Decoding Deception with TAXODIS – a Taxonomy of Disinformation Cues for Fine-grained Text Labeling

Isabel Bezzaoui [a,b,*], Pavlos Fafalios [c,d], Jonas Fegert [a,b], Konstantin Todorov [e] and Achim Rettinger [b,f]
[a] *KIT Karlsruhe Institute of Technology, Karlsruhe Germany*
E-mail: Bezzaoui@fzi.de
[b] *FZI, Research Center for Information Technology, Karlsruhe, Germany*
E-mail: fegert@fzi.de
[c] *Technical University of Crete, Chania Greece*
[d] *Institute of Computer Science, FORTH-ICS, Heraklion Greece*
E-mail: fafalios@ics.forth.gr
[e] *LIRMM, CNRS, University of Montpellier, Montpellier France*
E-mail: todorov@lirmm.fr
[f] *Trier University, Trier Germany*
E-mail: rettinger@uni-trier.de

Abstract. The ubiquity of disinformation on digital platforms poses a threat to democracy and social cohesion. Despite significant developments in machine learning for disinformation detection and more specific related tasks (such as fact-checking, check-worthiness detection, claim linking, propaganda and rumor detection), effectively applying empirical knowledge during the training of such models in a standardized and transparent way remains a challenge. In this paper, following the semantic web principles, we propose TAXODIS—the first of its kind openly available Taxonomy of Online Disinformation. It structures an interdisciplinary set of well-defined and analyzed linguistic features of online disinformation discourse and is meant to help annotate training data to nourish machine learning and computational models that deal with the above-mentioned tasks. The systematic clustering of linguistic features into a comprehensive and publicly available framework provides a basis for the empirically grounded training of models and enhances the understanding of disinformation on a textual and linguistic level. Demonstrating and evaluating the artifact, we find that it facilitates data labeling processes by offering annotators a compact yet empirically informed guide to identifying textual indicators of disinformation. This paper, proposing a structured taxonomy as a valuable tool for automated detection systems, contributes to disinformation detection by mapping nuanced linguistic characteristics in disinformation content.

Keywords: Disinformation Detection, Fact-checking, Data Annotation, Taxonomy, NLP models training

1. Introduction

As today's primary news sources, social media and news platforms suffer from inaccurate reporting and the distribution of unfounded opinions [69]. Especially in times of crises, the viral spread of disinformation poses a

*Corresponding author. E-mail: Bezzaoui@fzi.de.

central threat to political processes and social cohesion as the United Nations recently addressed in their disinformation report [59]. Disinformation is defined as false information and, unlike misinformation and malinformation [77], is spread with the intention to deceive [70]. Therefore, automated systems detecting disinformation on digital platforms are indispensable tools in the ongoing effort to maintain the integrity of information, protect democratic processes, and foster a more informed and cohesive society.

Research on disinformation detection using machine learning (ML) and natural language processing (NLP) is a rapidly expanding field that spans various disciplines, including computer science, social science, psychology, and information systems [11, 47, 81]. Most techniques focus on extracting multiple features, incorporating them into classification models, and then choosing the best classifier based on performance [6, 16]. Data suggests that disinformation content is difficult to identify [37] due to the variety of stylistic devices used in disinformation, creating a barrier for purely quantitative approaches to the problem [67]. The deceptive nature of disinformation, where the aim is to make the information appear to be authentic, may help to explain this difficulty [1]. Nevertheless, empirical evidence on the structure of disinformation demonstrates that legitimate and deceptive content differ significantly in their substance and sentiment [32, 35].

Thus, recognizing the need for a comprehensive understanding, this research delves into the clustering of linguistic features, creating a robust foundation for the empirical training of detection models. Accordingly, we are guided by the following research question: *How can a taxonomy of online disinformation characteristics be designed to support text classification and other downstream tasks in mis- and disinformation analysis, and be made available to facilitate the automated detection of disinformation?* In doing so, we aim to contribute to a shared understanding of disinformation at a linguistic level, providing a nuanced perspective that goes beyond conventional binary detection methodologies.

The focal point of this paper is the development, implementation, and demonstration of the Taxonomy of Online Disinformation (TAXODIS). We ground the implementation of this taxonomy in the principles and technology of the Semantic Web, with its structure represented as a SKOS thesaurus to ensure interoperability and reuse. Additionally, we show how the taxonomy can be used together with existing well-established vocabularies for the annotation of disinformation resources, mainly the Open Annotation Data Model (W3C Recommendation) and schema.org, and how the annotations can be then linked to existing knowledge bases such as Wikidata.

A well-structured semantic taxonomy of online disinformation can serve as a foundation for automated detection systems, providing scientific guidelines for more fine-grained annotation of disinformation datasets. Such annotated datasets can then be used to train classifiers and/or be published as Linked Data, enabling potential integration with other knowledge graphs.

The taxonomy builds on two earlier works: [13] outlines the methodology for taxonomy development, while [9] presents the DeFaktS dataset, demonstrating how the taxonomy can be operationalized for annotation purposes. The latter also presents an earlier (unimplemented) version of the taxonomy.

The paper is structured as follows. In Section 2, we review related work, before giving an overview of TAXODIS in Section 3. The methodology of building the taxonomy is given in Section 4, while examples of using and linking the resources to existing knowledge graphs are provided in Section 5. Several use case scenarios are presented in Section 6, before we conclude in Section 7.

2. Related Work

Recent research addresses both the benefits and drawbacks of different detection methods as well as their underlying theories [7, 66, 83]. Nevertheless, many disinformation classifiers presented in empirical papers lack explanations on how they were trained or how the datasets used for training were labeled [2, 23, 42]. Although these explanations are crucial to the transparency and traceability of the research process, only little research has accounted for this issue [53, 55]. Creating a succinct taxonomy that covers the wide-ranging attributes of disinformation regardless of the specific event while also being detailed enough to precisely categorize deceptive content may enhance the transparency of the manual classification process of disinformation datasets.

In the past years, there have been various endeavors to capture the phenomenon of disinformation with taxonomical frameworks. Alexander and Smith [4] base their approach to taxonomy development on a communication

model to illustrate how disinformation is spread to deceive its audiences. While they discuss illustrative examples of different strategies for modifying or distorting messages to subvert their initial meaning, the authors do not suggest a concise taxonomy providing a structured overview of indicators that help identify disinformation in social media. Tambini [72], on the other hand, provides generic categories that lead to overlapping definitions. The proposed categories encompass a wide range of sociopolitical phenomena such as 'falsehood to affect election results' and 'news that challenges orthodox authority'. These aspects primarily serve a descriptive rather than explanatory purpose, implying a need for more precision in classification. Parikh and Atrey [61] delineate disinformation features by relying on technical attributes or the structural format of news items. These categories encompass visual elements such as photoshopped images, user-based components involving fake accounts, and style-based aspects, among others. Their technical approach primarily introduces types of data in news, disinformation detection methods, and common disinformation datasets. While this approach proves valuable for developing automated detection tools, its technical orientation poses challenges when attempting to integrate it with broader frameworks equally focused on non-technical aspects of disinformation.

In adopting a detection-oriented approach to the issue, Kumar and Shah [39] present four broad categories: opinion-based, fact-based, misinformation, and disinformation, without delving into the finer nuances of the domain, such as clickbait, propaganda, and trolling. Their focus is limited to specific domains and they position the terms 'disinformation' and 'misinformation' at a more granular level, in contrast to the common practice of treating them as overarching umbrella terms. In their taxonomy, Lemieux and Smith [44] categorize disinformation and misinformation alongside more specific phenomena like hoaxes and rumors, placing them at a similar hierarchical level. Furthermore, they introduce the term 'mal-information' as an overarching category, on par with disinformation and misinformation. This approach makes it difficult to assign subphenomena, such as conspiracy theories, to overarching phenomena, such as disinformation. Molina et al. [55] differentiate various types of disinformation by employing four operational indicators: message, source, structure, and network. This approach extends beyond content-based methods and conventional definitions, instead centering on the dissemination of online information and offering insights into potential detection solutions. Their study provides an extensive overview of the characteristics of fabricated news. However, the proposed taxonomy lacks concision, resulting in nine extensive tables that are neither precise nor concise enough for handling large amounts of data [60]. Kapantai et al. [37] have designed a succinct taxonomy framework characterized by three fundamental dimensions: motive, facticity, and verifiability. These dimensions and their associated metrics prove crucial in the categorization of disinformation that has been previously identified as such, enabling differentiation between specific manifestations such as clickbait, trolling, and fake reviews. It is essential to note, however, that this taxonomy does not furnish discernible indicators intended to facilitate the proactive identification of disinformation content by human users. Finally, the DISARM framework provides an overview of several sub-frameworks for practitioners to describe and understand different parts of disinformation, including its actors, tactics, and countermeasures. While the framework is intended to help track and counter misinformation [21], it does not provide a hands-on and scientifically grounded scheme that can be applied to the recognition of disinformation via granular features and characteristics referring to language and content.

