Truth or Fake? Developing a Taxonomical Framework for the Textual Detection of Online Disinformation

Isabel Bezzaoui, Jonas Fegert and Christof Weinhardt
Information Process Engineering
FZI Research Center for Information Technology
Karlsruhe/Berlin, Germany
bezzaoui@fzi.de, fegert@fzi.de, weinhardt@fzi.de

***Abstract* — Disinformation campaigns have become a major threat to democracy and social cohesion. Phenomena like conspiracy theories promote political polarization; they can influence elections and lead people to (self-)damaging or even terrorist behavior. Since social media users and even larger platform operators are currently unready to precisely detect disinformation, new techniques for identifying online disinformation are urgently needed. In this paper, we present the first research insights of DeFaktS, an Information Systems research project, which takes a comprehensive approach to both researching and combating online disinformation with a special focus on enhancing media literacy and trust in explainable AI. Specifically, we demonstrate the first methodological steps towards the training of a machine learning-based system. This will be obtained by introducing the development and preliminary results of a taxonomy to support the labeling of a 'Fake News' dataset.**

Keywords - Fake News; Disinformation Detection; Machine Learning-Based Systems; Taxonomy.

I. INTRODUCTION

Online disinformation is currently regarded as one of the most serious challenges to democracy, journalism, and free expression, increasing the demand for research on the detection of fraudulent content. The present paper seeks to extend the findings of [1], a research project focusing on using explainable AI to understand and combat online disinformation. As the major news source of today, social media channels and online news portals suffer from non-fact-based reporting and opinion dissemination [2]. Spreading virally, disinformation poses a central threat to the political process and social cohesion. Disinformation is defined as false information, spread with the intention to deceive. 'Fake News' is an example of disinformation, which is why, in line with current literature in ICT research, we use these two terms interchangeably [3]. Deceptive information influences elections and tempts people to engage in (self-)damaging or even terrorist behavior. Accordingly, it displays a generally undesirable phenomenon in public information and opinion-forming processes [4,5]. Besides political radicalization [6], vaccination boycotts are increasingly attributed to disinformation campaigns and thereby present a threat to the general health system [7,8]. Therefore, on the one hand, there is a need for a comprehensive understanding of their mechanisms and spread, and on the other hand, based on this, methods to systematically combat them. People are naturally inclined to consume content with which they are familiar (familiarity bias), whose authors are similar to them (similarity bias), or whose statements they basically agree with (confirmation bias). In particular, confirmation bias is a decisive factor in the spread of disinformation [9]. Platforms such as WhatsApp and Telegram play a major role in the spread of disinformation and could take many preventive measures. They generally lack the appropriate approaches for this, since more emotional arousal and dissent lead to more activities on the platform, and in turn generate more revenue in advertising [10, 11]. Even though Twitter, for example, is experimenting with fact checking, these efforts are far from sufficient to limit the spread of 'Fake News' as those services do not operate across platforms. Thus, DeFaktS tries to empower actual users across various platforms to critically question news and social media content. For this purpose, the project will develop an explainable AI artifact for a participation platform that aims to combat online disinformation campaigns and foster critical media literacy among users by informing them about the occurrence of 'Fake News' in a transparent and trustworthy manner. Precisely, the DeFaktS project develops a data pipeline in which (i) messages are extracted by annotators in large quantities from suspicious social media and messenger groups. Based on this corpus, a machine learning-based system (ii) is trained that can recognize factors and stylistic devices characteristic of disinformation, which will be used for (iii) an explainable AI that informs users in a simple and comprehensible way about the occurrence of disinformation.

Machine data analytics remain challenged by the wide variety of stylistic devices utilized in fraudulent messages, which poses a barrier for merely quantitative approaches to the issue [12]. Empirical findings demonstrate that disinformation content is the hardest to detect. This seems

