DeFaktS: A German Dataset for Fine-Grained Disinformation Detection through Social Media Framing

Shaina Ashraf[1,4,5,6], **Isabel Bezzaoui**[2], **Ionut Andone**[3], **Alexander Markowetz**[1,3], **Jonas Fegert**[2], **Lucie Flek**[4,5,6]
[1]Department of Mathematics and Computer Science, University of Marburg
[2]FZI Forschungszentrum Informatik
[3]Murmuras GmbH
[4]Conversational AI and Social Analytics (CAISA) Lab
[5]Bonn-Aachen International Center for Information Technology (b-it), University of Bonn
[6]The Lamarr Institute for Machine Learning and Artificial Intelligence
ashrafs@staff.uni-marburg.de, Bezzaoui@fzi.de, johnny@murmuras.com
markowet@mathematik.uni-marburg.de, fegert@fzi.de, flek@bit.uni-bonn.de

Abstract

In today's rapidly evolving digital age, disinformation poses a significant threat to public sentiment and socio-political dynamics. To address this, we introduce a new dataset "DeFaktS", designed to understand and counter disinformation within German media. Distinctively curated across various news topics, DeFaktS offers an unparalleled insight into the diverse facets of disinformation. Our dataset, containing 105,855 posts with 20,008 meticulously labeled tweets, serves as a rich platform for in-depth exploration of disinformation's diverse characteristics. A key attribute that sets DeFaktS apart is its fine-grained annotations based on polarized categories. Our annotation framework, grounded in the textual characteristics of disinformation content, eliminates the need for external knowledge sources. Unlike most existing corpora that typically assign a singular global veracity value to news, our methodology seeks to annotate every structural component and semantic element of a news piece, ensuring a comprehensive and detailed understanding. In our experiments, we employed a mix of classical machine learning and advanced transformer-based models. The results underscored the potential of DeFaktS, with transformer models, especially the German variant of BERT, exhibiting pronounced effectiveness in both binary and fine-grained classifications.

Keywords: Disinformation, Twitter, X, German, Dataset, Fake News, Classification, Framing

1. Introduction

In the contemporary information era, the rapid proliferation of online platforms has reshaped communication paradigms. Social platforms have democratized information dissemination, ensuring real-time data sharing. This accessibility, however, is a double-edged sword. On one hand, it promotes knowledge sharing; on the other, it has become a conduit for the spread of disinformation or 'fake news' (Shu et al., 2017).

The implications of unchecked fake news are severe. Beyond the obvious erosion of public trust in media and institutions, disinformation can sway public opinion, influence election outcomes, and even catalyze real-world harm (Strömbäck, 2005; Groshek and Koc-Michalska, 2017). In the face of these challenges, ensuring the veracity of digital content has become an imperative. Empirical findings underscore the intricate complexity of disinformation, which, with its deceptive nature, strives to cloak itself as legitimate information, making its detection notably elusive (Shu et al., 2020). While studies emphasize that authentic and deceptive news articles demonstrate substantial disparities in their substantive content (Horne and Adali, 2017; Abonizio et al., 2020), the nuanced and multifaceted characteristics of disinformation amplify the challenge (Rosińska, 2021). Moreover, the lexical and structural features of disinformation often tend to be event-specific, suggesting that classifiers trained on one type of event or topic may underperform when faced with deceptive content derived from a different context (Shu et al., 2017). This multi-dimensional complexity and subtlety of disinformation necessitate innovative approaches that can navigate through its nuanced landscapes, offering a more holistic understanding and detection mechanism. In the realm of disinformation research, while English has been the primary focus, other significant languages like German have not received equivalent attention. This oversight is particularly evident in the scarcity of comprehensive annotated datasets dedicated to the German language, especially in the domain of disinformation analysis (Schreiber et al., 2021). Furthermore, Germany itself faces a pronounced challenge with disinformation, as indicated by its high number of QAnon members, ranking second globally. The unique linguistic characteristics and cultural con-

texts of German differentiate it from English, and the limited availability of annotated datasets for German compounds the complexities of disinformation detection in this language. This study navigates through these challenges by presenting a comprehensive approach to understanding and mitigating disinformation, especially within the German linguistic context, through three pivotal contributions:

- Introducing a richly curated and annotated dataset that encompasses a diverse array of topics and keywords from the German media, meticulously annotated with binary and fine-grained labels to serve as a foundational resource for developing and evaluating fake news detection algorithms.
- Recognizing the complex nature of disinformation, we propose a comprehensive and fine-grained taxonomy-based annotation scheme encompassing linguistic, semantic, psychological, and authenticity features formulated to facilitate a detailed and structured approach to analyzing and labeling tweets.
- The study further presents experiments employing both classical machine learning models and transformer-based models, providing initial insights into the dataset's utility and serving as a starting point for subsequent research endeavors to develop and refine disinformation detection models in the German language.