None of the mentioned efforts above propose a shared semantic model that would help lead toward a uniform and common understanding of the various categories of features. In that respect, several structured datasets with schema have been proposed to deal with the specific task of fact-checking or disinformation detection. The MultiFC [10] and the ClaimsKG [27, 28, 74] datasets both provide structured data of and about claims coming from established fact-checking portals, where claims are stored together with contextual metadata (such as authors, sources, claim reviews and other contextual information, including veracity labels). The two datasets are complementary in some respects. MultiFC focuses on evidence-based fact-checking in terms of downstream tasks, where via the Google Search API the ten most highly ranked search results per claim are retrieved and stored. ClaimsKG, on the other hand, provides a rich data model (an RDFS ontology) to represent check-worthy or fact-checked claims and related metadata, which is an important effort towards standardization and enables federated access to distributed data, where a specific search engine is provided in addition to a public Sparql endpoint [29]. MultiFC contains data in English, while ClaimsKG is multilingual, harvesting data from fact-checking portals in about ten languages. These datasets can be used to provide a pool of verified claims with additional metadata for fact-checking applications and to extract links to claims that are mentioned in fact-checking articles. However, they do not delve into the problem and nature of the linguistic and textual features that define disinformation.

In these terms, an important effort for annotating text with general linguistic features is the Linguistic Inquiry and Word Count tool (LIWC). LIWC is a gold standard for word-level text analysis, which has been used in large amounts of scientific publications.[1] It has also proven to be well-suited for web claim-related tasks (e.g., [51] ranked 2nd at the CheckThat! 2022 Fake News Detection Challenge and used LIWC in their pipeline). LIWC extracts features by using over 100 built-in dictionaries that encompass social and psychological states, emotional tones, linguistic properties, cognition processes, analytic speech patterns, punctuation marks, and several word-count-related features. Each dictionary can contain a list of words, a list of word stems, emoticons and other specific word constructions. The LIWC features can be divided into seven distinct categories: syntactic, analytic, sentiment, social, perceptual, informal language, and topic. However, although useful in claim-related analyses for fake news detection, LIWC has a more general focus. A specific subset of its features can be used to annotate disinformation-related data, but this selection has to be made manually, where this is additionally hindered by the fact that the vocabulary is not formally structured and queryable. In addition, access to LIWC is granted upon request, making it less easy to apply, as it is not openly available. In contrast, the proposed taxonomy in this paper is tailored to disinformation in particular, contains more specific and fine-grained categories and types of features for related downstream tasks, in addition to it being fully open and structured following the semantic web principles.

The current state of the art shows that what is missing so far is a fundamental but concise empirical overview of linguistic detection cues supporting the creation of labels for transparently annotating datasets on a granular level. By implementing a taxonomy encompassing such an overview, a classifier not only produces an output providing indications of content veracity but also furnishes more comprehensive information about prevalent characteristics in disinformation. The novel taxonomy is shaped and made openly available as a (SKOS-based) RDFS resource, which enhances re-usability, interoperability and FAIRness in general, with advantages such as easy access and federated queries over the vocabulary and the annotated datasets. Finally, this approach aims to enhance digital literacy among both annotators and end-users of the developed classifier.

While elements of the taxonomy have been discussed in earlier work, we emphasize that neither the work-in-progress taxonomy paper [13] nor the dataset-focused DeFaktS paper [9] present the final taxonomy as introduced and formalized in this manuscript. The earlier paper [13] primarily outlined the methodology for taxonomy development and provided illustrative examples, but did not introduce a finalized or structured resource. The DeFaktS dataset paper [9] demonstrates how the taxonomy can be operationalized for annotation purposes but does not describe or analyze the taxonomy itself in detail, nor does it provide an implementation of it. This manuscript is the first to consolidate, refine, and formally present the finalized taxonomy as a semantic resource, with specific attention to its structure, categories, and features. It also introduces the RDFS implementation following semantic web principles, which is novel and central to the FAIRness and reusability of the resource. Moreover, the taxonomy itself has been extended since earlier work: it now includes a sixth dimension, developed through additional conceptual work and literature review, along with updates to existing categories to improve clarity, coverage, and applicability.

3. Taxonomy Overview and Open Availability

Figure 1 depicts the TAXODIS taxonomy. The taxonomy contains (currently) 66 concepts, of which 48 are ‘leaf’ concepts (concepts with no narrower terms), organized in a hierarchical (tree-like) structure of maximum depth four. Its top concept is ‘disinformation characteristic’, which describes characteristics that are indicative of disinformation in a piece of content. This top term has three narrower terms: i) ‘detection feature’, which classifies the piece of content based on linguistic or stylistic features that are indicative of the detection of disinformation (e.g. length of the headline, lexical and contentual poorness, level of semantic incoherence, lack of new information, level of topicality, etc.), ii) ‘categorization’, which classifies the piece of content based on its theme or content type. e.g. social (theme), conspiracy theory (content type), and iii) ‘veracity’, which classifies the piece of content based on its veracity, e.g. mostly false, mixture, etc. A detailed explanation of the narrower terms of these three broad terms is provided in the next section.

[1] See https://www.liwc.app

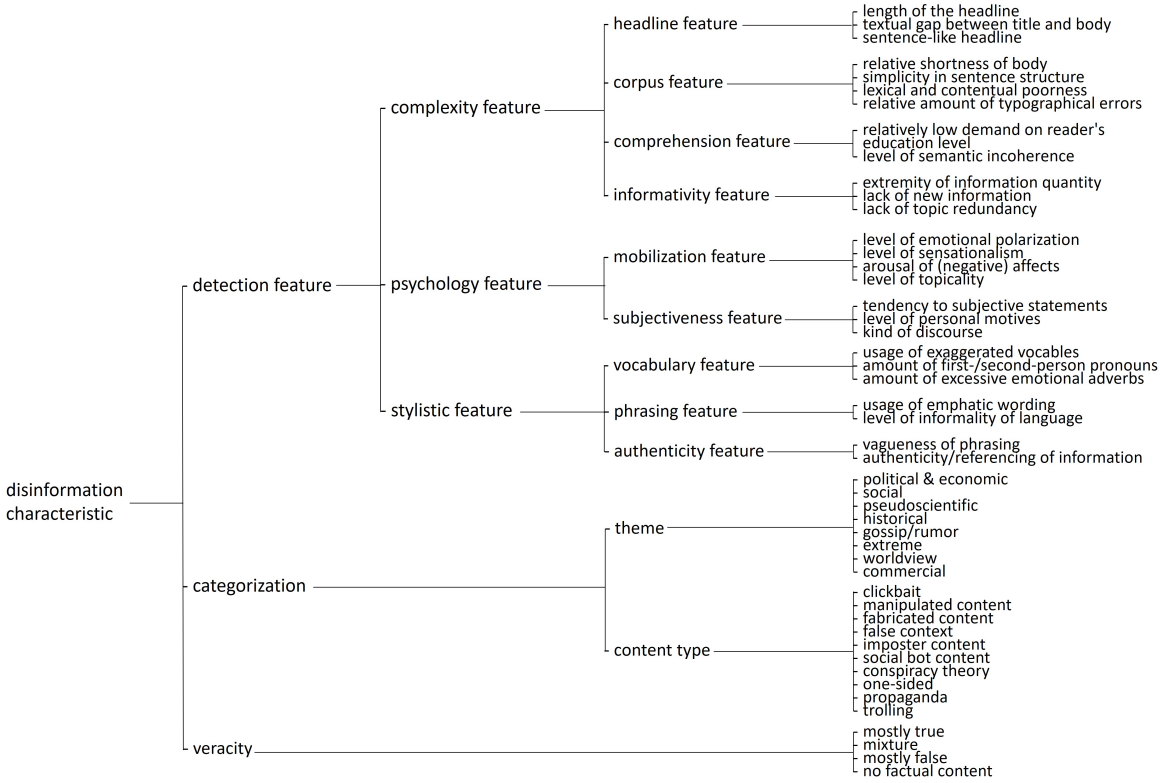

Fig. 1. The TAXODIS taxonomy.

We implemented the TAXODIS taxonomy as a SKOS vocabulary/thesaurus. SKOS[2] is a data model designed for the representation of thesauri, classification schemes, taxonomies, and other types of controlled vocabularies. It is a W3C recommendation built upon RDF and RDFS, and its main objective is to enable easy publication and use of controlled vocabularies across the web. The SKOS representation of TAXODIS provides for each term/concept: i) its preferred label in English (using the property *skos:prefLabel*), ii) its definition in English (using the property *skos:definition*), iii) its broader terms, if any (using the property *skos:broader*), iv) its narrower terms, if any (using the property *skos:narrower*), v) its notation, used to uniquely identify the term within the scope of a given concept scheme (using the property *skos:notation*), vi) the scheme (vocabulary/thesaurus) in which the term belongs to (using the property *skos:inScheme*).

We also provide the following metadata using properties of RDFS, DCMT (Dublin Core Metadata Terms) and other widely-used vocabularies: i) the title of the taxonomy (using the properties *rdfs:label* and *dct:title*), ii) the description of the taxonomy (using the properties *rdfs:comment* and *dct:description*), iii) the taxonomy's usage license (using the properties *dct:license* and *cc:license*), iv) the taxonomy's creation date (using the property *dct:issued*), v) the taxonomy's last modification date (using the property *dct:modified*), vi) the taxonomy's version (using the properties *owl:versionInfo* and *owl:versionIRI*), vii) the creators of the taxonomy (using the property *dct:creator*), viii) the taxonomy's namespace URI (using the property *vann:preferredNamespaceUri*), and ix) the taxonomy's namespace prefix (using the property *vann:preferredNamespacePrefix*).

