels, which provide little information about the grounds and system of such classifications. We propose a novel textual dataset GerDISDETECT for German disinformation. To provide comprehensive analytical insights, a fine-grained taxonomy guided annotation scheme is required. The goal of this dataset, instead of providing a direct assessment regarding true or false, is to provide wide-ranging semantic descriptors that allow for complex interpretation as well as inferred decision-making regarding information and trustworthiness of potentially critical articles. This allows this dataset to be also used for other tasks. The dataset was collected in the first three months of 2022 and contains 39 multilabel classes with 5 top-level categories for a total of 1,890 articles: General View (3 labels), Offensive Language (11 labels), Reporting Style (15 labels), Writing Style (6 labels), and Extremism (4 labels). As a baseline, we further pre-trained a multilingual XLM-R model on around 200,000 unlabeled news articles and fine-tuned it for each category.

Keywords: disinformation, dataset, NLP, German, multilabel

1. Introduction

Disinformation - also called ”fake news” - has become a constant of political life, as various types of actors propagate disinformation via online outlets or even as part of political campaigns. Fake news is often used as an umbrella term for disinformation and misinformation. The former being intentionally deceptive and the latter describing unintentional spread of false information (Sharma et al., 2019). From a societal point of view, a main problem is the increasing difficulty to distinguish between false and real statements (Figueira and Oliveira, 2017). From an academic and policy point of view, it has become necessary to be able to collect and analyze large quantities of data. Therefore, research started to focus on the automatic detection of disinformation in text.

The automatic detection of text can be done through the usage of machine learning (ML) models that are trained on a labeled dataset. Their output is a prediction of a given label with a certain confidence. To achieve this, one method is Natural Language Processing (NLP). Trained classification models are evaluated on their prediction of unseen data. At the same time, their performance depends on the quality of the annotated labels in the training dataset. With such systems users can have a first estimation to what could be a false news piece.

Most research on NLP has been conducted for the English language only. Currently, to our knowledge, there are only five datasets available (Köhler et al., 2022; Vogel and Jiang, 2019; Ströckl, 2020; Shahi and Nandini, 2020; Mattern et al., 2021) which provide labeled data in German. Furthermore, these datasets provide only binary (i.e. true or false) or multiclass (i.e. true, false, partly true, other) labels. Therefore, we propose a novel dataset **GerDISDETECT** for **Ger**man **DIS**information **DETECT**ion, which will be available publicly for the research community.

2. Dataset Overview & Motivation

The dataset contains 1,890 news articles from 75 distinct websites publishing in German. They were manually retrieved between January 2022 until March 2022 from media outlets as well as blogs and citizen journalist websites. In total, we created an annotation scheme with 5 top categories and 39 detailed multilabels: **General View**, **Writing Style**, **Reporting Style**, **Hate Speech**, and **Extremism**.

Multilabel defines that for each of the five categories in this dataset, the labels are not mutually exclusive. For example, the category *Reporting Style* has 15 labels. The annotator could flag any of the 15 labels at the same time. This leads to combinations of annotations, such as ”Clickbait”

LREC-COLING 2024, pages 7683–7695
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

together with ”False Science Reporting” and ”Propaganda”. This helps end-users to understand why an article is flagged as critical content, increases the accuracy of detection by the inclusion of several criteria, and simultaneously separates this approach from binary definitions. This more complex classification furthermore contributes to increased literacy in end-users, as simple definitions of ”fake” vs. ”real” do not portray real world circumstances, nor do they convey intrinsic credibility.

Coarse-grained labels give no indication to the end-user why an article could contain false information. That results in less trust in the prediction for the end-user, even though disinformation is a complex and oftentimes also a safety-critical domain. Particularly, disinformation campaigns can have an impact on the real world.

2.1. Disinformation

A more commonly known term for disinformation is ”fake news”. Fake news is usually defined as ”[...] news articles that are intentionally and verifiably false, and could mislead readers” (Allcott and Gentzkow, 2017). While seemingly a simple definition, its operationalization can be less straightforward and highly complex. First, whether or not something is labeled as true or false is highly dependent on the available facts and evidence at any given time, and which can shift at any given moment. Another issue with fake news is that there is no universal and constantly updated knowledge base against which to check the accuracy of information. Finally, classifying news is not always binary in nature, since articles can be generally true yet still contain some false information or vice versa (Fu et al., 2022; Oshikawa et al., 2020). This is underlined by the concept of hoaxes, which are often used to hide the truth behind a news story in a large scale (Sharma et al., 2019).

The term disinformation is itself usually described as fake articles that are intentionally spread by the author (Sharma et al., 2019). The term misinformation on the other hand describes the propagation of false information spread unintentionally (Sharma et al., 2019). In regard to these definitions, we define disinformation campaigns as intentionally spread activities to deceive consumers on a large scale. This relates to other research fields, such as propaganda detection. Propaganda is used to mislead the public and to propagate a biased view by the author. Often, this type of disinformation plays with the emotions of news consumers and can be written - for example - by governments, extremists or corporations (Khan et al., 2019). Extremists paint a distorted image of reality, usually involving the unjust oppression of the group they represent by some identifiable enemies. The latter are furthermore demonized and/or dehumanized in order to facilitate the acceptability of violent means to be used against them. Importantly, the success of extremist propaganda depends in part on the instrumentalization of real facts within the overall narrative (see the concept of empirical credibility), thus again underlying the complexity of disinformation beyond binary categories (Pisoiu and Hain, 2017).

Therefore, this work takes a broader view on disinformation than simply a question of true vs. false, and considers the contributions of adjacent fields of research such as hate speech and extremism. We have therefore defined offensive language and extremism detection to be closely related to the spread of disinformation on online platforms and media outlets. We intentionally not refer in this case to it as hate speech, since this is often found in short comments on social media and used for conversational analysis (Demus et al., 2022). For a broader view of news articles it is referred to as offensive language in general.