2. Related Work

Recent efforts in combating disinformation have largely centered around leveraging advanced machine learning techniques and developing datasets to facilitate the training and evaluation of models designed to discern the veracity of information disseminated online. Ali et al. (2022) focused on Arabic fake news detection related to COVID-19 on Twitter (now X) and Facebook. The authors introduced a new Arabic COVID-19 dataset and applied two pre-trained classification models, AraBERT and BERT base Arabic. Abd Rahim and Basri (2022) introduced MalCov, a dataset containing fake and valid news articles related to COVID-19 in the Malay language. The dataset, which comprises articles from social media platforms and has been validated by local authorities, was utilized to build classifiers using machine learning models such as Naive Bayes, SVM, and Logistic Regression. Suryavardan et al. (2023) introduced Factify 2, a multimodal fact-checking dataset that enhances its predecessor, Factify 1, by incorporating new data sources and adding satire articles. Factify 2 categorizes data into three broad categories (support, no-evidence, and refute) and subcategories based on the entailment of visual and textual data, providing a rich resource for developing and evaluating multimodal fake news detection models. Ciora and Cioca (2022) developed RoCoFake, a Romanian Covid-19 Fake News dataset, aggregating various online resources like tweets, news titles, and fact-checking news sites. RoCoFake addresses the scarcity of resources for fake news detection in the Romanian language, providing a valuable resource in the medical domain. Carrella et al. (2023) emphasized the importance of developing language-specific datasets and models to address the challenge of disinformation in Italian. (Plepi et al., 2022) conducted an in-depth analysis of users' time-evolving semantic similarities and social interactions, revealing that these patterns can be indicative of disinformation spread. Building on these findings, they proposed a dynamic graph-based framework that capitalizes on the fluidity of user networks to isolate fake news spreaders. Fatima et al. (2023) introduced YouFake, a multimodal dataset that includes both images and texts collected from popular YouTube channels, provides a comprehensive platform for developing and evaluating models that can handle multi-modal data (text, image, and video) for fake news classification.

These studies underscore the global and multilingual nature of the disinformation challenge, highlighting the importance of developing datasets and models that cater to various linguistic and cultural contexts. While these datasets provide valuable insights and resources for fake news classification Sakketou et al. (2022), it is evident that there is a gap in the availability of German-specific datasets for fake news detection, highlighting a potential area for contribution and development in the field. Moreover, the available datasets often exhibit a lack of diversity in topics and news categories, frequently concentrating on specific themes or health crises like the COVID-19 pandemic Mattern et al. (2021). This limitation potentially restricts the generalizability and applicability of models trained on such datasets to a broader spectrum of topics and contexts. Furthermore, there is a noticeable scarcity of datasets that provide transparent and comprehensive annotation schemes for labeling fake news Murayama et al. (2022). The meticulousness and granularity in labeling are pivotal for developing models that can discern and understand disinformation's nuanced and multifaceted nature. Many existing datasets Vogel and Jiang (2019), Ahuja and Kumar (2023) do not offer fine-grained labels or employ polar labeling schemes that enable annotators to adeptly identify and categorize various dimensions and spectrums of deceptive information.

In response to these gaps, we introduce "DeFaktS" [1], a dataset uniquely designed for German

[1] https://github.com/caisa-lab/

media. Our dataset not only offers a comprehensive understanding of misinformation within this specific linguistic context but also brings forth a novel approach in its annotation and structure. "DeFaktS" is thoroughly curated, emphasizing granularity in labels and ensuring that various dimensions of misinformation are adeptly captured. The annotation scheme, and correspondingly the labels utilized, are designed based on the Taxonomy of Online Disinformation developed by(Bezzaoui et al., 2022). Combining empirical findings from various fields such as computer science, linguistics, psychology, and media studies, the taxonomy gathers the many underlying linguistic features in disinformation into a schematic framework. Additionally, inspired by (Kellner and Share, 2005) critical media literacy framework and (Abonizio et al., 2020; Horne and Adali, 2017; Molina et al., 2021), our strategy for building an annotation framework revolves around addressing three key research objectives: First, the identification of specific linguistic cues that signify online disinformation, as highlighted in empirical literature. Second, the organization of these linguistic features into a coherent and comprehensive schema. Third, the integration of these dimensions and categories into a clearly defined, structured taxonomy. This positions "DeFaktS" not just as another dataset but as a contribution to the ongoing global effort to curb the influence of disinformation.

3. Dataset

Twitter[2] (now X) is a primary hub for real-time news dissemination. Its influence, coupled with the potential for spreading deceptive content that can mold public opinions, underscores its significance (Li and Su, 2020; Zhou et al., 2021). Therefore, we chose it as our primary data source.

3.1. Data Collection

Our "DeFaktS" dataset is methodically created, focusing on the German media domain, ensuring a robust and comprehensive collection suitable for in-depth analysis of various news topics. Initially, we compiled a list of 129 pertinent and diverse news topics, which were predominantly trending at the time of data collection. This included a range of controversial and high-impact topics such as elections, the energy crisis, lockdown measures, the war on Ukraine, the gender pay gap, immigrants, climate, and inflation, among others. A word cloud depicting the prominence of these topics within our dataset can be seen in Figure 5. In order to establish the topics, we started with a set of related keywords. We then collected German-language tweets that contained these keywords and added the first 2000 tweets that fit our criteria to our database. Given Twitter's (now X) dynamic nature and the prevalence of retweets, we removed duplicate entries to avoid any potential biases in our subsequent analyses.

3.2. Data Annotation

3.2.1. Fine-Grained Labels Annotation Scheme

The primary objective of the data annotation was to scrutinize the tweets, identifying and highlighting instances indicative of disinformation. In pursuit of this, a detailed annotation framework was designed, which has general category labels and more nuanced polar labels, each dissecting distinct facets of the tweets and pinpointing specific features potentially signaling disinformation. To ensure that current empirical knowledge on the text-based identification of disinformation is taken into account the annotation framework is based on the Taxonomy of Online Disinformation (TOD) (Bezzaoui et al., 2022). The taxonomy synthesizes scientific evidence from various disciplines into a concise overview covering dimensions ranging from more granular characteristics such as semantic aspects (Cardoso et al., 2021) of disinformation to broader aspects for categorization such as various content types.