The RDFS file (in Turtle format) of the SKOS implemenation of TAXODIS is publicly available under a creative commons license at: https://zenodo.org/records/14264593 (DOI: https://doi.org/10.5281/zenodo.14264593). The (resolvable) namespace of the taxonomy is https://hop.fzi.de/taxodis/.

[2]https://www.w3.org/2004/02/skos/

4. Building TAXODIS, the Taxonomy of Online Disinformation

This section outlines the methodology for constructing the taxonomy and describes its features, as initially introduced in [13] and [9]. It further provides an illustrative example and discusses annotation challenges.

4.1. Methodology

Our iterative approach consists of two major parts, integrating insights from multiple disciplines to construct a robust taxonomy. Initially, by conducting a systematic literature review [79], we gather a comprehensive range of linguistic features of online disinformation from various fields of study. This allows us to capture diverse perspectives on how disinformation manifests across different contexts. Subsequently, we cluster the empirical results in groups, supporting a linguistic-based disinformation detection approach. Categorizing objects aids in understanding and analyzing complex environments, making the creation of taxonomies essential for research and development [60]. Nickerson et al. [60] provided the first and well-conceived taxonomy-building methodology. Their approach has served as a blueprint of numerous taxonomy projects across various domains [41]. Building on these interdisciplinary foundations, we propose a novel six-dimensional taxonomy, based on the categorization criteria identified from the existing empirical literature.

4.1.1. Systematic Literature Review

We have conducted a systematic literature review following Webster and Watson's [79] methodological guidelines. A thorough review encompasses pertinent literature on the subject and is not confined to a particular research approach, set of journals, or geographical area [79]. Hence, we utilized large interdisciplinary databases to access all relevant research fields for our project. Upon careful examination of the literature concerning linguistic features and disinformation detection characteristics, we synthesized an overview of frequently used descriptions referring to various types and characteristics of disinformation content. However, the ad hoc definitions introduced by each study may give rise to conflicts or overlaps. Accordingly, the overarching objective of our literature review is to consolidate the existing knowledge on categorizing disinformation and to discern patterns and key concepts within the literature. Our aim is to advance prior research by synthesizing this knowledge into a cohesive taxonomy.

To achieve this goal, we followed a structured procedure for our review: Initially, we identified our sources from digital libraries and defined our search terms, which were subsequently applied to the selected sources. Afterward, we refined our selection of primary studies by employing inclusion and exclusion criteria on the search results. To further enhance the comprehensiveness of our review, we conducted both backward and forward searches based on the selected primary studies. An automated search was executed across five prominent scientific databases to identify relevant publications: IEEE Xplore Digital Library, Scopus, ACM Digital Library, Web of Science, and Springer Link. Initially, we conducted several pilot searches on our research topics to compile a preliminary list of papers. Based on these searches, we defined search terms that aligned with our research objectives. The selected search phrases, limited to abstract and title, were as follows: linguistic 'disinformation' OR 'fake news' AND 'classification' OR 'detection'.

For the next phase of our research, the following three inclusion and exclusion criteria were formulated: We excluded sources that solely address the issue of disinformation from a computational perspective, advocating technical solutions reliant on machine learning and statistical models to automatically categorize news articles into predefined categories, such as fake or real. Additionally, we omitted sources that primarily conducted performance evaluations of such models. Publications that mention specific categories or characteristics of false information without attempting systematic classification or providing explanations for the proposed categories were excluded. This criterion was applied to sources where the disinformation phenomenon is not a central concept, such as papers that incidentally use terms like 'fake news', or those that discuss specific types of false information without integrating them into a comprehensive framework, rendering them non-exhaustive or merely indicative. In the interest of promoting common scientific understanding, only papers written in English were included in our review. Our search yielded 29 primary studies across six different disciplines (e.g., computer science, linguistics, psychology, and media studies) introducing linguistic frameworks for disinformation detection. The selection process encompassed records obtained through database searching as well as those identified through additional backward and forward searches based on the initial records.

Figure 2 provides a detailed overview of the selection process (as a PRISMA flow diagram[3]), encompassing records obtained through database searching as well as those identified through additional backward and forward searches based on the initial records. In total, 34 papers were included in our review. Our initial objective was to

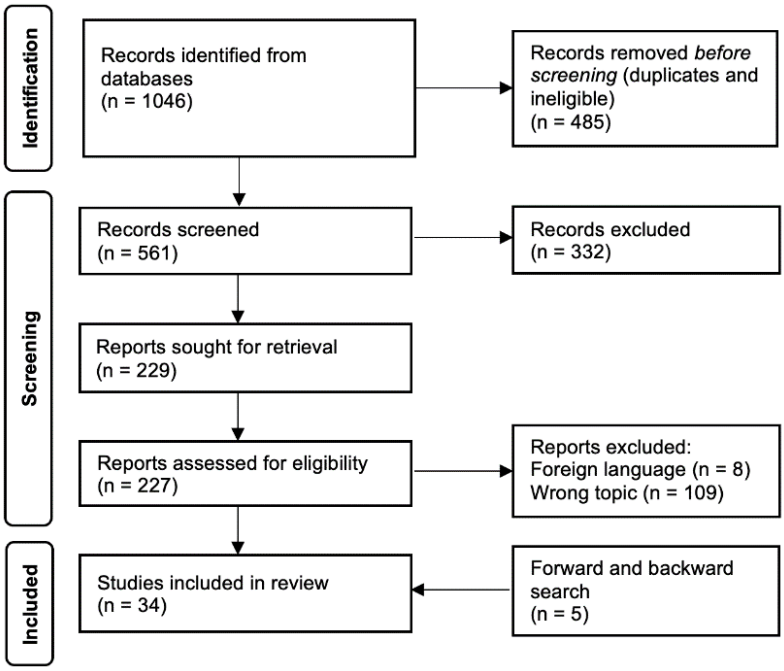

Fig. 2. PRISMA flow diagram.

identify linguistic-based cues of online disinformation in the empirical literature. Subsequently, we extracted the identified features of disinformation and organized them into clusters based on similarity to prepare our findings.

4.1.2. TAXODIS' Features

Our overall goal is to create a taxonomy of online disinformation that helps create a common understanding of what constitutes disinformation from a linguistic viewpoint, provides a list of categories and detection characteristics and can be used to develop labels that can be applied to diverse datasets. Based on the findings of our systematic literature review, we organized the identified features into a more fine-grained schema by clustering them according to their similarities. We observed many commonalities but also differences at both the category and dimension levels. In order to make sense of the patterns and contradictions, we applied several general rules during the processing of the data. First, we removed types and definitions that are either too generic (e.g., yellow press) or too technical (e.g., deep fakes). Second, we removed duplicates and synonyms to avoid repetitions and overlaps. Lastly, any types and definitions that were incorrectly categorized as disinformation (e.g., misinformation) were removed. After our fifth iteration, we did not identify any new characteristics and dimensions from the reviewed studies.

Our final framework (Table 1) consists of six dimensions, since, in our case, all ending conditions [60] were satisfied. The first dimension covers complexity features (1) that help to evaluate the complexity and readability of the text, splitting into headline, corpus, comprehension, and informativity. It allows TAXODIS users to evaluate the informational content and textual structure of the content under consideration. Our second dimension contains psychology features (2) that describe attitudes, behaviors, and emotions. This dimension, which splits into mobilization and subjectiveness, aids in illuminating and quantifying the cognitive process and individual concerns that underlie the writings. We added a third dimension, stylistic features (3), to reflect the writer's style and the syntax of the text, such as the number of verbs and nouns used, as well as the use of specific terminologies. This dimension splits into

[3]https://www.prisma-statement.org/prisma-2020-flow-diagram

vocabulary, phrasing, and authenticity. The fourth and fifth dimensions help to categorize disinformation content, as themes (4) contain categories such as 'pseudoscientific' or 'historical', and content type (5) allows differentiating between different types of content. Moreover, disinformation content can differ strongly in its deceitfulness. For this reason, our last dimension accommodates grades of veracity (6) to facilitate the evaluation of different kinds of disinformation corresponding with our fifth dimension. Below we provide details for the individual features.

Complexity Features. *Headline.* Unreliable sources try to convey as much information as possible in the title to draw the reader's attention. Thus, they use a higher amount of plain text or words in the headline [30] and often display a lower textual similarity between the title and the body of the article [14]. Titles of fake content often present sentence-like claims about people and entities associating them with actions [24, 35].

Corpus. Unreliable sources tend to have a lower level of plain text or number of words in relation to real articles [40], and their sentences exhibit a lower complexity in structure and a relatively low amount of words [30, 35]. Deceptive articles tend to have less diversity at the lexical and content level [11] and empirically exhibit a higher amount of typographical errors [82].

Comprehension. Reliable articles tend to be written in a more complex way regarding readability measures like the number of words per sentence. A higher complexity increases the effort to apprehend the content and thus works as an indicator for the demanded education level of the reader [75]. Real articles contain sentences that are correlated with each other while unreliable sources exhibit a lower level of sentence or word correlations defined as semantic incoherence [12].

Informativity. Articles containing false information often correspond with either a considerably low amount of information or a remarkable overload of information [83]. The body of such articles adds relatively little new information but serves to repeat and enhance the claims made in the title [11, 35]. Valid articles about a particular topic contain several direct or indirect references to this subject. One can interpret those as a kind of contextual redundancy which fake sources are usually missing [12].