2.2. Related Work

There are several datasets openly available for research in the context of the fake news, disinformation, and misinformation. Most of the published datasets contain English news articles or social media content. Only a few are available for the low-resource language German. In Table 1 we give an overview over related datasets in the field in German and English, some being multilingual. This list is created to the best of our knowledge regarding published datasets.

For fake news detection in English many datasets have been presented with focus not only fake news, but credibility detection, fact-checking and fact-extraction, as well as rumour, hoax and stance detection. Also the datasets vary heavily in the content that was extracted (i.e. different social media platforms, news articles, and fact-checking websites). Besides that, for the German language, a few specialized datasets exist: **GermanFakeNC** (Vogel and Jiang, 2019) contains 490 false and 4,500 real articles with the labels: *true with up to 25% false*, *half true*, *75% fabricated*, *100% untrue* and *lacked a claim*. Recently, another binary classification dataset in German was published on Kaggle (Ströckl, 2020). The **Fake News Dataset German** contains approximately 63,000 fake and non-fake news articles from the fields of economics and sports. A cross-lingual dataset **CT-FAN** (Köhler et al., 2022) was first introduced by the **CheckThat! 2022** shared task with 4 more fine-granular labels (*true*, *false*, *partially false*, *other*) and around 1,600 instances in English and German. However, the German data is sparse with 586 articles (Shahi et al., 2021; Nakov

Dataset	Language	Type	Content Size	Classes
CREDBANK (Mitra and Gilbert, 2015)	EN	Credibility	1,300	5
LIAR (Wang, 2017)	EN	Fake News	12,836	6
Some-like-it-hoax (Tacchini et al., 2017)	EN	Hoax	15,500	2
ISOT (Ahmed et al., 2017, 2018)	EN	Fake News	25,200	2
FEVER (Thorne et al., 2018)	EN	Fact Checking	185,455	3
BuzzFace (Santia and Williams, 2018)	EN	Fake News	2,282	4
FakeNewsNet (Shu et al., 2018)	EN	Fake News	22,589	2
Fakeddit (Nakamura et al., 2020)	EN	Fake News	1,063,106	2/3/6
NELA-GT-2019 (Gruppi et al., 2020)	EN	Fake News	1,120,000	3
Covid-19 (Patwa et al., 2020)	EN	Fake News	10,700	2
PHEME (Kochkina et al., 2018)	EN	Rumour	330	2
MuMiN (Nielsen and McConville, 2022)	+	Misinformation	21,000,000	nodes
Monant (Srba et al., 2022)	+	Facts & Stance	317,000	2+
GermanFakeNC (Vogel and Jiang, 2019)	GER	Fake News	4,990	4
KaggleGermanFake (Ströckl, 2020)	GER	Fake News	ca. 63,000	2
FANG-COVID (Mattern et al., 2021)	GER	Fake News	41,242	2
FakeCovid (Shahi and Nandini, 2020)	+	Fake News	5,182	11
CT-FAN (Köhler et al., 2022)	+	Fake News	ca. 1,600	4
GerDISDETECT	**GER**	**Disinformation**	**1,890**	**39**

Table 1: Related datasets in the disinformation research field for English (EN), German (GER), and multilingual (+). The content size is related to the text instances. The classes are shown in either binary (2), multiclass (3 or more) and multilabel for GerDISDETECT.

et al., 2022; Köhler et al., 2022; Shahi and Nandini, 2020). The largest German dataset available is **FANG-COVID** (Mattern et al., 2021), which contains real articles from three media outlets and ten publishers who are known to post false information. On top of that the authors added social media context information from Twitter for each article. In contrast to the presented datasets, the proposed **GerDISDETECT** dataset has the most amount of labels (39) and focused on a more exploratory approach instead of annotating the texts regarding their veracity in a binary or multi-class task.

3. Methodology

In this section we firstly describe how we collected the data and secondly give an overview over the annotation scheme and annotation process.

3.1. Data Collection

The dataset contains 1,890 news articles that were manually extracted. In contrast to other datasets (e.g. FakeNewsNet (Shu et al., 2018)) we decided against an automated extraction of text on websites. Many datasets – that were created in such way – contain large amounts of source code or error messages from websites. We started with a list of German websites that we suspect to post articles that can fall under the category of dis- or misinformation. Usually, not every website contains only false information (Fu et al., 2022). Therefore, we manually went through every website and looked for articles with similar topics (i.e Covid and the Russian-Ukranian war) but did not exclude other articles as well based on the availability.

In total, we collected articles from 75 different websites that publish in German. We did not differentiate between websites from i.e. Germany, Austria or Switzerland. Often, the imprint showed different countries than the aforementioned and could not be clearly assigned to a specific country. To have a balanced dataset, we did separate the articles based on expert opinions on which are rather reliable or questionable websites. Well-researched websites are for example known media outlets. Rather unreliable websites can be private blogs or websites with no clear journalistic background - both does not exclude the other. Similarly, the FANG-COVID authors Mattern et al. (2021) classified their articles based on trusted news publishers and a list of publishers that are confirmed to commonly publish false information regarding COVID . In comparison, we do not clearly state that an article is fake or real based on the publisher. Throughout this paper, we will use the terms ”questionable” and ”reliable” to refer to those concepts, which we added to the final dataset. The annotators were not aware of those categories during the labeling process.