The "DeFaktS" annotation scheme was specifically developed to dissect and identify the framing techniques utilized in the dissemination of disinformation through German social media. Our comprehensive labeling approach is geared towards detecting the nuanced ways in which information is framed, which can influence perceptions and propagate disinformation. Our annotation process is rooted in four principal dimensions: content type, authenticity, semantic, and psychological features, each chosen for its empirical association with disinformation. Semantic features helps to analyze the content for meaning and consistency, as disinformation is often riddled with contradictions or repeated content lacking new insights (Horne and Adali, 2017; Azevedo et al., 2021). Psychological features encompass tactics like polarization, emotionalization, and sensationalism. These features construct narrative frames that manipulate emotional biases to enhance engagement and dissemination (Jeronimo et al., 2019; Gruppi et al., 2018; Ribeiro Bezerra, 2021; Wang et al., 2019; Vicario et al., 2019). Authenticity features assess the authenticity of references and the clarity of phrasing, helping to determine whether the information is framed within a reliable context or crafted to mislead by obfuscating facts (Fernandez, 2019; Kumar

DeFaktS-Dataset-Disinformaton-Detection

[2] https://twitter.com/home

Category	Dimension	Feature	Code	Description
Polar Labels	semantic features	semantic inconsistency	infoincon	Disinformation exhibits a higher degree of contentual inconsistencies like semantic contradictions or logic errors throughout the text. (Cardoso et al., 2021)
		lack of (new) information	infonewinfo	The body of unreliable posts add relatively little new information, but serves to repeat and enhance the claims made at the beginning. (Horne and Adali, 2017), (Azevedo et al., 2021)
	psychology features	polarization	psychpolar	Unreliable posts frequently narrate in terms of a clear friend-foe-distinction with regard to specific national, ethical, or religious groups or elites as foes or perpetrators. The opposing group (often framed in a common "we", "ourselves", "the government") takes the part of the victim who needs to be protected. (Vicario et al., 2019)
		emotionalization	psychemo	Unreliable sources incline to use a more emotionally persuasive language and touch more often sensible subjects (like children, death and burial). (Wang et al., 2019), (Ribeiro Bezerra, 2021)
		sensationalism	psychsensa	Fake posts tend to be written in a hyperbolic way to attract the reader's attention, i.e. with a high usage of all-caps-words, exclamation marks or a general sentiment wording. (Gruppi et al., 2018), (Jeronimo et al., 2019)
		abasement	psychabas	Disinformation frequently entails stereotype narratives and ressentiments to denigrate targeted groups.
	authenticity features	topicality	psychtopic	Legitimate sources tend to report about past events whereas fake articles focus on highly recent topics.(Fernandez, 2019)
		vagueness of phrasing	authvague	Fake posts use a higher amount of hedging words (like 'possibly', 'usually', 'tend to be') to achieve a more indirect form of expression. Also they evoke a feeling of uncertainty by addressing the vagueness of information directly. (Mahyoob et al., 2020)
		authenticity/referencing of information	authrefer	Legitimate sources are considerably better referenced than unreliable articles. Unreliable sources tend to use none, false or wrong contextualized references. (Kumar et al., 2016)
	content type features	pseudoscientific	typpseudo	Content that calls on supposedly scientific research or reputable institutions without identifying concrete sources or by manipulating them to create a false theory. (Rosińska, 2021)
		forged content	typforged	Stories that lack any factual ground or manipulated information or image. The intention is to deceive and cause harm. Could be text or visual media. (Kapantai et al., 2021)
		false context	typfalcontex	Real information is being presented in a false context. The recipient is aware that the information is true, but he does not realize that the context has been changed. (Kapantai et al., 2021)
		conspiracy theory	typconspir	Stories without factual basis which usually explain important events as secret plots by government or powerful individuals. By definition their truthfulness is difficult to verify. Evidence refuting the conspiracy is regarded as further proof of the conspiracy. (Kapantai et al., 2021)
		propaganda	typpropa	Information that is created by a political entity to influence public opinion and gain support for a public figure, organization or government. (Rashkin et al., 2017)
		no factual content	typopinion	This rating is used for posts that are pure opinion, comics, satire, or any other posts that do not make a factual claim. This is also the category to use for posts that are of the "Like this if you think..." variety. (Tandoc Jr et al., 2018)
General Labels			corpkeyword	Keywords used to search tweets. This label doesn't indicate polarity but marks the span of text containing the search keyword.
			catposfake	Category indicating possible fake news. This label is attributed to posts that receive one of the polar labels
			catneutral	Indicates neutral posts where there's no indication of fake news. Such posts should never have any other polar labels.

Figure 1: Fine-Grained Annotation Framework

et al., 2016; Mahyoob et al., 2020). Content Type features address the thematic framing of content, including pseudo scientific claims, forged content, and propaganda. Such framing shapes audience perceptions and is an integral part of disinformation strategies (Rosińska, 2021; Kapantai et al., 2021; Bąkowicz, 2019; Tandoc Jr et al., 2018; Rashkin et al., 2017; Tandoc Jr et al., 2018).

To ensure fidelity and uniformity in our annotations, domain experts from the Center for Monitoring, Analysis, and Strategy (CeMAS)[3] conducted a rigorous training workshop. Here, annotators were equipped with guidelines and engaged in activities using sample data, which honed their ability to recognize text passages containing deceptive indicators aligned with our polar labels. Figure 1 and 2 illustrate the framework and provide examples of tweets annotated with these labels, demonstrating the application of our method and underscoring the role of each feature in pinpointing disinformation.