Psychology Features. *Mobilization.* Unreliable sources tend to use more emotionally persuasive language in general, leading to high levels of emotional polarization [65, 76]. Providing sensationalist content, articles containing false information tend to be written in a hyperbolic way to attract the reader's attention, i.e., with high usage of all-caps-words or exclamation marks [30, 36]. To cause an arousal of (negative) affects fake content uses a higher degree of words related to emotional actions, states, and processes [11, 50]. Legitimate sources tend to report about past events, whereas disinformation focuses on highly recent topics [24].

Subjectiveness. Exhibiting a tendency to subjective statements, deceptive articles are often written from a more personal view [36]. Creators of fake content are frequently driven by personal motives like raising profit, promoting ideology, or psychological aims [37]. Words and expressions of manipulative articles relate to a more argumentative discourse aiming to convince the reader of a specific point of view [11].

Stylistic Features. *Vocabulary.* Unreliable sources more often use hyperbolic words such as superlatives and subjectives [24, 47] and display more first-person and second-person pronouns than legitimate articles [24, 63]. To lure readers to the content, disinformation displays a higher amount of excessive emotional adverbs [14, 47].

Phrasing. Unreliable sources use a high level of exclamation marks, swear words, visual references, and are slightly more prone to emotional tones and higher polarity [11, 65]. The language of fake content tends to be less formal than reliable articles [35].

Authenticity. Disinformation use a higher amount of vague phrasing or hedging words to achieve a more indirect form of expression [47] while legitimate sources are considerably better referenced than unreliable articles [40].

Themes. The category 'political and economic' refers to content about specific politicians, or legal, political or economic actions. Content about social events, activists, public benefit and minority organizations, as well as dangers or threats to human and animal health is incorporated in the 'social' category. Pseudoscientific content calls on supposedly scientific research or reputable institutions without identifying concrete sources or by manipulating them to create a false theory. Content about historical events or the distant past of public figures is subsumed under the theme 'historical'. In addition to that, gossip or rumors may be spread about public figures without a political or activist profile. Extreme themes cover drastic, catastrophic or brutal events. The feature 'worldview' is applied to content about religion, faith, and spiritual figures as well as various non-religious ideologies, views, and beliefs.

Table 1
The TAXODIS taxonomy of online disinformation.

meta-characteristic	dimension	subdimension	feature	code	characteristic value	
					0	**1**
detection	complexity features	headline	length of the headline	headlength	low	high
			textual gap between title and body	headgap	no	yes
			sentence-like headline	headsent	no	yes
		corpus	relative shortness of body	corpshort	no	yes
			simplicity in sentence structure	corpsimpl	no	yes
			lexical and contentual poorness	corplex	no	yes
			relative amount of typographical errors	corperror	low	high
		comprehension	relatively low demand on reader's education level	compeduc	no	yes
			level of semantic incoherence	compincoh	low	high
		informativity	extremity of information quantity	infoextrem	low	high
			lack of new information	infonewinfo	no	yes
			lack of topical redundancy	infotopredun	no	yes
	psychology features	mobilization	level of emotional polarization	mobpolar	low	high
			level of sensationalism	mobsensat	low	high
			arousal of (negative) affects	mobaffect	low	high
			level of topicality	mobtopical	low	high
		subjectiveness	tendency to subjective statements	subjtenden	low	high
			level of personal motives	subjmotiv	low	high
			kind of discourse	subjdiscours	knowledge-based	opinion-based
	stylistic features	vocabulary	usage of exaggerated vocables	vocexagg	no	yes
			amount of first-/second-person pronouns	vocpronoun	low	high
			amount of excessive emotional adverbs	vocadverb	low	high
		phrasing	usage of emphatic wording	phrasemph	no	yes
			level of informality of language	prhasinformal	low	high
		authenticity	vagueness of phrasing	authvague	low	high
			authenticity/referencing of information	authrefer	frequently referenced	poorly referenced
categorization	themes	political & economic		thempoleco	no	yes
		social		themsoc	no	yes
		pseudoscientific		themscience	no	yes
		historical		themhisto	no	yes
		gossip/rumor		themgoss	no	yes
		extreme		themextrem	no	yes
		worldview		themworld	no	yes
		commercial		themcommer	no	yes
	content type	clickbait		typclick	no	yes
		manipulated content		typmanipul	no	yes
		fabricated content		typfabric	no	yes
		false context		typfalse	no	yes
		imposter content		typimpost	no	yes
		social bot content		typbot	no	yes
		conspiracy theory		typconspir	no	yes
		one-sided		typonesid	no	yes
		propaganda		typpropa	no	yes
		trolling		typtroll	no	yes
veracity grade	mostly true			vtrue	no	yes
	mixture of true and false			vtruefalse	no	yes
	mostly false			vfalse	no	yes
	no factual content			vnofact	no	yes

Themes can also be commercial, such as false product reviews, advertising campaigns or commercial clickbait aimed at accumulating views, likes and comments [67].

Content Type. Clickbait refers to sources that intentionally use exaggerated, misleading, or unverified headlines or thumbnails to attract readers to open the webpage [37]. Manipulated content involves altering information or an image to deceive the recipient, who receives it without being aware of its misuse [78]. Fabricated content encompasses entirely false stories lacking factual basis, with the intent to deceive and cause harm. Particularly severe forms of fabrication mimic the style of legitimate news articles to mislead recipients [37]. Real information may be presented in a false context, where the recipient acknowledges its truth but remains unaware that the context has been altered [78]. Imposter content involves genuine sources being impersonated by false, made-up sources to support a false narrative. This can include abusing a journalist's name, a logo, or a website [37]. A social bot is a computer algorithm that automatically produces and posts content, interacting with legitimate users and other bots to emulate and possibly alter their behavior [25, 76]. Conspiracy theory applies to stories without a factual basis that usually explain important events as secret plots by governments or powerful groups or individuals [37]. One-sided content is heavily biased, promoting division and polarization. It features imbalance, inflammatory, and emotionally charged information, often containing a mix of true and false or mostly false details [37]. Propaganda is information created by a political entity to influence public opinion and gain support for a public figure, organization or government [73]. Trolling is the intentional posting of offensive or inflammatory content to an online community with the intent of provoking readers or disrupting conversation [37].

Grade of Veracity. Following Potthast et al. [62], 'mostly true' indicates that a piece of content is based on factual information and accurately depicts it. This rating excludes unsupported speculation or claims. 'Mixture of true and false' describes content with some accurate and some inaccurate elements. It applies when speculation or unfounded claims are combined with real events, numbers, or quotes. 'Mostly false' is used when the majority or all of the information in a content piece is inaccurate. This rating also applies when the central claim is false. 'No factual content' is for posts expressing pure opinion, comics, satire, or anything without a factual claim. This adopted gradation follows a similar approach to knowledge graph 'ClaimsKG' [74], where the different veracity labels are mapped to four basic categories (i.e., true claims, false claims, mixture claims, other claims).

4.2. An Example

Consider the article published on *Before It's News* entitled *"RFK Jr: Fauci Must Be Prosecuted for 330K Murders, As Mass Graves Found Outside NYC (Video)"*.[4] This article has the following values on the TAXODIS detection features (manually annotated): length of the headline = high, textual gap between title and body = no, sentence-like headline = yes, relative shortness of body = yes, simplicity in sentence structure = no, lexical and contentual poorness = yes, relative amount of typographical errors = yes, relatively low demand on reader's education level = yes, level of semantic incoherence = high, extremity of information quantity = high, lack of new information = yes, lack of topical redundancy = yes, level of emotional polarization = high, level of sensationalism = high, arousal of (negative) affects = high, level of topicality = high, tendency to subjective statements = low, level of personal motives = high, kind of discourse = opinion-based, usage of exaggerated vocables = yes, amount of first-/second-person pronouns = high, amount of excessive emotional adverbs = high, usage of emphatic wording = yes, level of informality of language = high, vagueness of phrasing = high, authenticity/referencing of information = poorly referenced. As regards the categorization features, the article falls under the themes 'political & economic' and 'extreme', and the content type 'fabricated content', while its veracity grade is 'mostly false'.

We recognize that some categories in TAXODIS, such as "emotional polarization" versus "sensationalism", exhibit conceptual proximity that may challenge consistent annotation. In such cases, multiple labels to a single content item can be assigned, even within the same dimension where categories are not strictly mutually exclusive. For example, content may be simultaneously labeled as both "clickbait" and "propaganda" when relevant. This multilabel approach reflects the complex and often overlapping nature of disinformation phenomena. For this reason, the

[4] https://beforeitsnews.com/alternative/2024/09/rfk-jr-fauci-must-be-prosecuted-for-330k-murders-as-mass-graves-found-outside-nyc-video-3821353.html (accessed on October 30, 2024)

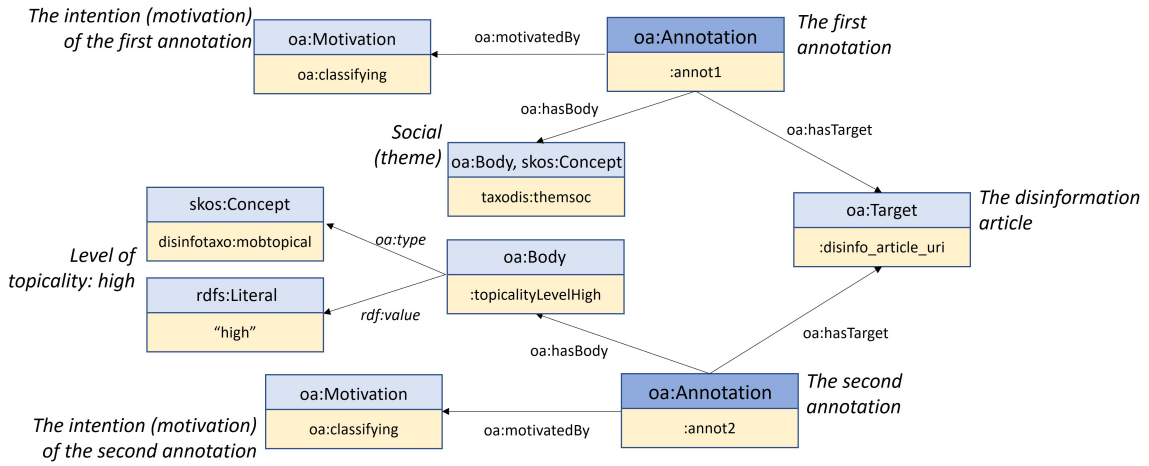

Fig. 3. An annotation example using TAXODIS together with the Open Annotation Data Model in which an article is categorized as of social theme and as having a high topicality level.

precise annotation protocols should be designed and shared with annotators according to the specific task at hand, where attention should be paid to clearly defining the task and giving examples of annotated text, including difficult cases, which may help disambiguate.