In total we collected 965 articles that can be considered as reliable information (51.06 %) and 925 articles (48.94%) that could possibly contain questionable content. Split by websites, it would mean that 65.33% of sources are annotated as questionable (49 websites in total) and 34.67% are poten-

tially reliable (26 websites in total). We also extracted the authors and publication dates, which will not be added to the dataset. However, this was done automatically by the annotation tool, in contrast to the extraction of the content. The annotation tool was provided by a project partner and is not publicly available. The tool is a web-based framework, where each annotator got the articles assigned to label. The annotated articles were extracted as a CSV file. We found a total of 617 different authors without cleaning. After manually going through the data (i.e. deleting "Author not mentioned" and extractions of website names) we ended up with 594 distinct author names or author combinations. In total we had 856 articles without any author (519 articles categorized as reliable, 337 articles categorized as questionable). In comparison, the 1,034 articles that had an author mentioned have 446 articles categorized as reliable and 588 as questionable.

The highest count of published articles was on March 16, 2022 unrelated to any event or specific topic that day. We believe this is due to a larger collection of articles during that week. 444 articles were published before that date, starting in the 2000s. 21 articles were published before January 1st, 2015. The accessible dataset contains the following information:

- *_id:* the identifier for each article.
- *Text_content:* the body text of the article.
- *Text_title:* the title of the article.
- *Text_subtitle:* the subtitle of the article.
- *Multilabels:* a column for every label in the dataset. The top-level category is abbreviated in front of each label name. Values 0 and 1 indicate the presence of the label in the article.
- *Binary_cat:* binary class showing, whether an article is potentially reliable or questionable.

The binary classification refers to the assessment of the articles based on their source, not labeled by the annotators.

3.2. Annotation Scheme & Process

To create a sustainable dataset for a standard classification approach, the data has to be labeled. Therefore, we created a comprehensive annotation scheme that will be described in detail in this section. The concepts we found as most important, based on former research, we split into five top-level categories: 1) **General View**, 2) **Writing Style**, 3) **Reporting Style**, 4) **Offensive Language**, and 5) **Extremism**. Each of the top-level categories has multilabel classes on document level.

Label	Total	Rel.	Quest.
Author Mentioned	1,161		
Publication Date	1,763		
Sources	462		

Table 2: Category 1 - General View.

Each article was labeled by one trained person and each annotation checked and confirmed by two additional people. If there was a disagreement between the original annotator and the final person to check the article, this person adjusted the annotation. But even such a confirmation loop is prone to errors. Due to time constraints - annotating 39 highly challenging labels and the length of an article - we could not have more people annotate interchangeably. The 11 annotators who labeled the datasets were trained by experts - with backgrounds in disinformation, computer linguistics, data science, and social sciences - on each category and labels extensively. After a few test rounds of labeling the data, we started with 200 articles each week over a time period of around three months. Furthermore, "Rel." refers to "Reliable" and "Quest." refers to "Questionable".

3.2.1. General View

The first category "General View" (Table 2) has three labels that could be annotated: *Author Mentioned*, *Publication Date Mentioned*, *Sources Mentioned*. Those can - but not necessarily - give first clues about the credibility of an article, since there are two types of deceiving websites: 1) either intentionally deceiving, trying to imitate traditional websites; 2) or partisan websites that usually contain a mixture of false and real information, and opinions (Allcott and Gentzkow, 2017).

3.2.2. Writing Style

The "Writing Style" (Table 3) describes, *how* an article is written. This can of course change depending on the media outlet and each author. However, in many studies (Shu et al., 2017; Zhou and Zafarani, 2019; Conroy et al., 2015; Zhou and Zafarani, 2018; Afroz et al., 2012) it is found that the writing style is important to distinguish between possible disinformation and well-researched articles. Examples for a *Populistic* writing style are texts with words written in caps or a series of exclamation marks (i.e. "this CANNOT be done!!!") (Molina et al., 2021). *Polarization* on the other hand describes to which extent opinions on an issue are opposed. Another writing style is *Exaggeration*. If an author writes exaggerated, the topic in the article is shown in usually the opposite way (i.e. more positive, negative) (Martino et al., 2019).

Label	Total	Rel.	Quest.
Exaggerated	323	15	308
Humoristic	31	4	27
Polarized	312	21	291
Populistic	132	3	129
Subjective	634	130	504
Unprofessional	57	5	52

Table 3: Category 2 - Writing Style.

Humoristic style could hint to satire or similar (Jr. et al., 2018). *Subjective* articles usually rather depict the opinion of the author - i.e. opinion journalism and opinion pieces - instead neutral reporting (Rashkin et al., 2017). This does not mean an article lacks of journalistic proficiency, in contrary to *Unprofessional Writing Style*. The latter is used as an umbrella term in this context for bad readability, grammar mistakes, unprofessional wording, typos and the general style of the article.

3.2.3. Reporting Style

The "Reporting Style" (Table 4) is the most valuable category in the annotation scheme. *Claims Stated* describes that a statement was made or asserted by the author (Risch et al., 2021). A stated claim does not necessarily mean that also the evidence for the given claim was provided. Therefore, there is an additional label *Lack of Evidence* that is exactly used for the case that an author is missing some underlying information that is needed to understand the larger context (Molina et al., 2021). However, not every claim necessarily needs evidence, such as statements connected to general knowledge. If scientific evidence is quoted but somehow taken out of context or misrepresented, we define it as *False Science Reporting*. Similarly, if an article only presents and favors one opinion of a topic or argument, it can underline a biased view. This is labeled with *One-Sided Reporting* (Molina et al., 2021). This can also relate to questioning the credibility - for example of media platforms (Reis et al., 2019; Martino et al., 2019) - of someone or something, which can stir up *Doubt* in the end-user (Martino et al., 2019).

Another concept is *Logic Flaws* in the structure of an article (Molina et al., 2021) or false contextual information (*False Context* (Sharma et al., 2019)). To push end-users to click on an article, often *Clickbait* headlines are used. Such headlines usually create a knowledge gap for the reader. Therefore, the reader opens the article but does not receive the promised answer to the given question in the headline (Chen et al., 2015). However, if the body content of an article is written to provoke or intend excitement (i.e. tabloids, gossip), it can be labeled as *Sensationalism* (Chen et al., 2015). Another label is *Repetition* which states that the same statements or thoughts are repeatedly used in the article (Horne and Adali, 2017; Martino et al., 2019).