[3] https://cemas.io/en/

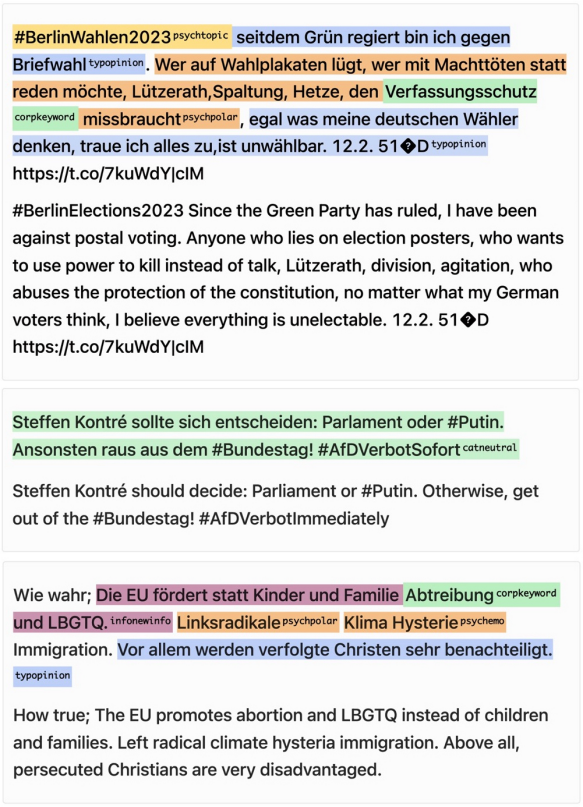

Figure 2: Annotated Samples: Original German and Translated English Text for Three Tweets

3.2.2. Binary Labels

In addition to the multi-label annotation scheme that categorizes posts into an array of polar and general labels, a binary classification scheme is also employed to demarcate between two primary categories:

- **Real News** is dedicated to posts that are regarded as neutral in nature. Specifically, posts under this umbrella contain exclusively the label 'catneutral'.
- **Fake News** represents posts that exhibit traits indicative of potential disinformation or bias. Posts allocated to this category contain at least one of the polar labels but are devoid of the label 'catneutral'.

3.2.3. Annotation Platform

In this study, we utilized Doccano (Nakayama et al., 2018), an open-source annotation tool, to facilitate our annotation process, primarily owing to its user-friendly interface and capability to streamline collaborative efforts. Doccano is well-equipped with features tailored to our task requirements, thereby making it an apt choice for managing our annotation activities. The project was configured as a sequence labeling task, enabling the annotators to select specific text spans and assign labels to them, supporting multiple labeling. Furthermore, annotators had the flexibility to select the entirety of the text to assign general category labels. Prior to uploading the data to Doccano, default labels with the code 'corpkeyword' were assigned to highlight keywords within the text, which were initially used for filtering tweets during the data collection process (as also mentioned in the annotation scheme). Additionally, comprehensive annotation guidelines were uploaded to the platform, serving as a readily available reference for annotators during the text annotation process, ensuring consistency and adherence to the specified labeling criteria.

3.2.4. Cross Annotation

To fortify the robustness and dependability of our annotations, we undertook a process of cross-annotation. A subset of 767 samples was independently annotated by two annotators, ensuring a thorough examination of both our fine-grained labels and binary labels. Consequently, inter-annotator agreement (IAA) (McHugh, 2012) was computed for both labeling methods to gauge the level of concordance between the annotators. In the cross-annotation subset, we observed disagreements across the labels: 53 for binary labels and 95 for fine-grained labels. Given that the fine-grained labels span 17 categories, higher contradictions were seen compared to binary labels. To quantify the IAA, we employed Cohen's Kappa metric, unveiling a substantial agreement with a score of 0.72 for binary labels. For fine-grained labels, which naturally present a more complex annotation scenario, the average score across multiple labels was 0.56, indicating a moderate level of agreement. In an additional layer of evaluation, and to assess the similarity in the sets of fine-grained labels assigned by the annotators to each instance, we calculated Jaccard Similarity Score, achieving a noteworthy score of 0.88. This score, paired with Cohen's Kappa metric, affirms the robustness and reliability of the annotations across our dataset, ensuring a solid foundation for the subsequent experiments and analyses.

3.2.5. Dataset Statistics

The dataset comprises a total of 105,855 posts, where 20,008 tweets are labeled with the class distribution of 11,776:8,232 of Real News and Fake News, respectively. The dataset encapsulates a variety of attributes for each tweet, enabling analyses related to temporal patterns, identifying topics, trends, and user engagements. Upon curating the "DeFaktS" dataset, a thorough exploratory data analysis was conducted to comprehend the underlying patterns and characteristics inherent to the

collected attributes. All the other polar labels have

Data	Stats
Unique Users	44,486
Average Tweet Length (chars)	187
Average Tweet Length (words)	24
Average Likes	22
Average Retweets	4
Average Replies	3
Average Quotes	0.4
Average Tweets/User	3
Number of tweets with URLs	65,889

Table 1: Basic Data Stats

varying counts associated with them; the most frequently associated polar label is typo-pinion, with 5,354 occurrences, followed by 'psychsensa' with 2,056 occurrences. There are no specific polar labels associated with Real News in the dataset. This means the dataset's Real News entries do not have any polar labels from the annotation guidelines, which aligns with the notion that these polar labels are indicators of fake or unreliable information. The label 'typopinion' has the highest occurrence, suggesting that many fake news tweets in the dataset are opinion-based without factual content. Labels like 'psychsensa' (indication of sensationalism) and 'psychemo' (indication of emotionalization) also have significant occurrences, indicating common features of sensationalism and emotional language in fake news. Given this analysis, we can infer that fake news in the dataset frequently exhibits features such as sensationalism, emotionalization, lack of proper referencing, and more. To better understand this, we can visualize a bar graph of the polar label distribution for the tweets in Figure 3, similarly the distribution across binary labels is shown in Figure 4.