5. Taxonomy Usage and Linking to Related Vocabularies

The taxonomy can be used together with existing, established vocabularies for the annotation of (disinformation) resources. We suggest the exploitation of the Web Annotation Data Model,[5] which is a W3C recommendation for the structured representation of annotations that can be shared and reused across different platforms. In this model, an annotation (instance of class oa:Annotation) is considered to be a set of connected resources, typically including a *body* (instance of class oa:Body) and a *target* (instance of class oa:Target), and conveys that the body is related to the target. The exact nature of this relationship changes according to the intention of the annotation, but the body is most frequently somehow "about" the target (the intention of the annotation can be represented using the class oa:Motivation). In our case, the body of the annotation is a taxonomy term, accompanied by a value (level or degree) for the terms that are under 'detection feature', and the target is a disinformation piece of content or resource.

Figure 3 shows an example in which an article (instance of class oa:Target) is linked to two annotations: one which categorises the article as of social theme (taxodis:themsoc) and one which categorises the article as having 'high' topicality level (taxodis:mobtopical). The intension (motivation) of both annotations is classification (oa:classifying). Notice that the first annotation is directly linked to the taxonomy term taxodis:themsoc through multiple instantiation (the term is an instance of both oa:Body and skos:Concept). This annotation method can be applied for all taxonomy terms that are under 'categorisation' and 'veracity', since these terms do not accept a degree value or level like the terms that are under 'detection feature'.

Figure 4 shows how we can link the annotated resource with rich (meta)data using another established vocabulary, namely schema.org. Schema.org[6] is a collaborative, community activity with a mission to create, maintain, and promote schemas/vocabularies for structured data on the web. It provides classes and properties for embedding structured data to web resources. In the example of Figure 4, the annotated article is both an instance of oa:Target and an instance of schema:CreativeWork. This allows using properties of schema.org for providing more information about the article, such as its URL (instance of schema:URL), its publication date (instance of schema:DateTime), its headline (instance of schema:Text), its author (instance of schema:Person), and its content (instance of schema:Text).

[5] https://www.w3.org/TR/annotation-model/
[6] https://schema.org/

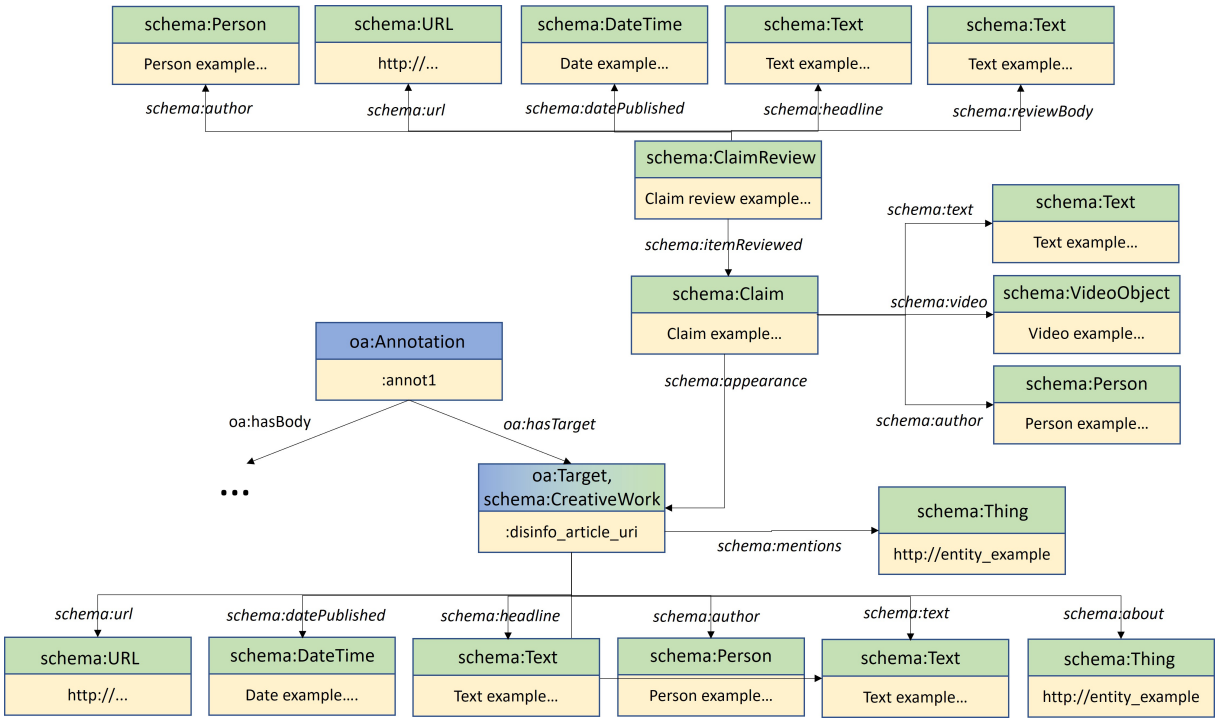

Fig. 4. Enriching the annotated resource with rich information using schema.org.

We can also link the article with entities of different types mentioned in it, such as persons, places, etc., using the property schema:mentions. This enables connecting the annotation to widely used knowledge graphs, such as DBpedia and Wikidata, which provide descriptions of millions of entities across different domains. In addition, we can link claims (instances of schema:Claim) to the articles using the property schema:appearance. A claim can be then linked to its text, video/audio (if any) and author (using the properties schema:text, schema:video/schema:audio, and schema:author, respectively), as well as with claim reviews (instances of schema:ClaimReview). In a similar way, a claim review can be linked with related data such as its author, URL, publication date, headline, review body, etc.

Another well-known vocabulary that can be used together with the taxonomy is the SIOC Core Ontology,[7] a data model that provides the main concepts and properties required to describe information from social media and online communities. Linking to such established vocabularies supports the integration of annotation data with existing knowledge bases that make use of the same or similar data models, such as ClaimsKG [74], CimpleKG [17], TweetsKB [22], and TweetsCov19 [20].

Queries that can be answered using TAXODIS annotations include:

- retrieve all resources classified as of social theme and which have a high level of emotional polarization
- retrieve all resources with imposter content together with the values of all features that are under ‘psychology feature’
- retrieve the number of resources per content type having high usage of emphatic wording
- retrieve all resources published on a specific time period containing claims that have been reviewed and have received a veracity score ‘mostly false’
- retrieve all resources mentioning a specific person which are mostly false, together with the values of all features that are under ‘detection feature’

The first query of the above list is translated to SPARQL as follows:

[7] https://www.w3.org/submissions/sioc-spec/

```
PREFIX taxodis: <https://hop.fzi.de/taxodis/>
PREFIX oa: <http://www.w3.org/ns/oa#>
PREFIX schema: <http://schema.org/>
SELECT ?resourceUri ?resourceHeadline ?resourceAuthor WHERE {
  ?annot1 oa:hasTarget ?resourceUri ; oa:hasBody taxodis:themsoc .
  ?annot2 oa:hasTarget ?resourceUri ; oa:hasBody ?annot2Body .
  ?annot2Body oa:type taxodis:mobpolar ; rdf:value "high" .
  OPTIONAL { ?resourceUri schema:headline ?resourceHeadline }
  OPTIONAL { ?resourceUri schema:author ?resourceAuthor } }
```

Obtaining the Feature Values. For a given piece of content, we can estimate the value of each feature either manually or using dedicated software. Each approach has its pros and cons. The manual approach provides annotations of very high quality. However, it is very laborious, time-consuming, and not scalable (the annotation time is proportional to the number of texts/documents we want to annotate and the number of considered features). On the contrary, using a software system, we can obtain annotations for large corpora, with no human effort. However, the accuracy of the annotations is questionable and depends on several factors, such as the overall quality and performance of the software system, the availability of training data, the language used in the input texts, etc. Furthermore, there might be a monetary cost for using the system.

Existing software systems that can be used to obtain values for one or more of the TAXODIS features include: i) linguistic and word usage analysis tools (such as LIWC [15]) for the 'detection' features, ii) topic and theme extraction tools [19] for the 'categorization' features, and iii) fact-checking, disinformation detection and claim linking tools (such as ClaimLinker [48]) for the 'veracity' features. Moreover, if enough training (annotation) data is available, dedicated classifiers per feature can be built and used for larger text corpora. Surveying such software systems and evaluating their performance is out of the scope of this paper.