reasonable considering that the false class is dispersed and layered over other classes. The deceptive nature of 'Fake News', where the goal is to make the information appear to be a legitimate piece, may help to explain this [13]. Nevertheless, studies on the structure of disinformation indicate that the substance of authentic and deceptive news articles differs significantly [14]. The DeFaktS project seeks to face this challenge in the following way: The development of a taxonomy of online disinformation (TOD) that entails linguistic features and dimensions of disinformation content shall facilitate and ensure the quality of the data labeling process. One of the difficulties in detecting false news is that some terms and expressions are unique to a particular type of event or topic. When a 'Fake News' classifier is trained on fake versus real articles based on a certain event or topic, the classifier learns event-specific features and may not perform well when used to identify deceptive content based on a different type of event. As a result, 'Fake News' classifiers must be generalized to be event-independent [3]. Another challenge is that the majority of datasets are in English, and German-language datasets are rare [15]. Since the spread of disinformation is not bound to language barriers, creating functioning datasets in other languages is crucial. Recent research addresses the opportunities of different detection methods and their underlying theories [16-19]. What is lacking, however, is a fundamental empirical overview of concrete detection cues supporting the creation of labels for annotating datasets. Furthermore, even though there are numerous empirical papers presenting disinformation classifiers, they offer no explanations on how these classifiers were trained and how exactly the datasets used for training were labeled [20-23]. Although these explanations are critical to the transparency and traceability of the overall research process and prove that current scientific knowledge is considered in the labeling process of the data, little research has addressed this issue. These observations call for the creation of a taxonomy of online disinformation that encompasses broad and event-independent dimensions and characteristics of disinformation, which is still specific enough to precisely identify and label deceiving content. The systematic coordination of knowledge is an ongoing issue in information systems research [24]. The classification of items helps understanding and analyzing complex settings, and therefore, the creation of taxonomies are crucial for research and development [25]. Furthermore, by using taxonomies, a domain's (e.g., disinformation) knowledge body can be organized and given structure [26]. According to the design science epistemology [27], which also covers descriptive knowledge and prescriptive knowledge, taxonomies are examples of conceptual knowledge. Our final taxonomy will display a design artifact in and for itself that will be demonstrated and evaluated before as well as within the labeling process. After the artifact undergoes iterations, it will be made accessible, and thus extendable, to other researchers through scientific publications or open access services. In this paper, we extend our findings [1] by providing insights into the methodological approach of developing a taxonomical framework for the textual detection of online disinformation as well as an overview of our preliminary results. The paper is structured as follows: Section II will give an introductory overview on the current knowledge base on the efforts of combating disinformation as well as the concepts of critical media literacy and trust. Subsequently, the scientific method and first research activities in the project will be presented in Section III. Thereafter, we provide a short overview of our project's preliminary results in Section IV. Finally, the paper concludes with a summary of the project's research endeavors and an outlook on future work related to the project in Section V.

II. THEORETICAL FOUNDATION

A. Combating Disinformation Using Machine Learning-Based Systems

The fact that nowadays almost anyone can publish content on the internet not only increases the possibility of social participation – it also creates new opportunities for spreading disinformation and propaganda. The COVID-19 pandemic has already produced a flood of false reports and demonstrated the importance of being able to distinguish reliable information from mis- and disinformation. The war on Ukraine also demands a special confrontation with disinformation distributed by state entities [28]. Currently, research on 'Fake News' detection using machine learning-based systems (MLS) is a rapidly expanding field that spans numerous disciplines, including computer science, social science, psychology, and information systems [29-31]. Synoptically, empirical efforts to detect and combat disinformation can be divided into four categories: data-oriented, feature-oriented, model-oriented and application-oriented [2]. The majority of methods concentrate on extracting multiple features, putting them into classification models, such as naive Bayes, logistic regression, or decision trees, and then selecting the best classifier based on performance [32-35]. What is missing from the previous work, however, are empirical evaluations of when the classifiers are put into practice with real users and of what benefits and impact the presented tools may have. For instance, Guess et al. [36] showed that promoting media literacy can help people judge the authenticity of online content more accurately. Their findings suggest that a lack of critical media literacy is a major factor in why people fall for disinformation. Pennycook and Rand [37] found that susceptibility to 'Fake News' is driven mostly by insufficient critical thinking rather than by partisan bias per se. Thus, in order to counter false news, more critical media competence is needed on the part of users. From this point of view, it seems crucial to investigate the potential of MLS

detection tools for promoting critical media literacy among social media users.

Furthermore, previous research has demonstrated the importance of trust for the acceptance and perceived usefulness of ICT tools, and MLS in particular [38,39]. Trust is one of the vital components to fostering active, engaged and informed citizens [40]. Transparency is therefore an important aspect when it comes to dealing with disinformation. In this regard, the challenge of how to positively affect and build trust when developing tools for 'Fake News' detection arises. The implementation of an XAI-approach into the development process seeks to make the system's internal dynamics more transparent, as well as the analysis' conclusions more understandable and hence trustworthy to the user. These observations give rise to the need to examine the effect of XAI (Explainable Artificial Intelligence) elements on user trust and thus acceptance and perceived usefulness of the final tool. In order to fill the two above-mentioned research gaps, we would therefore like to address the following research questions in the DeFaktS project:

How to design an artifact for the detection of online disinformation that helps to foster informed and critical thinking?

i. How does the tool promote critical media literacy by helping users identify disinformation more accurately?

ii. How does the tool's XAI-component assist users to trust the algorithm's assessment?