Label	Total	Rel.	Quest.
Claims Stated	516	17	499
Clickbait	87	7	80
Conspiracy	286	1	285
Doubt	496	8	488
False Context	177	0	177
False Science Reporting	173	1	172
Lack of Evidence	282	1	281
Logic Flaws	91	0	91
Misinformation	364	1	363
Misleading	453	2	451
One-Sided Reporting	439	18	421
Political Agenda	339	7	332
Propaganda	150	3	147
Repetition	23	0	23
Sensationalism	257	7	250

Table 4: Category 3 - Reporting Style

The label *Political Agenda* describes whether an article pushes political programs or motives (Jr. et al., 2018). This is also highly related to *Propaganda*, which is usually used to help a certain cause. This sometimes can be misused by political leaders or governments that want to deceive the reader (Khan et al., 2019). *Propaganda* itself is more or less systematic to manipulate other people's beliefs, attitudes, or actions (Jr. et al., 2018). Another strong label related to the whole concept of disinformation is *Misleading*, which describes the action of deceiving the reader into information or statements that the author knows are not true (Sharma et al., 2019). If the false information is unintentionally spread by using low-quality journalism or unverified sources, it is called *Misinformation* (Bara et al., 2019). Lastly, *Conspiracy* theories are attempts to explain the ultimate causes of significant social and political events and circumstances with claims of secret plots by two or more powerful actors (Douglas et al., 2019).

3.2.4. Offensive Language

Offensive or *Abusive Language* (Table 5) is a very specific phenomenon in the world of online-articles and social media content. Some studies have especially made it clear that offensive language is not essentially only against groups, but can also target individuals and their specific characteristics (Ayo et al., 2021). Therefore, we use the concept of offensive language to define articles that might bother people based on certain characteristics. Most characteristics are derived from the DeTox project on abusive and hateful language (Demus et al., 2022) and expanded or adjusted. In

Label	Total	Rel.	Quest.
Culture	62	7	55
Ethnicity	27	0	27
Anti-Semitism	17	1	16
Anti-LGBTQIA+	91	13	78
Nationality	25	0	25
Physical, Psychological	45	1	44
Political View	460	9	451
Race	19	1	18
Religion	31	3	28
Sex	43	9	34
Social Status	49	1	48

Table 5: Category 4 - Offensive Language.

this dataset, the labels Anti-LGBTQIA+ and Anti-Semitism were added. The former is an abbreviation for "lesbian, gay, bisexual, transgender, queer/questioning (one's sexual or gender identity), intersex, and asexual/aromantic/agender" by Merriam-Webster[1]. Additionally, we introduced *Ethnicity* as a more broad term for all differences that are not directly fitting into more narrow-terms such as race, culture, or religion.

3.2.5. Extremism

The category "Extremism" (Table 6) assesses the articles for leaning into ideas of extreme political views in one of four sub-categories. This includes several ideological variations in the field of *Right-Wing* political views in forms, such as the New Right, sub-cultural forms, white supremacy, and accelerationism (Young and Boucher, 2022). *Left-Wing*: this category describes extreme left-wing political views in forms of anarchism and communism (Jungkunz, 2019). *Religious*: such as Islamism, Christian fundamentalism and Hindu fundamentalism. *Single-Issue*: this describes movements with an overarching goal that are willing to resort to extraordinary means to achieve it (Ackerman and Koulouganes, 2019) (e.g. using extreme means for the purposes of environmental protection, animal rights, anti-abortion, anti-vaccination). While there are more variations of extremism within these sub-categories, we kept this particular category rather simple in order to match the level of knowledge of the annotators.

4. Data Exploration

In this section, we will show a more detailed overview of the data. Firstly, we will go more into detail about the first top-level category: General View. As described in Table 2, the publication

[1] https://www.merriam-webster.com/dictionary/LGBTQIA

Label	Total	Rel.	Quest.
Left-Wing	15	6	9
Religious	11	2	9
Right-Wing	219	8	211
Single-Issue	306	3	303

Table 6: Category 5 - Extremism.

Subset	Mean	Median	Min.	Max.
All	742.251	605.5	28	2,464
Rel.	624.356	504.0	28	2,410
Quest.	865.245	726.0	29	2,464

Table 7: Length of the body text in comparison to the complete data, and possible reliable and questionable articles. Length shown in number of words. Min. is short for minimum; Max. is short for maximum.

date is labeled 1,856 times out of 1,890 articles in the dataset. This shows that most articles - reliable or not - have a publication date mentioned. We conclude that this is not helpful to distinguish between credible or unreliable news sources. Similarly, most websites have an author mentioned, as this is also underlined by the automatic extraction by the annotation tool. However, only 489 articles had additional sources mentioned: 308 of those articles are tagged reliable and 181 articles are tagged questionable in the dataset.

Secondly, we have a closer look at the length distributions to find any differences between reliable and questionable articles in the dataset (Table 7). In the case of the proposed **GerDISDETECT** dataset, there is significant difference in length. The longest questionable article is three times as long as the longest reliable article. This leads to a more comprehensive analysis of the distributions to see if the minimum and maximum values are outliers that push the mean and median values. In total, there were only five questionable articles that were longer than the longest reliable article (ca. 5,000 words). As mentioned before, the minimum points in body content length however are extraction errors (length of under 10 tokens) during the annotation process. This stands in contrast to related work, where they analyzed the length distribution for English fake and real articles and found that reliable articles are around 300 characters longer than fake ones (mean value) (Schütz et al., 2021). After cleaning the outliers (deletion of 18 articles) the dataset seems rather balanced regarding the minimum and maximum of tokens. Taking a closer look at the median and mean however the questionable articles are longer.