6. TAXODIS Evaluation, Use Cases, and Maintenance

6.1. Taxonomy Use and Evaluation

The taxonomy was initially introduced through an internal workshop in 2023, during which a group of interested researchers (from sociology, computer science, and political science) and practitioners (from NGOs and industry) utilized it to create labels for identifying different types of disinformation for the research project DeFaktS.[8] This evaluation, as described in prior work [9], involved participants from diverse backgrounds, who used the taxonomy to annotate real-world disinformation-related social media content. During the workshop, the participants applied TAXODIS to scrutinize real-world data, i.e. numerous social media posts derived from various platforms (e.g., Telegram and Twitter/X) containing disinformation. These labels were then used in annotating a comprehensive dataset for training a classifier to detect deceptive messages [9]. The workshop, focusing on textual detection of disinformation, involved 15 researchers and practitioners from relevant fields. They used TAXODIS to assess whether a given content was disinformation or not and utilized it as a baseline to create suitable labels for data annotation. Two groups of workshop participants approached the task in different ways, testing the taxonomy's usefulness during group work. Jointly they generated a list of 15 polar labels, 13 of which were selected either directly from the taxonomy or created with its assistance, such as by merging two features into one label for the annotation process. To enhance the robustness and reliability of our annotations conducted through the annotation platform Doccano [58], we implemented a cross-annotation process. Specifically, a subset of 767 data samples underwent independent annotation by two teams, each consisting of two annotators. This approach ensured a comprehensive evaluation of the labeling process facilitated by the taxonomy. Subsequently, we computed the inter-annotator agreement [52] to assess the level of concordance between the annotators. To quantify this agreement, we utilized Cohen's Kappa

[8] https://defakts.de/2023/02/24/start-der-annotation-mit-dem-defakts-workshop/

metric, revealing a substantial score of 0.72. This result confirms the strength and dependability of the annotations throughout the dataset, establishing a robust foundation for training a model based on TAXODIS. While the user study and annotation procedure have been published previously [9], we have briefly summarized them here to contextualize the taxonomy's practical use and applicability in real-world annotation workflows.

6.2. Use-Case Scenarios

We do not provide a specific annotation protocol or guidelines for using this taxonomy, because such protocols are usually highly task- and domain-dependent. For example, a text may be annotated as half-true by one group but as false or misleading by another. This can be specifically relevant when dealing to science-related misinformation as opposed to generic claims [31]. In the following, we present prominent interdisciplinary use-case scenarios where TAXODIS can be applied, leaving the design of annotation guidelines to the specific community and teams according to their needs.

Computer Science and AI. In the field of computer science, and in particular AI and supervised learning, the resource can be of use to build and/or fine-tune language models to perform various downstream tasks related to disinformation detection and analysis. The taxonomy enables fine-grained annotation of text with relevant linguistic features, while the use of standards and semantic web technology allows to query and access specific sub-sets of annotated data in a centralized manner, even if they come from different sources. In that way, this technology provides the possibility of extracting data for precisely training or fine-tuning of machine learning models that correspond to specific criteria, according to a specific selection of TAXODIS labels.

The automatic extraction of the taxonomy features from text via dedicated tools could facilitate annotation. Certain features from the taxonomy can be linked to some of the features from the LIWC vocabulary (discussed above), for which LIWC provides tools for their automatic extraction. However, the majority of the vocabulary terms being specific to the disinformation context, dedicated tools for their extraction need to be created. Taking it a step further, the taxonomy can facilitate the annotation of new text with reduced reliance on human labor by incorporating examples into prompts for generative AI systems.

The features can contribute to contextualize the outcomes and predictions in tasks, such as disinformation detection. Indeed, the resource can be useful in enhancing explainability of language models, such as BERT. A language model fine-tuned on corpora annotated by TAXODIS can be applied to perform various downstream tasks, such as classifying texts as disinformation or not. However, the model as such will struggle to provide an interpretation of its prediction, where understanding why a specific piece of information is classified as flawed or not is crucial for journalists or social scientists (cf. below), as well as ordinary users. A major challenge in AI research is indeed the interpretation of the features used by language models, e.g. by extracting the most predictive tokens [49, 71], or by understanding the implicit semantics carried by the embedding layers [18]. In our case, if the corpora that are used to train/fine-tune the model is annotated by the high-level linguistic features coming from TAXODIS, one could be able to conduct explicability analysis by identifying the taxonomy features that contribute most to a specific class prediction. In addition, the vocabulary can help to match the low-level BERT (or BERT-like model) features to high-level, meaningful, and human-curated linguistic features, hence contributing largely to the explainability challenge of language models. A potential way of conducting that analysis is performing independent classification by using a language model with automatically embedded features and then by using a simple binary classifier (like a decision tree) by using the TAXODIS features only and then applying an explainability system, such as SHAP [45] on both in order to identify groups of features on both sides that contribute most to the specific classification outcome.

Social Science. In the field of social sciences, understanding and analyzing online disinformation is crucial for examining its impact on public opinion, behavior, and societal dynamics [5, 26]. Researchers studying the effects of disinformation on social behavior can utilize the taxonomy to systematically categorize and analyze linguistic features within disinformation content. This structured approach allows for more precise measurement and comparison of how different types of disinformation affect various demographic groups and societal segments. Computational social scientists often rely on annotated datasets to train models and conduct analyses [3, 64]. The taxonomy's comprehensive framework aids in the consistent labeling of disinformation instances, ensuring that datasets are uniformly annotated. This uniformity may enhance the reliability of statistical analyses and the generalizability

and long-term validity of research findings. Furthermore, providing a standardized taxonomy may facilitate collaboration between social scientists and computational experts. Researchers can leverage the resource to align their qualitative insights with quantitative analyses, fostering interdisciplinary studies that combine linguistic features with social theories. Finally, social scientists can use insights derived from the taxonomy to inform policy recommendations. Understanding the specific linguistic markers of disinformation enables the development of targeted interventions and strategies for mitigating the adverse effects of disinformation on public discourse and democratic processes [46, 56].

Journalism. In journalism, the taxonomy may serve as a practical tool for improving the accuracy and effectiveness of disinformation detection and fact-checking. Journalists and fact-checkers can use the taxonomy to streamline their verification processes. By referring to the taxonomy's linguistic features, they can more effectively identify and analyze disinformation in news content, ensuring that false claims are quickly and accurately addressed. Additionally, the taxonomy may support journalists in analyzing patterns of disinformation across different media sources. By categorizing linguistic features, journalists can detect recurring themes and tactics used by disinformation campaigns, leading to more informed reporting and deeper investigative insights [38]. In education, the taxonomy may provide a valuable resource for training journalists and media professionals. Offering a clear, empirically grounded guide to recognizing disinformation, the resource may equip journalists with a tool needed to navigate complex information environments and maintain high standards of journalistic integrity. Finally, journalists may use the taxonomy to create educational content that raises public awareness about disinformation. By demonstrating how specific linguistic features indicate false or misleading information, they can help readers become more discerning consumers of news and reduce the spread of disinformation.

6.3. Monitoring and Maintenance

In the near future, we will be both closely monitoring and actively advertising the use of the taxonomy in the aforementioned use-cases and beyond, targeting both semantic web and natural language processing communities. We will actively engage with specific initiatives and projects related to dis- and mis-information detection, such as the AI4Sci[9] and ClaimsKG[10] projects, as well as CLEF's Check That! shared tasks lab[11]. This and other related initiatives will help assess the usefulness of the resource in the short and mid-term, and guide its future evolution with respect to the feedback of the scientific communities that will use it.

7. Conclusion and Future Work

The widespread phenomenon of disinformation, understood as deliberately deceptive or false information, presents important risks to political stability and social cohesion, particularly during times of crises. Automated disinformation detection systems, leveraging machine learning and natural language processing, are essential in the fight against disinformation as tools assisting journalists and social sciences in their efforts. Given the complex and nuanced nature of disinformation, this study contributes a structured taxonomy, named TAXODIS, to aid automated systems in annotating corpora and recognizing linguistic markers of disinformation with high precision. TAXODIS is presented as a SKOS vocabulary, leveraging the semantic web technology and principles. It is, hence, the first resource of its kind that is openly available, reusable, and interoperable, aiming to play the role of a standard, useful for annotation and classification tasks, fostering both scholarly and practical advancements in automated disinformation detection in fields such as computer science, journalism, and social sciences.

In the near future, we aim to enrich the taxonomy to account for disinformation cues. Several categories already lend themselves to such an extension: "Headlines" can be reflected in thumbnails or meme captions, "mobilization" and "subjectiveness" are often conveyed through shocking or emotionally charged imagery, while "authenticity" can be compromised via fake screenshots, logos, or deepfakes. Considering these multimodal markers is particularly

[9] https://ai4sci-project.org/
[10] https://data.gesis.org/claimskg/
[11] https://checkthat.gitlab.io/clef2025/

important in crisis situations [8], where viral multimodal posts may play a major role in spreading disinformation and triggering negative emotions. Another important direction for future work concerns the integration of TAXODIS into disinformation detection tools or annotation platforms, such as ClaimBuster[12] [34] and MAAM[13] [68], enabling the efficient production of annotation data that can then be used for training dedicated classification models.