B. Critical Media Literacy

Disinformation is producing uncertainty in the process of information procurement, endangering the public's ability to make informed decisions [41]. In order to foster a critical comprehension of both manipulative communications and the internet as a distribution medium, users must have broad knowledge and a deeper understanding of social media functionalities [42]. Critical media literacy encourages people to consider why a message was sent and where it came from [43]. Following [44], critical media literacy entails developing skills in analyzing media codes and conventions, and the ability to critique stereotypes, dominant values, and ideologies, as well as the competence to interpret media texts' multiple meanings and messages. Furthermore, it assists individuals to use media responsibly, to discern and assess media content, to critically examine media forms, to explore media effects, and based on those abilities to deconstruct alternative media. However, a systematic evaluation of the effects of the usage of MLS 'Fake News' detection tools on the cultivation of critical media literacy is scarce [45]. Schmitt et al. [45] define three dimensions of critical media literacy that can be referred to the critical handling of online disinformation:

1. Awareness: Awareness in this case means to become aware of the existence of disinformation. This includes knowledge of various forms of disinformation (picture, text, or video form, distorted articles, and pseudo media outlets) as well as a deeper understanding of how media, and online media in particular, operate.
2. Reflection: Reflection in the context of critical media literacy is about applying analytical criteria to internet content and determining whether or not it is deceptive. The conscious consideration, reflection, of content with the character of news is relevant, the thorough thinking before an article is liked, shared or the claim of a headline is taken at face value. As a result, reflection utilizes an individual's knowledge, abilities, and attitudes to critically evaluate (media-communicated) information based on specific criteria including credibility, source, and quality.
3. Empowerment: Individuals' confidence in their ability to detect manipulative messages, participate in social discourses, and actively position themselves against disinformation is cultivated through empowerment strategies and methods. In this context, empowerment can be defined as a certain form of behavior that encompasses a person's ability to recognize and express doubts about specific content as well as express their own thoughts.

In the DeFaktS project, these three dimensions will be used to investigate whether and to what extent the developed MLS can make a positive contribution to the cultivation of critical media competence among social media users. To this end, it will be analyzed whether and to what extent awareness, reflection, and empowerment are strengthened by the use of the artifact.

C. Trust

Niklas Luhmann [46] understands trust in the broadest sense as an elementary component of social life, interpreting it as a form of security, which can only be gained and maintained in the present. First and foremost, trust is needed to reduce a future of more or less undetermined complexity. According to Luhmann's understanding, the constant technical progress of society brings with it a simultaneous increase in complexity, which subsequently results in an increased need for trust. Thus, trust is a necessary condition to live and act with growing complexity in relation to modern events and dynamics [46]. However, trust is severely shaken by negative experiences [47], for instance caused by deception through disinformation. As MLS

systems and algorithms become more complex, people increasingly regard them as 'black boxes' that defy comprehension in the sense that understanding an MLS's decision requires growing amounts of specialized expertise and knowledge. Non-expert end-users are not able to retrace how the algorithmic code cascades led to a given decision [48]. Accordingly, there has been increased demand to offer the proper explanation for how and why a particular result was obtained [49]. Recent empirical evidence on algorithm acceptance [50] insinuates that explainability plays a heuristic role in algorithm and MLS service acceptance. Currently, however, research gives light to a controversy over whether the implementation of XAI-features actually helps increase user-trust or not. Shin [51] analyzed the impact of explainability in MLS on user trust and attitudes towards MLS and concluded that the inclusion of causability and explanatory features in MLS assists to increase trust as it helps users understand the decision-making process of MLS algorithms by providing transparency and accountability. In contrast, through their experiment on transparency and trust in MLS, [52] found that transparency features can actually affect trust negatively. These recent contradictory observations give rise to the need for further investigation of the effect of explainability on user trust. In the DeFaktS project, this research gap will be addressed through the evaluation of whether, and if so which, XAI elements increase user trust in the application.

III. METHODOLOGY

The goal of DeFaktS is to develop an artifact that is as close as possible to the needs of the subsequent user so that it contributes precisely to solving the above-mentioned issues. To implement this, the project is embedded in a design science research approach according to Peffers et al. [53], dividing the process into six steps: problem identification and motivation, definition of the objectives of a solution, design and development, demonstration, evaluation, and communication. Our research methodology for developing a taxonomy is based on the design science research paradigm, which seeks to address new knowledge about artificial objects that are designed to meet specific goals and benefit their users [54]. After identifying the problem and our motivation to develop a taxonomy of online disinformation, we defined the objectives of a solution; building an artifact that is intended to support researchers during the process of identifying and classifying (online) disinformation. Furthermore, the artifact serves as a basis for future design science research projects, the purpose of which is to investigate online disinformation and extend the given taxonomy [25]. This section gives insights into the design and development phase that researchers in the DeFaktS project are currently concerned with. Based on a structured literature review, we first build an artifact (TOD) for identifying and labeling disinformation. Then we evaluate the artifact by using it to create labels for a real-world dataset of factual news and 'Fake News'. To this end, a group of experts evaluates the taxonomy by assessing its efficacy in developing labels for classifying social media content of interest in our specific domain. After the steps of demonstration and evaluation are completed, the artifact will be communicated via scholarly publications.