label distribution for the tweets in Figure fig:figure2-polar; similarly,

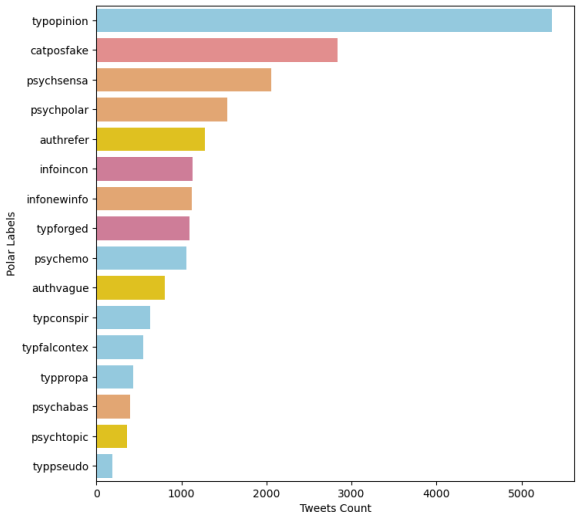

Figure 3: Distribution of Polar Labels in Fake News

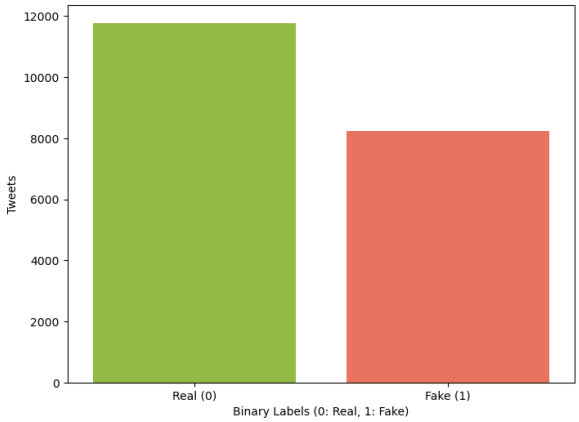

Figure 4: Distribution of Binary Labels (Real vs. Fake)

4. Methodology

4.1. Preprocessing

In our research, preprocessing was crucial to mitigate noise and ensure data quality. We executed several steps, including stopword removal, lowercase conversions, tokenization and lemmatization . Additionally, we stripped URLs to eliminate potential source link biases, ensuring a cleaner dataset for feature extraction and model training.

4.2. Features and Text Encoding

To represent our text data, the following features and embeddings were utilized for model training.

- **Bag of Words (BOW)**: A vector representation counting word occurrences, ignoring grammar and word order (Qader et al., 2019).
- **Term Frequency-Inverse Document Frequency (TF-IDF)**: Highlights word frequency in a document relative to its frequency across all documents, offering a measure of its importance (Havrlant and Kreinovich, 2017).
- **Word2Vec**: Embeddings that capture semantic meanings of words, using pre-trained models on German Wikipedia with 100-dimensional representations (Yamada et al., 2020).
- **GerVADER Sentiment (GVSent)**: Sentiment-based features derived using GerVader (Tymann et al., 2019) to determine word polarity, providing overall sentiment scores of tweets.

4.3. Traditional ML Classifiers

We utilized the following classical machine learning models in our baseline experiments:

- **Support Vector Machines (SVM):** A supervised algorithm recognized for its effectiveness

in text classification by finding the optimal hyperplane for data separation (Wang, 2005).

- **Random Forest (RndFor):** Constructs multiple decision trees for high accuracy and can handle large datasets as well as missing values.
- **Logistic Regression (LogReg):** Commonly used for binary classification but adaptable for multilabel tasks using methods like the one-vs-rest (OvR) approach.

4.4. Deep Learning models

Acknowledging the prowess of language models in diverse Natural Language Processing (NLP) tasks and their ability to grasp the contextual relationships between words, we fine-tuned several state-of-the-art pre-trained language models using our dataset.

- **BERT-Base:** Pretrained on English data, it is recognized for capturing deep contextual word relationships (Devlin et al., 2018).
- **BERT-Multilingual:** Trained on 104 languages, this variant of BERT is adept at handling linguistic diversity, making it suitable for diverse languages, including German (Pires et al., 2019).
- **BERT-German:** Tailored for German, it captures linguistic nuances specific to the language while also understanding cross-lingual patterns.
- **Xlm-RoBERTa:** An advanced BERT variant trained on a vast corpus, known for its high performance in various NLP tasks (Conneau et al., 2019).