In the next step, to get a first view of the main topics of the dataset, wordclouds are used to visualize the top 50 most common words in the data.

The wordclouds are split into reliable (Figure 2) and questionable (Figure 1). The questionable wordcloud seems to rather have no specific theme, even though some words hint to Covid being the main topic (e.g. ”Impfung” - ”vaccine”). In contrast, the reliable wordcloud rather shows keywords regarding the Russian-Ukranian war and Covid. For the removal of German stopwords and tokenization the NLTK library was used.

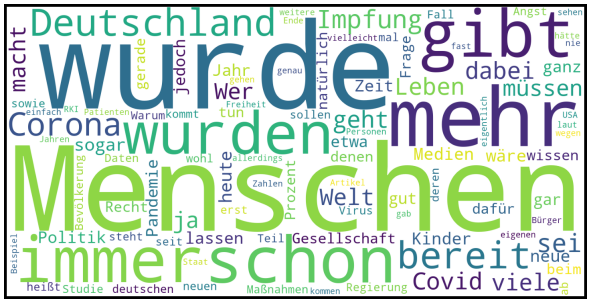

Figure 1: Wordcloud of the most common words without stopwords for questionable articles.

Figure 2: Wordcloud of the most common words without stopwords for reliable articles.

Lastly, the co-occurrence matrix for questionable articles (Figure 3) for the top 12 labels (for better readability) is shown. The darker the shade, the higher the correlation between each label. Notably, the ”General View” is not shown, since this category is not directly analyzing the content itself but gives a general information.

5. Experiments

Since the proposed dataset is mainly created to be used in ML applications, we provide a benchmark baseline. This section will give an overview of the experimental setup and results. We propose a feature-based approach on the textual content of the news articles with a transformer model. Transformers are the current state-of-the-art models in NLP and most often outperform standard approaches in NLP downstream tasks, especially in fake news detection (Yang et al., 2019; Levi et al., 2019; Rodríguez and Iglesias, 2019; Aggarwal et al., 2020; Antoun et al., 2020; Cruz et al., 2020). Transformers are pre-trained on large text corpora and can be fine-tuned on a specific task (Devlin et al., 2019), such as classification, machine translation or text generation. The unsupervised pre-training is done on millions of data entries without any labels (Devlin et al., 2019). However, during fine-tuning a model for classification tasks, labels are necessary. For the benchmark baseline, we propose the multilingual transformer model XLM-R (Conneau et al., 2019). The used XLM-R model is composed of two other transformer architectures: RoBERTa (Liu et al., 2019) and XLM (Conneau and Lample, 2019).

	WS-Subjective	RS-Claims stated	RS-Doubt	RS-Misleading	OL-Political view	RS-One-sided reporting	RS-Misinformation	RS-Political agenda	WS-Exaggerated	E-Single-issue	WS-Polarized	RS-Conspiracy
WS-Subjective	504	347	333	297	246	252	202	167	140	158	98	170
RS-Claims stated	347	499	318	269	215	203	211	169	148	139	135	183
RS-Doubt	333	318	488	284	241	226	182	155	184	125	139	168
RS-Misleading	297	269	284	451	249	228	209	154	160	149	133	149
OL-Political view	246	215	241	249	451	256	163	214	158	139	174	141
RS-One-sided reporting	252	203	226	228	256	421	152	175	179	135	187	119
RS-Misinformation	202	211	182	209	163	152	363	81	106	154	95	146
RS-Political agenda	167	169	155	154	214	175	81	332	119	80	154	77
WS-Exaggerated	140	148	184	160	158	179	106	119	308	126	92	92
E-Single-issue	158	139	125	149	139	135	154	80	126	303	92	116
WS-Polarized	98	135	139	133	174	187	95	154	92	92	291	84
RS-Conspiracy	170	183	168	149	141	119	146	77	92	116	84	285

Figure 3: Co-occurrence matrix of ”Questionable” articles.

5.1. Unsupervised Pre-Training

We further pre-trained the existing XLM-R model with additional data. The dataset we used for pre-training was collected over a period of 1,5 years - as a part of a nationally funded research project - and therefore not publicly available. It contains 194,332 gathered news articles from different sources. The articles are multilingual, however the majority are either in English or German. The articles are not annotated in terms of whether they contain disinformation and are only used for pre-training the transformer model. We trained the available XLM-R model provided by HuggingFace[2]

[2] https://huggingface.co/xlm-roberta-base

for 5 epochs, with a batch size of 16 and a learning rate of 2e-5. The probability for masked language modeling was 15% as used in the original BERT paper (Devlin et al., 2019). The training time took roughly 55 hours on one GPU.

5.2. Supervised Fine-Tuning

For each top-level category the pre-trained model was fine-tuned on the dataset for 8 epochs, a batch size of 8, a maximum sequence length of 512 and a learning-rate of 2e-5, as in Schütz et al. (2022). Each model was trained on a 90% training and 10% validation split of the dataset (same split for each model). We did not use any preprocessing steps before training nor any cleaning processes as this mostly has no significant difference in the results (Liakhovets et al., 2022; Schütz et al., 2021; Böck et al., 2021). However, we concatenated titles, subtitles and the body context in this order as an input, since we believed this could push the final accuracy of the predictions (Schütz et al., 2021).