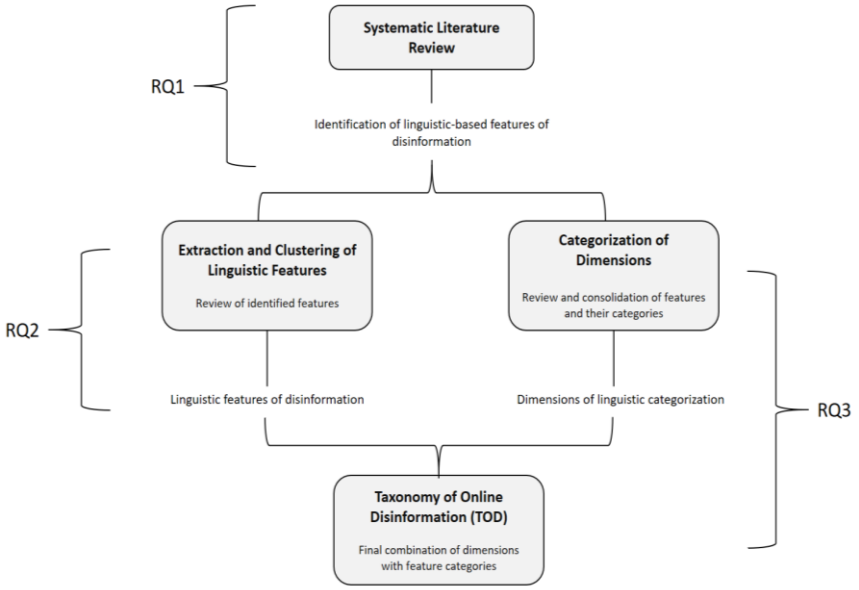

Figure 1. Overall research outline.

Our approach, visualized in Figure 1, consists of two major parts that will be presented in the following. Initially, by conducting a systematic literature review [55], we gather all types of linguistic features of online disinformation in the literature. Subsequently, we cluster the empirical results in groups, supporting a linguistic-based 'Fake News' detection approach. Finally, we propose a novel, five-dimensional taxonomical framework, based on the categorization criteria found in the existing empirical literature. Our proceeding is guided by the following research questions:

1. What linguistic-based cues of (online) disinformation can be found in the empirical literature?
2. How can the linguistic features be clustered in an overarching schema?
3. How can the dimensions and categories resulting from the schema be conjugated into a taxonomy?

A. Systematic Literature Review

To comprehensively address our first research question, we conducted a systematic literature review based on Webster and Watson's [55] methodological guidelines. A thorough review contains pertinent literature on the subject and is not restricted to a particular research approach, collection of journals, or geographical area [55]. For this reason, we make use of large interdisciplinary databases to access all research fields relevant to our project. After carefully examining the literature on linguistic features and disinformation detection characteristics, we end up with an overview of descriptions that are frequently used to refer to different kinds and characteristics of disinformation content. However, the ad hoc definitions that each study introduces can cause conflicts or overlaps. Accordingly, the overall goal of our literature review is to make sense of the accumulated knowledge on categorizing disinformation, as well as to find patterns and identify key concepts in the literature in order to extend past research by synthesizing said knowledge into a useful taxonomy. For our review, we applied the following procedure:

1. Selection of our sources (digital libraries)
2. Definition of search terms
3. Application of each search term on selected sources
4. Selection of primary studies by use of inclusion and exclusion criteria on search results
5. Backward and forward search based on the selected primary studies

An automatic search was based on the following five primary sources of scientific databases to identify relevant publications: IEEE Xplore Digital Library, Scopus, Web of Science, Springer Link and Google Scholar.

We conducted several pilot searches based on our research topics to compile a preliminary list of papers. The search terms that best suited our research objectives were then defined using those as the foundation for the systematic review. The utilized search phrases restricted to abstract and title are listed in the following:

a. "fake news classification"
b. "disinformation classification"
c. "linguistic fake news detection"
d. "linguistic disinformation detection"
e. "linguistic fake news classification"
f. "linguistic disinformation classification"

For the next phase of our research, the following three inclusion and exclusion criteria were formulated:

1. We excluded sources that approached the issue of disinformation solely from a computational standpoint, proposing technical solutions based on, for instance, machine learning and statistical models to categorize news articles into predefined categories automatically, such as fake or real, as well as mere performance evaluations of such models.
2. Publications that mention specific categories or characteristics of false information without making an effort to classify them systematically or even to explain the proposed categories were excluded. This is used to describe sources where the disinformation phenomenon is either not a central concept (such as papers that happen to use terms like 'Fake News'), or they mention specific types of false information outside of a general framework or classification model and are therefore non-exhaustive or indicative.