5. Experimental Setup

In our experiments, we evaluated the models across two distinct classification paradigms: binary (distinguishing between Fake and Real News) and fine-grained (categorizing across 17 labels). Confronted with a pronounced class imbalance in our dataset between Real and Fake news instances, we resorted to downsampling the 'Real News' category. This strategy was instrumental in ensuring parity in representation between the Real and Fake news categories, a balance we maintained for both classification tasks. However, we refrained from further downsampling when transitioning to the fine-grained classification. Given the varied distribution across the 17 labels, additional downsampling could risk discarding valuable data, particularly for polar labels with limited samples. As the next step in our process, we employed a consistent preprocessing pipeline across all models. We established a 5-fold cross-validation for our classical ML models to assess their performance and ensure robustness in our analysis. For features like BOW and TF-IDF, the vectorizer was restricted to a maximum of 5,000 features, considering both unigrams and bigrams. For our transformer-based models, We partitioned the dataset into training (80%), validation (10%), and test (10%) sets. The training set was utilized to fine-tune the pre-trained models, the validation set to tune hyperparameters and prevent overfitting, and the test set to evaluate model performance. The models were trained using a batch size of 32 across 10 epochs. We employed early stopping, monitoring the validation loss. Training would halt if no loss improvement was observed over 3 consecutive epochs. The AdamW optimizer was utilized, configured with a learning rate of $2e-5$

Baselines			
Binary Class			
Features	**SVM**	**RndFor**	**LogReg**
TF-IDF	0.76	0.74	**0.81**
BOW	0.78	0.72	0.80
GVSent	0.47	0.52	0.46
Word2Vec	0.64	0.44	0.58
Fine-grained Class			
Features	**SVM**	**RndFor**	**LogReg**
TF-IDF	0.40	0.48	**0.54**
BOW	0.50	0.48	**0.54**
GVSent	0.23	0.27	0.29
Word2Vec	0.27	0.28	0.29

Table 2: F1-Scores for Experiments with Feature Based Models

Language Models		
	Binary	**Fine-grained**
BERT-simple	0.78	0.49
BERT-Multi	0.80	0.61
BERT-German	**0.86**	**0.65**
Roberta	0.82	0.58

Table 3: F1-Scores for Experiments with Deep Learning Models

6. Results

To evaluate the effectiveness of both our classical and transformer-based models, we computed several metrics including accuracy, precision, recall, and F1-score. The F1-scores for our experiments are presented in Table 2 and Table 3.

6.1. Binary Classification:

Using feature-based models, the best performance for the binary classification task was achieved with TF-IDF representations, closely followed by BOW. This indicates that count-based representations effectively capture distinguishing features between Real and Fake categories.

Transformer-based models, particularly BERT-German, outperformed feature-based models, highlighting their robust ability to discern Real from Fake News in German content. The detailed classification report reveals that the model is adept at identifying fake news instances (evident from a high recall) but occasionally misclassifies other content as fake news.

6.2. Fine-Grained Classification:

Feature-based models like TF-IDF and BOW exhibited satisfactory performance in the fine-grained classification task, albeit lower than their binary classification counterparts. This drop in performance is anticipated due to the intricate nature of distinguishing among numerous categories. A closer examination of the detailed classification report reveals that labels like 'catneutral' and 'typopinion' are predicted with higher precision and recall, suggesting these categories possess distinct features easily identifiable by the model. However, classes such as 'psychsensa', 'psychpolar', and 'authrefer', despite having ample instances, didn't fare as well. This might hint at these classes sharing overlapping features with others or being inherently more challenging to classify. Sparse classes, like 'typconspir' and 'psychabas', predictably struggled, emphasizing the challenges of classifying underrepresented categories.

Transformer-based models, especially BERT-German, continued to outpace feature-based models in the fine-grained classification task. However, a detailed label-wise analysis uncovers significant performance variance across labels. For instance, while labels like 'infonewinfo' and 'typfalcontex' were accurately predicted, others such as 'typpseudo' and 'psychemo' encountered difficulties. This discrepancy might arise from dominant labels overshadowing subtler ones in multi-label contexts.

6.3. Analysis and Discussion

The empirical results underscore the unparalleled advantages provided by language-specific models, such as BERT-German. Their adeptness at understanding linguistic intricacies, grammar, and vocabulary specific to the German language is pivotal. The timeless efficacy of TF-IDF and BOW representations was evident, even when combined with classical models. However, the sentiment scores from German Vader (GerVader) underperformed compared to other features. The brevity of tweets, often filled with slang and abbreviations, can impede accurate sentiment analysis. Tools like Vader provide generalized sentiment features, which may be inadequate for intricate tasks like fake news detection. Exploring sentiment computation using advanced language models might offer more nuanced insights.

Our results show that binary classification, while challenging, is simpler than fine-grained classification. Both feature-based and deep learning models exhibited superior performance in binary classification. This observation aligns with expectations, as discerning between two broad categories (Fake vs. Real) is intuitively simpler than distinguishing among 17 nuanced categories. The model has found challenges in categorizing them, as some classes might have overlapping features with others, making it hard for the model to distinguish between them. For example, 'psychpolar' and 'psychsensa' both deal with emotional or sensational content in the text. The potential overlap in their features might be causing misclassifications. Some labels might differ in very nuanced ways which are hard to capture with the given features. For instance, 'auth-refer' and 'auth vague' both deal with the authenticity of the content, but one might be about poor referencing, while the other is about vague claims. Capturing such subtle differences is challenging.

Incorporating external knowledge from knowledge graphs, ontologies, or trusted news databases is essential for validating claims and providing the necessary context, especially for aspects concerning authenticity and references. While models such as BERT-German have shown effectiveness, the integration of advanced Large Language Models (LLMs) can take this a step further. LLMs, renowned for their excellence in context learning and prompting-based techniques, can tap into their extensive linguistic capabilities and world knowledge to cross-reference and validate claims against established facts. By fine-tuning these models or employing precise prompts that reflect the context and intent of the content, LLMs become powerful tools for uncovering subtle disinformation cues that may bypass more traditional detection methods.

7. Linguistic Analysis

The word cloud representation in Figure 5 depicts the frequency of news topics within the tweets from our dataset, offering a glimpse into the most prominent themes and discussions within the German media. The size of each word indicates its frequency in the tweets, with larger words appearing more frequently.