5.3. Results & Discussion

In Table 8 the results of all trained models are shown. For the evaluation of the trained models standard evaluation metrics for classification tasks are used (threshold 0.5 for each output probability for each label), such as accuracy, precision, recall and the F1-score. The results show that no model reached the 70% mark for prediction accuracy on the validation set during training. It is also notable, that even though the *Extremism* category has one of the least amount of labels (4) recall and precision are the lowest. This could be based on the class distribution that was shown in Table 6. Especially, the classes *Left-Wing* and *Religious* are underrepresented in the dataset due to the covered topics in media beginning of 2022. Similarly, we expected the Offensive Language model to overfit on the class *Political View*, since it is overly present in comparison to all other labels (under 100 times for each label in ratio to almost 500 times for *Political View*). Lastly, the binary (based on the source of the article) classification on the other hand reaches more than 80% in all metrics.

6. Label Fusion for Inference

Since this dataset was created to be also useful for helping end-users to understand the concept of disinformation better than binary classification approaches, we propose a rule-based fusion strategy to combine the predictions of the trained models during inference. Inference in this case means a prediction on unseen data that might either be a batch of text or one article at a time.

Cat.	Acc.	Rec.	Pre.	F1	Loss
GV	0.482	0.876	0.805	0.839	0.494
OL	0.630	0.475	0.495	0.485	0.086
RS	0.485	0.398	0.503	0.444	0.272
WS	0.515	0.432	0.454	0.443	0.212
EX	0.670	0.117	0.219	0.152	0.022
B	0.844	0.848	0.856	0.827	0.280

Table 8: Results on the validation set for each category (Cat.): GV (General View), OL (Offensive Language), (RS) Reporting Style, (WS) Writing Style, (EX) Extremism, and (B) binary. (Acc.) Accuracy, (Rec.) recall, (Pre.) precision and the F1 are micro-averaged for multilabel - because of the imbalance of the labels - and macro-averaged for binary. Except (B) all models were multilabel classification tasks (top-level categories).

Therefore, we propose four categories that are distinct from each other: ”seems reliable”, ”slightly questionable”, ”questionable”, and ”highly questionable”. Each of the labels is put into one of four categories and three colors for the rule-based approach. The rules are described as the following:

- **Highly Questionable:**
  - more than or exactly 4 of red and yellow together
  - more than or exactly 2 of red
- **Questionable:**
  - more than or exactly 3 of yellow
  - 1 red, less than or exactly 2 of yellow, and green
  - more than or exactly 2 of red, 1 or none of yellow, 1 or none of green
- **Slightly Questionable:**
  - at least 1 of red
  - less than or exactly 2 of yellow
  - no red, less than 2 of yellow, and green
- **Seems Reliable:**
  - 1 or more of green, no red, no yellow
  - no labels found at all

The three categories of the labels that are needed for the rule-based approach are color-coded: ”red”, ”yellow”, and ”green”. We define red-colored labels as highly important ones for the determination whether an article could contain false information. For example, when ”highly questionable” is flagged, the models predicted more than 2

of the red-colored labels. This could be in that specific case ”Propaganda” and ”Right-Wing” extremism. Another example would be the combination of ”False Science Reporting”, ”Religious Extremism” and ”Lack of Evidence”. The color coding for each labels are the following:

- **Red:** *Propaganda, Conspiracy, Misleading, Misinformation, False Science Reporting, Lack of Evidence, Left-Wing, Right-wing, Religious Extremism, Single issue, Anti-Semitism, Populistic, Polarized*
- **Yellow:** *Subjective, False Context, Logic Flaws, One-Sided Reporting, Political Agenda, Religion, Nationality, Ethnicity, Race, Anti-LGTBQIA+, Exaggerated, Sex*
- **Green:** *Clickbait, Sensationalism, Doubt, Claims Stated, Culture, Political View, Physical and Psychological Characteristics, Humoristic, Unprofessional Writing Style, Social Status, Repetition*

7. Conclusion

In this work, we presented a novel multilabel dataset for the detection of disinformation in German news articles. The dataset has in total 39 labels in 5 top-level categories. The categories include a vast amount of concepts to cover other related research fields such as offensive language, and extremism . Because of the broad set of labels, each combination of labels can give clues whether an article might contain false information and can be easily assessed by human experts. This sets a basis for an automatic detection system, where experts can have an initial and fast overview over potentially critical content, especially in large data collections.

8. Acknowledgements

This contribution has been funded by the FFG projects “Defalsif-AI” (grant no. 879670), ”RAIDAR” (grant no. 886364) and ”DesinFact” (grant no. 48961119) in the Austrian security research program KIRAS of the Federal Ministry of Finance (BMF)).

9. Ethical Considerations & Limitations

There are multiple limitations we are aware of - also amid the creation of this dataset. During the whole process and the modeling of the annotation scheme there were experts of multiple domains involved to ensure a high quality of the labels and helpful overview of the concept of disinformation. This includes social scientists, journalists and fact-checkers as well as feedback regarding data protection and legal issues. However, creating a language resource is usually coupled with drawbacks.

Regarding the annotation of the dataset, there could be a possible annotator-bias, since there was only one person to annotate each article, which is highly subjective. Even though each article got checked and confirmed by two more persons, there could be wrongly annotated instances. The preferred solution to such issue would be an approach for tracking inter-annotator agreements. Another bias could be the manual retrieval of the dataset, often leading to 100 to 200 articles by one website for a week of annotation. However, this cannot be measured, since the articles got randomly assigned to one of the 11 annotators. Lastly, the annotators could see the website domain during the annotation process, which could already lead to an influence on the labeling process by the annotator’s opinion. This is largely based on the half-automated annotation tool provided. Blogs and lesser known websites can already give a certain impression. However, the task for the annotators was not to decide whether an article is reliable or not. Certainly, all those issues are known factors of human annotation processes and are mostly dependent on resources. In future work, we plan to have a more stable labeling system, the data should be annotated by two more trained annotators. This must be taken into account when using this dataset - the intention behind it is rather to help data analysis experts during exploratory tasks and getting an overview of large scale unlabeled data than labeling an article as definitively fake.