References

[1] H.Q. Abonizio, J.I. de Morais, G.M. Tavares and S. Barbon Junior, Language-independent Fake News Detection: English, Portuguese, and Spanish Mutual Features, *Future Internet* **12**(5) (2020), 87, Publisher: Multidisciplinary Digital Publishing Institute.

[2] B. Akinyemi, O. Adewusi and A. Oyebade, An Improved Classification Model for Fake News Detection in Social Media, *international journal of Information Technology and Computer Science (IJITCS)* **12**(1) (2020), 34–43.

[3] M. Alassad, B. Spann and N. Agarwal, Combining Advanced Computational Social Science and Graph Theoretic Techniques to Reveal Adversarial Information Operations, *Information Processing & Management* **58**(1) (2021), 102385.

[4] J.M. Alexander and J.M. Smith, Disinformation: A Taxonomy, *Department of Computer & Information Science* **Technical Reports (CIS)** (2010).

[5] H. Allcott and M. Gentzkow, Social Media and Fake News in the 2016 Election, *Journal of economic perspectives* **31**(2) (2017), 211–236.

[6] H. Alsaidi and W. Etaiwi, Empirical Evaluation of Machine Learning Classification Algorithms for Detecting COVID-19 Fake News, *Int. J. Advance Soft Compu. Appl* **14**(1) (2022).

[7] W. Ansar and S. Goswami, Combating the Menace: A Survey on Characterization and Detection of Fake News from a Data Science Perspective, *International Journal of Information Management Data Insights* **1**(2) (2021), 100052, Publisher: Elsevier.

[8] I. Arcos, P. Rosso and R. Salaverría, Divergent Emotional Patterns in Disinformation on Social Media? An Analysis of Tweets and TikToks about the DANA in Valencia, *arXiv preprint arXiv:2501.18640* (2025).

[9] S. Ashraf, I. Bezzaoui, I. Andone, A. Markowetz, J. Fegert and L. Flek, DeFaktS: A German Dataset for Fine-Grained Disinformation Detection through Social Media Framing, in: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 4580–4591.

[10] I. Augenstein, C. Lioma, D. Wang, L. Chaves Lima, C. Hansen, C. Hansen and J.G. Simonsen, MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng and X. Wan, eds, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 4685–4697. doi:10.18653/v1/D19-1475. https://aclanthology.org/D19-1475/.

[11] L. Azevedo, M. d'Aquin, B. Davis and M. Zarrouk, 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)*, Association for Computational Linguistics, 2021, pp. 41–56.

[12] S. Badaskar, S. Agarwal and S. Arora, Identifying Real or Fake Articles: Towards Better Language Modeling, in: *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*, 2008.

[13] I. Bezzaoui, J. Fegert and C. Weinhardt, Truth or Fake? Developing a Taxonomical Framework for the Textual Detection of Online Disinformation, *International journal on advances in internet technology* **15**(3/4) (2022), 53–63.

[14] P. Biyani, K. Tsioutsiouliklis and J. Blackmer, " 8 Amazing Secrets for Getting More Clicks": Detecting Clickbaits in News Streams Using Article Informality, 2016.

[15] R.L. Boyd, A. Ashokkumar, S. Seraj and J.W. Pennebaker, The Development and Psychometric Properties of LIWC-22, *Austin, TX: University of Texas at Austin* **10** (2022).

[16] L. Bozarth and C. Budak, Toward a Better Performance Evaluation Framework for Fake News Classification, 2020, pp. 60–71. ISBN 2334-0770.

[17] G. Burel, M. Mensio, Y. Peskine, R. Troncy, P. Papotti and H. Alani, CimpleKG: A continuously updated knowledge graph on misinformation, factors and fact-checks, in: *International Semantic Web Conference*, Springer, 2024, pp. 97–114.

[18] E. Chersoni, E. Santus, C.-R. Huang and A. Lenci, Decoding Word Embeddings with Brain-based Semantic Features, *Computational Linguistics* **47**(3) (2021), 663–698.

[19] A. Dhar, H. Mukherjee, N.S. Dash and K. Roy, Text Categorization: Past and Present, *Artificial Intelligence Review* **54**(4) (2021), 3007–3054.

[20] D. Dimitrov, E. Baran, P. Fafalios, R. Yu, X. Zhu, M. Zloch and S. Dietze, TweetsCov19-a knowledge base of semantically annotated tweets about the COVID-19 pandemic, in: *Proceedings of the 29th ACM international conference on information & knowledge management*, 2020, pp. 2991–2998.

[21] DISARM, DISARM Framework Explorer, 2023. doi:https://disarmframework.herokuapp.com.

[12] https://idir.uta.edu/claimbuster/

[13] https://mever.gr/tools/media-asset-annotation-management/

[22] P. Fafalios, V. Iosifidis, E. Ntoutsi and S. Dietze, Tweetskb: A Public and Large-scale rdf Corpus of Annotated Tweets, in: *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, Springer, 2018, pp. 177–190.

[23] M. Fayaz, A. Khan, M. Bilal and S.U. Khan, Machine Learning for Fake News Classification with Optimal Feature Selection, *Soft Computing* (2022), 1–9, Publisher: Springer.

[24] A.C.T. Fernandez, Computing the Linguistic-Based Cues of Credible and Not Credible News in the Philippines Towards Fake News Detection (2019).

[25] V.C. Ferreira, S. Kundu and F.M. França, Analysis of Fake News Classification for Insight into the Roles of Different Data Types, in: *2022 IEEE 16th International Conference on Semantic Computing (ICSC)*, IEEE, 2022, pp. 75–82.

[26] D. Freelon and C. Wells, Disinformation as Political Communication, *Political Communication* (2020), 145–156.

[27] S. Gangopadhyay, K. Boland, D. Dessí, S. Dietze, P. Fafalios, A. Tchechmedjiev, K. Todorov and H. Jabeen, Truth or dare: Investigating Claims Truthfulness with Claimskg, in: *D2R2 2023-2nd International Workshop on Linked Data-driven Resilience Research*, Vol. 3401, 2023.

[28] S. Gangopadhyay, S. Schellhammer, S. Hafid, D. Dessi, C. Koß, K. Todorov, S. Dietze and H. Jabeen, Investigating Characteristics, Biases and Evolution of Fact-Checked Claims on the Web, in: *Proceedings of the 35th ACM Conference on Hypertext and Social Media*, 2024, pp. 246–258.

[29] M. Gasquet, D. Brechtel, M. Zloch, A. Tchechmedjiev, K. Boland, P. Fafalios, S. Dietze and K. Todorov, Exploring Fact-checked Claims and their Descriptive Statistics, in: *Proceedings of the ISWC 2019 Satellite Tracks (Posters & Demonstrations, Industry, and Outrageous Ideas) co-located with 18th International Semantic Web Conference (ISWC 2019), Auckland, New Zealand, October 26-30, 2019*, CEUR-WS.org, 2019.

[30] M. Gruppi, B.D. Horne and S. Adali, An Exploration of Unreliable News Classification in Brazil and the US, *arXiv preprint arXiv:1806.02875* (2018).

[31] S. Hafid, S. Schellhammer, Y.S. Kartal, T. Papastergiou, S. Dietze, S. Bringay and K. Todorov, An In-depth Analysis of the Linguistic Characteristics of Science Claims on the Web and their Impact on Fact-checking, *ACM Transactions on the Web* **19**(3) (2025), 1–31.

[32] S.K. Hamed, M.J. Ab Aziz and M.R. Yaakub, Fake News Detection Model on Social Media by Leveraging Sentiment Analysis of News Content and Emotion Analysis of Users' Comments, *Sensors (Basel, Switzerland)* **23**(4) (2023), 1748. doi:10.3390/s23041748. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9960438/.

[33] H. Hanke and D. Knees, A Phase-Field Damage Model Based on Evolving Microstructure, *Asymptotic Analysis* **101** (2017), 149–180.

[34] N. Hassan, G. Zhang, F. Arslan, J. Caraballo, D. Jimenez, S. Gawsane, S. Hasan, M. Joseph, A. Kulkarni, A.K. Nayak et al., Claimbuster: The first-ever end-to-end fact-checking system, *Proceedings of the VLDB Endowment* **10**(12) (2017), 1945–1948.

[35] B. Horne and S. Adali, 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*, Vol. 11, 2017, pp. 759–766, Issue: 1. ISBN 2334-0770.

[36] C.L.M. Jeronimo, L.B. Marinho, C.E. Campelo, A. Veloso and A.S. da Costa Melo, Fake News Classification Based on Subjective Language, in: *Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services*, 2019, pp. 15–24.

[37] E. Kapantai, A. Christopoulou, C. Berberidis and V. Peristeras, A Systematic Literature Review on Disinformation: Toward a Uified Taxonomical Framework, *New Media & Society* **23**(5) (2021), 1301–1326.

[38] A.Y. Kebede, A.C. Ali and M.A. Moges, Examining Journalists Organizational Trust Pursuant to Predictive Variables in the Ethiopian Media Industry: The Case Study of Amhara Media Corporation, *Cogent Social Sciences* **8**(1) (2022), 2068271. doi:10.1080/23311886.2022.2068271.

[39] S. Kumar and N. Shah, False Information on Web and Social Media: A Survey, *arXiv preprint arXiv:1804.08559* (2018).