   In the interest of common scientific understanding, only papers written in English were included.

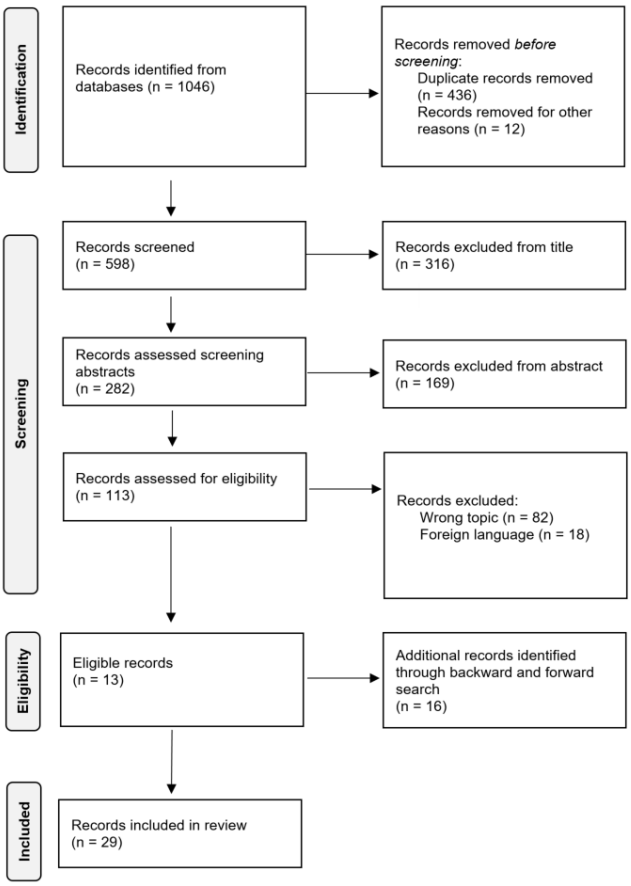

Figure 2. Prisma flow diagram.

Our search results, including the citations from the mentioned libraries, identify thirteen primary studies from six different disciplines (e.g., computer science, linguistics, psychology and media studies) where linguistic frameworks for disinformation detection are introduced. Figure 2 presents in detail the selection process of both records found through database searching and records identified through an additional backward and forward search based on the initial records, resulting in 29 papers included in our review. Our first goal was to identify linguistic-based cues of online disinformation in the empirical literature (RQ1). For addressing RQ2, we then extracted the identified features of disinformation and clustered them by similarities in an overarching schema in order to prepare our findings for RQ3.

B. Taxonomy of Online Disinformation

The overall goal of our research is to create a taxonomy of online disinformation, called TOD, that helps create a common understanding of what constitutes 'Fake News' or disinformation, provides a list of categories and detection characteristics and can be used to develop labels that can be applied to German using diverse 'real world' datasets (RQ3). After examining the findings from RQ1 and RQ2, we identify and extract the dimensions and categorization criteria resulting from our examination to select relevant and recurring ones. Our conclusive step is to systematically organize and map them into a taxonomy. In the context of our project, the final taxonomy shall support researchers to precisely label the datasets that will be used to train the DeFaktS AI. Beyond the scope of the DeFaktS project, offering a comprehensive and fine-grained taxonomy could also be utilized for educational purposes. Since online disinformation may influence people's actions [56], considering the issue of classifying online disinformation from a global viewpoint could help to prevent having significant repercussions in real-life settings. Additionally, our research may also be useful in fields like artificial intelligence, where it is crucial to encode real-world concepts and entities consistently and methodically. A fact-checking or disinformation detection system will be able to produce the most accurate and understandable results the more clearly defined a particular type of disinformation is. The 'Liar, Liar Pants on Fire' dataset [57] and the 'Fake News Corpus' [58] are just two examples of the numerous 'Fake News' datasets that are currently available and used to research and develop detection models with utterly different labeling schemes. So far, the performance of computational models using various conceptual frameworks is not directly comparable, which makes it difficult to define the state of the art in science and industry and ultimately hinders the advancement of research. However, it is crucial for academics and professionals from various fields to come to consensus on this complicated subject, not only about the macro-level notions but also, if feasible, regarding the lower level of more specific attributes and subcategories.

As Figure 3 shows, the process starts with defining the meta-characteristic (1) based on the purpose of the taxonomy, that is linguistic cues of disinformation. Subsequently, the ending conditions are determined (2). In our case, we chose an empirical-to-conceptual approach, gathering empirical results through our systematic literature review from which we identify and extract common characteristics (3, 4). These characteristics are then grouped into dimensions to create and potentially revise the taxonomy (5). Practically, once a set of traits has been determined through the review, they can be formally categorized using statistical methods or arbitrarily by means of a manual or graphical process. The resulting groups define the taxonomy's initial dimensions. Since our method is iterative, conditions must be met in order to know when to stop (6).