10. Bibliographical References

Gary Ackerman and Anastasia Kouloganes. 2019. Single-issue terrorism. *The Oxford handbook of terrorism*, pages 316–330.

Sadia Afroz, Michael Brennan, and Rachel Greenstadt. 2012. Detecting hoaxes, frauds, and deception in writing style online. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy*, SP ’12, page 461–475, USA. IEEE Computer Society.

Akshay Aggarwal, Aniruddha Chauhan, Deepika Kumar, Mamta Mittal, and Sharad Verma. 2020. Classification of fake news by fine-tuning deep bidirectional transformers based language model. *EAI Endorsed Transactions on Scalable Information Systems: Online First*, 7.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31:211–236.

Wissam Antoun, Fady Baly, Rim Achour, Amir Hussein, and Hazem Hajj. 2020. State of the art models for fake news detection tasks. In *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*, pages 519–524.

Femi Emmanuel Ayo, Olusegun Folorunso, Friday Thomas Ibharalu, Idowu Ademola Osinuga, and Adebayo Abayomi-Alli. 2021. A probabilistic clustering model for hate speech classification in twitter. *Expert Systems with Applications*, 173:114762.

George Bara, Gerhard Backfried, and Dorothea Thomas-Aniola. 2019. Fake or fact? theoretical and practical aspects of fake news. In Éloi Bossé and Galina L. Rogova, editors, *Information Quality in Information Fusion and Decision Making*, pages 181–206. Springer.

Jaqueline Böck, Daria Liakhovets, Mina Schütz, Armin Kirchknopf, Djordje Slijepčević, Matthias Zeppelzauer, and Alexander Schindler. 2021. AIT_FHSTP at GermEval 2021: Automatic fact claiming detection with multilingual transformer models. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 76–82, Duesseldorf, Germany. Association for Computational Linguistics.

Yimin Chen, Niall J. Conroy, and Victoria L. Rubin. 2015. Misleading online content: Recognizing clickbait as "false news". In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, WMDD '15, page 15–19, New York, NY, USA. Association for Computing Machinery.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Niall Conroy, Victoria L. Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52.

Jan Christian Blaise Cruz, Julianne Agatha Tan, and Charibeth Cheng. 2020. Localization of fake news detection via multitask transfer learning. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2596–2604, Marseille, France. European Language Resources Association.

Christoph Demus, Jonas Pitz, Mina Schütz, Nadine Probol, Melanie Siegel, and Dirk Labudde. 2022. Detox: A comprehensive dataset for German offensive language and conversation analysis. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 143–153, Seattle, Washington (Hybrid). Association for Computational Linguistics.

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.

Karen M Douglas, Joseph E. Uscinski, Robbie M. Sutton, Aleksandra Cichoka, Turkay Nefes, Chee Siang Ang, and Farzin Deravi. 2019. Understanding conspiracy theories. pages 3–35.

Álvaro Figueira and Luciana Oliveira. 2017. The current state of fake news: challenges and opportunities. *Procedia Computer Science*, 121:817–825.

Dongqi Fu, Yikun Ban, Hanghang Tong, Ross Maciejewski, and Jingrui He. 2022. Disco: Comprehensive and explainable disinformation detection.

Benjamin D. Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news.

Edson C. Tandoc Jr., Zheng Wei Lim, and Richard Ling. 2018. Defining "fake news": A typology of scholarly definitions. *Digital Journalism*, 6(2):137–153.

Sebastian Jungkunz. 2019. Towards a measurement of extreme left-wing attitudes. *German Politics*, 28(1):101–122.

Sayeed Ahsan Khan, Mohammed Hazim Alkawaz, and Hewa Majeed Zangana. 2019. The use and abuse of social media for spreading fake news.

In *2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*, pages 145–148.

Juliane Köhler, Gautam Kishore Shahi, Julia Maria Struß, Michael Wiegand, Melanie Siegel, and Thomas Mandl. 2022. Overview of the CLEF-2022 CheckThat! lab task 3 on fake news detection. In *Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum*, CLEF '2022, Bologna, Italy.

Or Levi, Pedram Hosseini, Mona Diab, and David Broniatowski. 2019. Identifying nuances in fake news vs. satire: Using semantic and linguistic cues. *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*.

Daria Liakhovets, Mina Schütz, Jaqueline Böck, Medina Andresel, Armin Kirchknopf, Andreas Babic, Djordje Slijepčević, Jasmin Lampert, Alexander Schindler, and Matthias Zeppelzauer. 2022. Transfer learning for automatic sexism detection with multilingual transformer models. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*, A Coruna, Spain. CEUR-WS.org.

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.

Giovanni Martino, Alberto Barrón-Cedeño, and Preslav Nakov. 2019. Findings of the nlp4if-2019 shared task on fine-grained propaganda detection.

Maria D. Molina, S. Shyam Sundar, Thai Le, and Dongwon Lee. 2021. "fake news" is not simply false information: A concept explication and taxonomy of online content. *American Behavioral Scientist*, 65(2):180–212.

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Julia Maria Struß, Thomas Mandl, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, Gautam Kishore Shahi, Hamdy Mubarak, Alex Nikolov, Nikolay Babulkov, Yavuz Selim Kartal, Javier Beltrán, Michael Wiegand, Melanie Siegel, and Juliane Köhler. 2022. Overview of the CLEF-2022 CheckThat! lab on fighting the COVID-19 infodemic and fake news detection. In *Proceedings of the 13th International Conference of the CLEF Association: Information Access Evaluation meets Multilinguality, Multimodality, and Visualization*, CLEF '2022, Bologna, Italy.

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, Marseille, France. European Language Resources Association.