Upon analyzing the textual content of the tweets, we notice that tweets classified as Real news tend to be slightly more extensive, both in terms of character length and word count, compared to fake news as depicted in Figure 6. This might suggest that Real News endeavors to provide more detailed and thorough information, possibly requiring addi-

Figure 5: Distribution of Topics in Dataset

tional words or incorporating URLs to convey accurate information. Conversely, a peak in character usage in fake news indicates that such posts might occasionally employ a more verbose narrative compared to Real news, potentially crafting a compelling, albeit deceptive, story line.

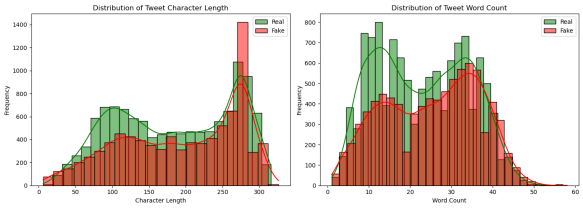

Figure 6: Textual Distribution in Real vs Fake news

8. Conclusion

In this research, we presented DeFaktS, a unique dataset tailored for disinformation analysis within the context of German political discussions on Twitter (now X) . Through a comprehensive annotation scheme, our dataset facilitates the precise identification and labeling of deceptive content. Beyond binary labels of Real and Fake News, DeFaktS incorporates fine-grained labels that signify polarized information in textual spans. Our experimental benchmarks, established using both traditional ML classifiers and state-of-the-art deep learning methods, highlight the efficacy of transformer-based models, especially the BERT-German variant, in discerning disinformation patterns. The insights derived from our study pave the way for further nuanced analysis and the development of more robust detection methodologies in the domain of disinformation. Overall, DeFaktS serves as a resource for the German media research community, promoting further exploration into refined analysis and detection techniques against disinformation.

9. Ethical Considerations and Limitations

Our research heavily relies on tweets, a publicly accessible form of data. While this data is public, ensuring the anonymity of the individuals and preventing potential misuse is paramount. All user data is kept separately on protected servers, linked to the raw text and network data solely through anonymous IDs. This precaution ensures that any personal information, such as user handles or profile details, is isolated from the research data, thereby respecting user privacy and safeguarding against potential breaches. It is important to note that conducting further analyses on Twitter (now X) data for future research endeavors is now limited to the greatly restricted access for researchers to data generated and distributed by the platform. Additionally, engaging human annotators for the labeling of data containing mentally and emotionally harmful content displays a challenge that researchers should handle responsibly. In the context of this project, to safeguard the annotators' well-being, different safety measures, such as group meetings and mood polls, were applied. While our research aims to detect and combat disinformation, there is potential for misuse. The tools and methods could be appropriated to suppress genuine information or target certain narratives. We emphasize that the primary goal is to detect disinformation and not to suppress freedom of expression.

10. Acknowledgements

This research work is part of the DeFaktS project, funded by the Federal Ministry of Education and Research (BMBF) on the basis of a decision of the German Bundestag. The project embarks on a comprehensive approach to the research and combat of disinformation by leveraging the capabilities of AI. The models are trained on messages extracted from potentially suspicious social media, enabling them to discern the unique characteristics and stylistic nuances of disinformation.

The leadership and expertise of the FZI Research Center for Information Technology, which spearheaded the consortium, have been instrumental to the project's success. Furthermore, we extend our gratitude to our project partners: Murmuras GmbH, Liquid Democracy e.V., and Philipps-Universität Marburg for their invaluable contributions, insights, and collaborative efforts throughout this research endeavor. This research was also

supported by the state of North Rhine-Westphalia as part of the Lamarr Institute for Machine Learning and Artificial Intelligence.

11. Bibliographical References

Hugo Queiroz Abonizio, Janaina Ignacio De Morais, Gabriel Marques Tavares, and Sylvio Barbon Junior. 2020. Language-independent fake news detection: English, portuguese, and spanish mutual features. *Future Internet*, 12(5):87.

Lucas Azevedo, Mathieu d'Aquin, Brian Davis, and Manel Zarrouk. 2021. Lux (linguistic aspects under examination): Discourse analysis for automatic fake news classification. In *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, pages 41–56. Association for Computational Linguistics.

Katarzyna Bąkowicz. 2019. Introduction to the definition and classification of the fake news. *Media Studies/Studia Medioznawcze*, 78(3).

Isabel Bezzaoui, Jonas Fegert, and Christof Weinhardt. 2022. Truth or fake? developing a taxonomical framework for the textual detection of online disinformation. *International journal on advances in internet technology*, 15(3/4):53.

Nathália Fraga Cardoso et al. 2021. Misconstruction of covid-19: an analysis of semantic concepts created by fake news.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Aaron Carl T Fernandez. 2019. Computing the linguistic-based cues of credible and not credible news in the philippines towards fake news detection.

Jacob Groshek and Karolina Koc-Michalska. 2017. Helping populism win? social media use, filter bubbles, and support for populist presidential candidates in the 2016 us election campaign. *Information, Communication & Society*, 20(9):1389–1407.

Mauricio Gruppi, Benjamin D Horne, and Sibel Adali. 2018. An exploration of unreliable news classification in brazil and the us. *arXiv preprint arXiv:1806.02875*.

Lukáš Havrlant and Vladik Kreinovich. 2017. A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation). *International Journal of General Systems*, 46(1):27–36.

Benjamin 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. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 759–766.