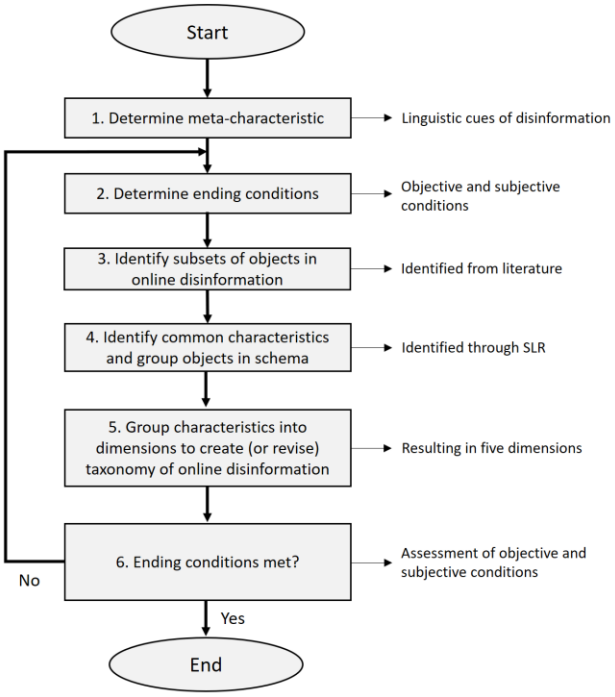

Figure 3. Taxonomy development process [25].

There are both objective and subjective conditions [25]. Throughout the procedure, we check to see if the ending conditions have been satisfied with the taxonomy's current iteration at the conclusion of either of these steps (Tables 1 and 2). Conditions must be examined on both the objective and subjective level. If the objective conditions have been satisfied, it is necessary to investigate the subjective conditions. Both the objective and the subjective conditions must be satisfied for the method to be complete. These characteristics make up the prerequisites for a taxonomy's usefulness, but they do not always specify the sufficient requirements. Nevertheless, by crafting strong justifications for a taxonomy's utility, they can provide researchers with direction and serve as bases for descriptive evaluations based on reasoned argument. The method's output, or the taxonomy it produces after the design science building phase, needs to be assessed for usefulness. However, establishing the necessary conditions for usefulness is challenging, and ultimately, determining usefulness may depend on whether or not others find it useful [25].

Table 1 displays the current status regarding the objective ending conditions. While most of the requirements are satisfied, the assessment of the classification of a sample of objects is still ongoing and therefore marked as ‘under evaluation’.

TABLE I. ASSESSMENT OF THE OBJECTIVE ENDING CONDITIONS FOR TOD.

Objective Ending Condition	Yes	No	Under Evaluation
All objects or a representative sample of objects have been examined			×
No object was merged with a similar object or split into multiple objects in the last iteration	×		
At least one object is classified under every characteristic of every dimension			×
No new dimensions or characteristics were added in the last iteration	×		
Every dimension is unique and not repeated	×		
Every characteristic is unique within ist dimension	×		
Each cell (combination of characteristcs) is unique and is not repeated	×		

In addition to the objective ending conditions, [25] suggests that a useful taxonomy has the following subjective attributes:

1. It is concise: Because an extensive classification scheme with many dimensions and many characteristics may exceed the cognitive load of the researcher and be challenging to understand and apply, a taxonomy should only contain a small number of dimensions and a small number of characteristics in each dimension.
2. It is robust: A useful taxonomy should have sufficient dimensions and attributes to distinguish the objects of interest. A taxonomy with few dimensions and traits might not be able to distinguish between objects effectively.
3. It is comprehensive: There are two possible interpretations for this condition. One interpretation is that all known objects within the domain under consideration can be classified by a useful taxonomy (requirement of completeness). The second interpretation holds that all of an object's dimensions should be included in a taxonomy that is useful.
4. It is extendible: When new kinds of objects are discovered, a useful taxonomy should permit the inclusion of new dimensions and characteristics within an existing dimension. A taxonomy that cannot be expanded may quickly become outdated. In other words, it is dynamic rather than static.
5. It is explanatory: A useful taxonomy includes dimensions and traits that aid in our understanding of the objects by usefully elucidating their nature rather than exhaustively describing every aspect of the objects under study or the objects of the future.

TABLE II. ASSESSMENT OF THE SUBJECTIVE ENDING CONDITIONS FOR TOD.