Daniela Pisoiu and Sandra Hain. 2017. *Theories of Terrorism: An Introduction*.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Julio C. S. Reis, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benevenuto. 2019. Explainable machine learning for fake news detection. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, page 17–26, New York, NY, USA. Association for Computing Machinery.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12, Duesseldorf, Germany. Association for Computational Linguistics.

Álvaro Ibrain Rodríguez and Lara Lloret Iglesias. 2019. Fake news detection using deep learning.

Mina Schütz, Alexander Schindler, Melanie Siegel, and Kawa Nazemi. 2021. Automatic fake news detection with pre-trained transformer models. In *Pattern Recognition. ICPR International Workshops and Challenges*, pages 627–641, Cham. Springer International Publishing.

Mina Schütz, Jaqueline Böck, Medina Andresel, Armin Kirchknopf, Daria Liakhovets, Djordje Slijepčević, and Alexander Schindler. 2022. Ait_fhstp at checkthat! 2022: Cross-lingual fake news detection with a large pre-trained transformer. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum*, pages 660–670, Bologna, Italy.

Mina Schütz, Jaqueline Böck, Daria Liakhovets, Djordje Slijepčević, Armin Kirchknopf, Manuel Hecht, Johannes Bogensperger, Sven Schlarb, Alexander Schindler, and Matthias Zeppelzauer. 2021. Automatic sexism detection with multilingual transformer models. In *Proceedings of the*

Iberian Languages Evaluation Forum (IberLEF 2021), pages 346–355, Málaga, Spain.

Gautam Kishore Shahi and Durgesh Nandini. 2020. FakeCovid – a multilingual cross-domain fact check news dataset for covid-19. In *Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media*.

Gautam Kishore Shahi, Julia Maria Struß, and Thomas Mandl. 2021. Overview of the clef-2021 checkthat! lab task 3 on fake news detection. *Working Notes of CLEF*.

Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2019. Combating fake news: A survey on identification and mitigation techniques. *ACM Trans. Intell. Syst. Technol.*, 10(3).

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.

Kai-Chou Yang, Timothy Niven, and Hung-Yu Kao. 2019. Fake news detection as natural language inference. *CoRR*, abs/1907.07347.

Helen Young and Geoff Boucher. 2022. Digital conspiracies and accelerationist fictions. *Arena*, (11):69–74.

Xinyi Zhou and Reza Zafarani. 2018. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 0(ja).

Xinyi Zhou and Reza Zafarani. 2019. Fake news detection: An interdisciplinary research. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 1292, New York, NY, USA. Association for Computing Machinery.

11. Language Resource References

Ahmed, Hadeer and Traore, Issa and Saad, Sherif. 2017. *Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques*. Springer International Publishing.

Hadeer Ahmed and Issa Traoré and Sherif Saad. 2018. *Detecting opinion spams and fake news using text classification*.

Maurício Gruppi and Benjamin D. Horne and Sibel Adalı. 2020. *NELA-GT-2019: A Large Multi-Labelled News Dataset for The Study of Misinformation in News Articles*.

Elena Kochkina and Maria Liakata and Arkaitz Zubiaga. 2018. *PHEME dataset for Rumour Detection and Veracity Classification*.

Juliane Köhler and Gautam Kishore Shahi and Julia Maria Struss and Michael Wiegand and Melanie Siegel and Mina Schütz. 2022. *Overview of the CLEF-2022 CheckThat! Lab: Task 3 on Fake News Detection*.

Mattern, Justus and Qiao, Yu and Kerz, Elma and Wiechmann, Daniel and Strohmaier, Markus. 2021. *FANG-COVID: A New Large-Scale Benchmark Dataset for Fake News Detection in German*. Association for Computational Linguistics.

Tanushree Mitra and Eric Gilbert. 2015. *CREDBANK: A Large-Scale Social Media Corpus With Associated Credibility Annotations*.

Nakamura, Kai and Levy, Sharon and Wang, William Yang. 2020. *Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection*. European Language Resources Association.

Nielsen, Dan S. and McConville, Ryan. 2022. *MuMiN: A Large-Scale Multilingual Multimodal Fact-Checked Misinformation Social Network Dataset*. Association for Computing Machinery, SIGIR '22.

Parth Patwa and Shivam Sharma and Srinivas PYKL and Vineeth Guptha and Gitanjali Kumari and Md. Shad Akhtar and Asif Ekbal and Amitava Das and Tanmoy Chakraborty. 2020. *Fighting an Infodemic: COVID-19 Fake News Dataset*.

Santia, Giovanni and Williams, Jake. 2018. *BuzzFace: A News Veracity Dataset with Facebook User Commentary and Egos*.

Gautam Kishore Shahi and Durgesh Nandini. 2020. *FakeCovid - A Multilingual Cross-domain Fact Check News Dataset for COVID-19*.

Kai Shu and Deepak Mahudeswaran and Suhang Wang and Dongwon Lee and Huan Liu. 2018. *FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media*.

Srba, Ivan and Pecher, Branislav and Tomlein, Matus and Moro, Robert and Stefancova, Elena and Simko, Jakub and Bielikova, Maria. 2022. *Monant Medical Misinformation Dataset: Mapping Articles to Fact-Checked Claims*. Association for Computing Machinery, SIGIR '22.

Andreas Ströckl. 2020. *Fake News Dataset German*.

Eugenio Tacchini and Gabriele Ballarin and Marco L. Della Vedova and Stefano Moret and Luca de Alfaro. 2017. *Some Like it Hoax: Automated Fake News Detection in Social Networks*.

Thorne, James and Vlachos, Andreas and Christodoulopoulos, Christos and Mittal, Arpit. 2018. *FEVER: a Large-scale Dataset for Fact Extraction and VERification*. Association for Computational Linguistics.

Inna Vogel and Peter Jiang. 2019. *Fake News Detection with the New German Dataset "German-FakeNC"*.

William Yang Wang. 2017. *"Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection*.