Caio Libanio Melo Jeronimo, Leandro Balby Marinho, Claudio EC Campelo, Adriano Veloso, and Allan Sales da Costa Melo. 2019. Fake news classification based on subjective language. In *Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services*, pages 15–24.

Eleni Kapantai, Androniki Christopoulou, Christos Berberidis, and Vassilios Peristeras. 2021. A systematic literature review on disinformation: Toward a unified taxonomical framework. *New media & society*, 23(5):1301–1326.

Douglas Kellner and Jeff Share. 2005. Toward critical media literacy: Core concepts, debates, organizations, and policy. *Discourse: Studies in the cultural politics of education*, 26(3):369–386.

Srijan Kumar, Robert West, and Jure Leskovec. 2016. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*, pages 591–602.

Jianing Li and Min-Hsin Su. 2020. Real talk about fake news: Identity language and disconnected networks of the us public's "fake news" discourse on twitter. *Social Media+ Society*, 6(2):2056305120916841.

Mohammad Mahyoob, Jeehaan Al-Garaady, and Musaad Alrahaili. 2020. Linguistic-based detection of fake news in social media. *Forthcoming, International Journal of English Linguistics*, 11(1).

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

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.

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.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Joan Plepi, Flora Sakketou, Henri-Jacques Geiss, and Lucie Flek. 2022. Temporal graph analysis of misinformation spreaders in social media. In *Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing*, pages 89–104.

Wisam A Qader, Musa M Ameen, and Bilal I Ahmed. 2019. An overview of bag of words; importance, implementation, applications, and challenges. In *2019 international engineering conference (IEC)*, pages 200–204. IEEE.

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.

Jose Fabio Ribeiro Bezerra. 2021. Content-based fake news classification through modified voting ensemble. *Journal of Information and Telecommunication*, 5(4):499–513.

Klaudia A Rosińska. 2021. Disinformation in poland: Thematic classification based on content analysis of fake news from 2019. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 15(4).

David Schreiber, Cristina Picus, David Fischinger, and Martin Boyer. 2021. The defalsif-ai project: protecting critical infrastructures against disinformation and fake news. *Elektrotechnik Und Informationstechnik*, 138(7):480.

Kai Shu, Amrita Bhattacharjee, Faisal Alatawi, Tahora H Nazer, Kaize Ding, Mansooreh Karami, and Huan Liu. 2020. Combating disinformation in a social media age. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(6):e1385.

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

Jesper Strömbäck. 2005. In search of a standard: Four models of democracy and their normative implications for journalism. *Journalism studies*, 6(3):331–345.

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.

Karsten Tymann, Matthias Lutz, Patrick Palsbröker, and Carsten Gips. 2019. Gervader-a german adaptation of the vader sentiment analysis tool for social media texts. In *LWDA*, pages 178–189.

Michela Del Vicario, Walter Quattrociocchi, Antonio Scala, and Fabiana Zollo. 2019. Polarization and fake news: Early warning of potential misinformation targets. *ACM Transactions on the Web (TWEB)*, 13(2):1–22.

Lipo Wang. 2005. *Support vector machines: theory and applications*, volume 177. Springer Science & Business Media.

Liqiang Wang, Yafang Wang, Gerard De Melo, and Gerhard Weikum. 2019. Understanding archetypes of fake news via fine-grained classification. *Social Network Analysis and Mining*, 9:1–17.

Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. 2020. Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 23–30. Association for Computational Linguistics.

Cheng Zhou, Kai Li, and Yanhong Lu. 2021. Linguistic characteristics and the dissemination of misinformation in social media: The moderating effect of information richness. *Information Processing & Management*, 58(6):102679.

12. Language Resource References

Abd Rahim, NH and Basri, MSH. 2022. *MalCov: Covid-19 Fake News Dataset in the Malay Language*. IEEE.

Ahuja, Nishtha and Kumar, Shailender. 2023. *MulFaD: attention based detection of multiLingual fake news*. Springer.

Ali, Sghaira Ben and Kechaou, Zied and Wali, Ali. 2022. *Arabic fake news detection in social media Based on AraBERT*. IEEE.

Carrella, Fabio and Miani, Alessandro and Lewandowsky, Stephan. 2023. *IRMA: the 335-million-word Italian coRpus for studying MisinformAtion*.

Ciora, Radu A and Cioca, Adriana L. 2022. *RoCoFake-A Romanian Covid-19 Fake News Dataset*. IEEE.

Fatima, Syeda Arooj and Zafar, Adeel and Malik, Khalid Mahmood. 2023. *YouFake: A Novel Multi-Modal Dataset for Fake News Classification*. IEEE.

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*.

Murayama, Taichi and Hisada, Shohei and Uehara, Makoto and Wakamiya, Shoko and Aramaki, Eiji. 2022. *Annotation-Scheme Reconstruction for" Fake News" and Japanese Fake News Dataset*.

Sakketou, Flora and Plepi, Joan and Cervero, Riccardo and Geiss, Henri-Jacques and Rosso, Paolo and Flek, Lucie. 2022. *Factoid: A new dataset for identifying misinformation spreaders and political bias*.

Suryavardan, S and Mishra, Shreyash and Patwa, Parth and Chakraborty, Megha and Rani, Anku and Reganti, Aishwarya and Chadha, Aman and Das, Amitava and Sheth, Amit and Chinnakotla, Manoj and others. 2023. *Factify 2: A multimodal fake news and satire news dataset*.

Vogel, Inna and Jiang, Peter. 2019. *Fake news detection with the new German dataset "GermanFakeNC"*. Springer.