Subjective Ending Condition	Yes	No	Under Evaluation
Concision	×		
Robustness	×		
Comprehensiveness			×
Extensibility	×		
Explanation	×		

Once again, Table 2 shows that most of the required subjective ending conditions are met, while the taxonomy's comprehensiveness is currently still under critical evaluation using real data. Furthermore, we reviewed the results of our systematic literature review considering the more granular level of their proposed features. 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 some general rules during the processing of the data.

a. Removal of types and definitions that are either generic (e.g., yellow press) or too technical (e.g., deep fakes).
b. Removal of duplicates and synonyms to avoid repetitions and overlaps.
c. Removal of types and definitions that were incorrectly categorized as disinformation (e.g., misinformation).

The fact that not all types of disinformation have the same degree of deceitfulness or harmful effects made the step of refining the disinformation taxonomy one of our biggest challenges, and some of them could not be categorized as disinformation. For instance, 'fabrication' is a more serious offense than 'hyperpartisanship' or 'clickbait', the latter of which has generated a lot of discussion. In our most recent iteration, this was addressed by adding a new dimension that deals with the degree of veracity. We completed our research goal, developing a taxonomy of online disinformation after taking the aforementioned information into account. As our goal is to create a useful taxonomy [59], our final test, then, is to examine the resulting taxonomy for its usefulness for the intended users and the intended purpose. The users of the TOD were projected to be researchers, journalists and developers of tools for disinformation detection, and their purpose was to distinguish among truthful and deceptive online content based on linguistic assessment. Nickerson et al. [25] claim that under some circumstances, such as possible collisions in the requirements for a taxonomy to be useful, conflicting criteria may need to be resolved by the researcher. This factor will be taken into account in a later research phase testing the usability of our framework.

IV. PRELIMINARY RESULTS

After our last iteration, we cannot identify any new characteristics and dimensions from the studies under review. Since in our case all ending conditions that can be met before putting the TOD into practice are satisfied, our final framework consists of five dimensions. The first dimension covers **different types of 'Fake News'**, splitting into subtypes (e.g., Trolling or Clickbait) and themes (e.g., pseudoscientific, commercial or political). Our second dimension contains **complexity features** that help to calculate the complexity and readability of the text, giving hints on its truthfulness. It serves users of the TOD to assess the informational content and textual structure of content under examination. A third dimension encompassing **psycho-linguistic features** describes attitudes, personas, behaviors, and emotions. This dimension, which includes the frequency of emotion words and informal language, helps to illustrate and quantify the cognitive process and individual concerns that underlie the writings. With a fourth dimension, we added **stylistic features** that shall reflect the style of the writers and syntax of the text, such as the number of verbs and nouns as well as the usage of certain terminologies. As mentioned before, disinformation content can differ strongly in its deceitfulness. For this reason, our fifth dimension accommodates **grades of veracity** ranging from 'No factual content' to 'Mostly true' to facilitate the evaluation of different kinds of 'Fake News' corresponding with our first dimension.

In the next steps, the current version of our taxonomy will be evaluated by a group of experts in the domain of research on online disinformation. Following on from this, it will be the subject of a workshop in which researchers will develop labels based on the TOD, which will then be used to label a 'Fake News' dataset that will form the basis for training the DeFaktS AI.

V. CONCLUSION AND FUTURE WORK

The research presented in this paper seeks to provide novel perspectives on the rapidly expanding field of combating online disinformation in a methodical and organized manner. Our goal was to discover and categorically define the many underlying linguistic features in the sphere of deceptive information, which was motivated by the lack of a conventionally accepted domain language. The concrete benefit of the developed TOD is, on the one hand, to make the phenomenon of disinformation as such more tangible, to achieve a common understanding of disinformation among researchers in the DeFaktS project, and to help answer the question of how disinformation in social media can be recognized as such. On the other hand,

by unifying numerous study results on the linguistic detection of ‘Fake News’, our taxonomy offers researchers and actors in educational work a framework that provides a systematic overview of the scientific findings from the domain to date. By publishing the TOD and sharing it with a broad community at a later stage, we also hope to contribute to simplifying and standardizing the labeling of data for ‘Fake News’ detection, and thereby making it more transparent.

As we approached this intricate and vast field, we faced some substantial challenges. An issue we encountered was the large amount of research output produced by the latest wave of Big Data, AI, and MLS tools. Despite the abundance of scientific studies in the area, we discovered that the majority of them introduce singular and ad hoc solutions, leading to a fragmentation issue. The main objective of this type of research is still to suggest effective and precise algorithmic approaches as well as to evaluate their performance, so in most cases the justification and conceptual model are not sufficiently explained. In addition, we discovered that depending on its nature, disinformation can vary greatly in its veracity, which may cause difficulties in classification by means of a schema. To resolve this concern, we added a dimension to the TOD called ‘grades of veracity’, allowing us to address the various subtypes and topics that fall under the definition of disinformation. Yet, we anticipate the emergence of new types of disinformation and their associated characteristics given the dynamic nature of the domain, potentially causing the need for a revision of our taxonomy and the dimensions it entails. Because of this, we invite researchers to evaluate and validate the framework in the future to identify potential new dimensions or categories that may alter or extend our work.