[40] S. Kumar, R. West and J. Leskovec, Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes, in: *Proceedings of the 25th international conference on World Wide Web*, 2016, pp. 591–602.

[41] D. Kundisch, J. Muntermann, A.M. Oberländer, D. Rau, M. Röglinger, T. Schoormann and D. Szopinski, An Update for Taxonomy Designers, *Business & Information Systems Engineering* **64**(4) (2022), 421–439, Publisher: Springer.

[42] Y. Lasotte, E. Garba, Y. Malgwi and M. Buhari, An Ensemble Machine Learning Approach for Fake News Detection and Classification Using a Soft Voting Classifier, *European Journal of Electrical Engineering and Computer Science* **6**(2) (2022), 1–7.

[43] E. Lefever, A Hybrid Approach to Domain-Independent Taxonomy Learning, *Applied Ontology* **11**(3) (2016), 255–278.

[44] V. Lemieux and T.D. Smith, Leveraging Archival Theory to Develop a Taxonomy of Online Disinformation, IEEE, 2018, pp. 4420–4426. ISBN 1-5386-5035-5.

[45] S.M. Lundberg and S.-I. Lee, A Unified Approach to Interpreting Model Predictions, *Advances in neural information processing systems* **30** (2017).

[46] B. Lutz, M. Adam, S. Feuerriegel, N. Pröllochs and D. Neumann, Which Linguistic Cues Make People Fall for Fake News? A Comparison of Cognitive and Affective Processing, *Proceedings of the ACM on Human-Computer Interaction* **8**(CSCW1) (2024), 1–22. doi:10.1145/3641030.

[47] M. Mahyoob, J. Algaraady and M. Alrahaili, Linguistic-Based Detection of Fake News in Social Media, *International Journal of English LInguistics* **11**(1) (2021).

[48] E. Maliaroudakis, K. Boland, S. Dietze, K. Todorov, Y. Tzitzikas and P. Fafalios, ClaimLinker: Linking Text to a Knowledge Graph of Fact-checked Claims, in: *Companion Proceedings of the Web Conference 2021*, 2021, pp. 669–672.

[49] I. Malkiel, D. Ginzburg, O. Barkan, A. Caciularu, J. Weill and N. Koenigstein, Interpreting BERT-based Text Similarity via Activation and Saliency Maps, in: *Proceedings of the ACM Web Conference 2022*, 2022, pp. 3259–3268.

[50] D.M. Markowitz and J.T. Hancock, Linguistic Traces of a Scientific Fraud: The Case of Diederik Stapel, *PloS one* **9**(8) (2014), e105937, Publisher: Public Library of Science San Francisco, USA.

[51] J.R. Martinez-Rico, J. Martinez-Romo and L. Araujo, NLP &IRUNED at CheckThat! 2022: Ensemble of Classifiers for Fake News Detection, *Working Notes of CLEF* (2022).

[52] M.L. McHugh, Interrater Reliability: The Kappa Statistic, *Biochemia medica* **22**(3) (2012), 276–282.

[53] P. Meel and D.K. Vishwakarma, Fake News, Rumor, Information Pollution in Social Media and Aeb: A Contemporary Survey of State-of-the-Arts, Challenges and Opportunities, *Expert Systems with Applications* **153** (2020), 112986, Publisher: Elsevier.

[54] P.S. Meltzer, A. Kallioniemi and J.M. Trent, Chromosome alterations in human solid tumors, in: *The Genetic Basis of Human Cancer*, B. Vogelstein and K.W. Kinzler, eds, McGraw-Hill, New York, 2002, pp. 93–113.

[55] M.D. Molina, S.S. Sundar, T. Le and D. Lee, "Fake News" is not Simply False Information: A Concept Explication and Taxonomy of Online Content, *American behavioral scientist* **65**(2) (2021), 180–212, Publisher: SAGE Publications Sage CA: Los Angeles, CA.

[56] L. Munn, Angry by Design: Toxic Communication and Technical Architectures, *Humanities and Social Sciences Communications* **7**(1) (2020), 1–11, Publisher: Palgrave.

[57] P.R. Murray, K.S. Rosenthal, G.S. Kobayashi and M.A. Pfaller, *Medical Microbiology*, 4th edn, Mosby, St. Louis, 2002.

[58] H. Nakayama, T. Kubo, J. Kamura, Y. Taniguchi and X. Liang, doccano: Text Annotation Tool for Human, 2018, Software available from https://github.com/doccano/doccano. https://github.com/doccano/doccano.

[59] U. Nations, Countering Disinformation for the Promotion and Protection of Human Rights and Fundamental Freedoms., Report of the Secretary-General., 2022. N2245924.pdf(un.org).

[60] R.C. Nickerson, U. Varshney and J. Muntermann, A Method for Taxonomy Development and its Application in Information Systems, *European Journal of Information Systems* **22** (2013), 336–359.

[61] S.B. Parikh and P.K. Atrey, Media-rich Fake News Detection: A Survey, in: *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*, IEEE, 2018, pp. 436–441.

[62] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff and B. Stein, A Stylometric Inquiry into Hyperpartisan and Fake News, in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, I. Gurevych and Y. Miyao, eds, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 231–240. doi:10.18653/v1/P18-1022. https://aclanthology.org/P18-1022/.

[63] H. Rashkin, E. Choi, J.Y. Jang, S. Volkova and Y. Choi, 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*, 2017, pp. 2931–2937.

[64] C. Rauh and J. Schwalbach, The ParlSpeech V2 Data Set: Full-text Corpora of 6.3 Million Parliamentary Speeches in the Key Legislative Chambers of Nine Representative Democracies, *Harvard Dataverse* **1** (2020), 1.

[65] J.F. Ribeiro Bezerra, Content-based Fake News Classification through Modified Voting Ensemble, *Journal of Information and Telecommunication* **5**(4) (2021), 499–513, Publisher: Taylor & Francis.

[66] D. Rohera, H. Shethna, K. Patel, U. Thakker, S. Tanwar, R. Gupta, W.-C. Hong and R. Sharma, A Taxonomy of Fake News Classification Techniques: Survey and Implementation Aspects, *IEEE Access* (2022), Publisher: IEEE.

[67] K.A. Rosińska, Disinformation in Poland: Thematic Classification Based on Content Analysis of Fake News from 2019, *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* **15**(4) (2021).

[68] M. Schinas, P. Galopoulos and S. Papadopoulos, MAAM: Media Asset Annotation and Management, in: *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, 2023, pp. 659–663.

[69] K. Shu, A. Sliva, S. Wang, J. Tang and H. Liu, Fake News Detection on Social Media: A Data Mining Perspective, *ACM SIGKDD explorations newsletter* **19**(1) (2017), 22–36.

[70] K. Shu, A. Bhattacharjee, F. Alatawi, T.H. Nazer, K. Ding, M. Karami and H. Liu, Combating Disinformation in a Social Media Age, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **10**(6) (2020), e1385.

[71] M. Szczepański, M. Pawlicki, R. Kozik and M. Choraś, New Explainability Method for BERT-based Model in Fake News Detection, *Scientific reports* **11**(1) (2021), 23705.

[72] D. Tambini, Fake News: Public Policy Responses (2017), Publisher: The London School of Economics and Political Science.

[73] E.C. Tandoc Jr, Z.W. Lim and R. Ling, Defining "Fake News" A Typology of Scholarly Definitions, *Digital journalism* **6**(2) (2018), 137–153.

[74] A. Tchechmedjiev, P. Fafalios, K. Boland, M. Gasquet, M. Zloch, B. Zapilko, S. Dietze and K. Todorov, ClaimsKG: A Knowledge Graph of Fact-Checked Claims, in: *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings*, Springer, Berlin, Heidelberg, 2019, pp. 309–324.

[75] P.K. Verma, P. Agrawal, I. Amorim and R. Prodan, WELFake: Word Embedding over Linguistic Features for Fake News Detection, *IEEE Transactions on Computational Social Systems* **8**(4) (2021), 881–893, Publisher: IEEE.

[76] L. Wang, Y. Wang, G. de Melo and G. Weikum, Understanding Archetypes of Fake News via Fine-grained Classification, *Social Network Analysis and Mining* **9**(1) (2019), 1–17, Publisher: Springer.

[77] C. Wardle, A New World Disorder, *Scientific American* **321**(3) (2019), 88–95.

[78] C. Wardle and H. Derakhshan, *Information Disorder: Toward an Interdisciplinary Framework for Research and Policymaking*, Vol. 27, Council of Europe Strasbourg, 2017.

[79] J. Webster and R.T. Watson, Analyzing the Past to Prepare for the Future: Wiriting a Literature Review, *MIS Quarterly* **26**(2) (2002), 13–23.

[80] E. Wilson, Active vibration analysis of thin-walled beams, PhD thesis, University of Virginia, 1991.

[81] S. Yu and D. Lo, Disinformation Detection using Passive Aggressive Algorithms, *ACM Southeast Conference, Session 4* (2020), 324f..

[82] L. Zhou, J.K. Burgoon, J.F. Nunamaker and D. Twitchell, Automating Linguistics-based Cues for Detecting Deception in Text-based Asynchronous Computer-mediated Communications, *Group decision and negotiation* **13**(1) (2004), 81–106, Publisher: Springer.
[83] X. Zhou, A. Jain, V.V. Phoha and R. Zafarani, Fake News Early Detection: An Interdisciplinary Study, *arXiv preprint arXiv:1904.11679* (2019).