We also concluded that multidisciplinary approaches are essential for comprehending and developing strategies and tools to combat the spread of deceptive information. Despite the field's close ties to political communication theory, we think that modern disinformation demonstrates traits that necessitate the use of additional analytical tools. Digital communities that exhibit distinctive traits that are difficult to compare to the past are where disinformation is flourishing [56]. Furthermore, disinformation also encompasses forms outside of the political sphere, such as fake reviews and pseudoscience. Finally, the development of (semi-)automated fact-checking tools is predicted by the recent impressive advancements in technologies like machine learning. These currently observable dynamics call for more interdisciplinary research on the domain that we would like to encourage with our contribution.

REFERENCES

[1] I. Bezzaoui, J. Fegert, and C. Weinhardt, “Distinguishing Between Truth and Fake: Using Explainable AI to Understand and Combat Online Disinformation”, The 16th International Conference on Digital Society, 2022.

[2] 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, 2017.

[3] K. Shu, A. Bhattacharjee, F. Alatawi, T.H. Nazer, K. Ding, M. Karami, and H. Liu, “Combating Disinformation in a Social Media Age”, WIREs Data Mining and Knowledge Discovery, 10, 1–23, 2020.

[4] D. McQuail, “Media performance: Mass communication and the public interest”, Thousand Oaks, CA: Sage. M. Young,The Technical Writer’s Handbook, Mill Valley, CA: University Science, 1992.

[5] J. Strömbäck, “In search of a standard: Four models of democracy and their normative implications for journalism”, Journalism Studies, (6:3), pp. 331-345, 2005.

[6] J. Groshek and K. Koc-Michalska, “Helping populism win? Social media use, filter bubbles, and support for populist presidential candidates in the 2016 US election campaign”, Information Communication and Society, (20:9), pp. 1389-1407, 2017.

[7] H. Holone, “The filter bubble and its effect on online personal health information”, Croatian Medical Journal, (57:3), pp. 298–301, 2016.

[8] K. Sharma, Y. Zhang, and Y. Liu, “COVID-19 vaccines: characterizing misinformation campaigns and vaccine hesitancy on twitter”, Retrieved May 2022. arXiv preprint arXiv:2106.08423, 2021.

[9] E.C. Tandoc Jr, “The facts of fake news: A research review”, Sociology Compass, 13(9), e12724, pp. 1-9, 2019.

[10] L. Munn, “Angry by design: toxic communication and technical architectures”, Humanities and Social Sciences Communications, 7(1), pp. 1-11, 2020.

[11] K. Nelson-Field, E. Riebe, and K. Newstead, “The emotions that drive viral video”, Australasian Marketing Journal, 21(4), pp. 205–211, 2013.

[12] 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.

[13] 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), 87, 2020.

[14] 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”, 11(1), pp. 759–766, 2017.

[15] D. Schreiber, C. Picus, D. Fischinger, and M. Boyer, “The defalsif-AI project: Protecting critical infrastructures against disinformation and fake news/Das Projekt defalsif-AI: Schutz kritischer Infrastrukturen vor Desinformation und Fake News”, Elektrotechnik und Informationstechnik, Vol. 138 (7), pp. 480–484, 2021.

[16] X. Zhou and R. Zafarani, “A survey of fake news: Fundamental theories, detection methods, and opportunities”, ACM Computing Surveys (CSUR), 53(5), pp. 1-40, 2021.

[17] D. Rohera et al., “A Taxonomy of Fake News Classification Techniques: Survey and Implementation Aspects”, IEEE ACCESS, 10, 2022.

[18] W. Shahid et al., “Detecting and Mitigating the Dissemination of Fake News: Challenges and Future Research Opportunities, IEEE Transactions on Computational Social Systems, 2022.

[19] 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.

[20] F. I. Adiba, T. Islam, M.S. Kaiser, M. Mahmud, and M.A. Rahman, “Effect of corpora on classification of fake news using naive Bayes classifier”, International Journal of Automation, Artificial Intelligence and Machine Learning, 1(1), pp. 80–92, 2020.

[21] 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), pp. 34–43, 2020.

[22] M. Fayaz, A. Khan,M. Bilal, and S.U. Khan, “Machine learning for fake news classification with optimal feature selection”, Soft Computing, pp. 1–9, 2022.

[23] 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), pp. 1–7, 2022.

[24] R.A. Hirschheim, H.K. Klein, and K. Lyytinen, “Information Systems Development and Data Modeling: Conceptual and Philosophical Foundations”, Cambridge University Press, Cambridge, 1995.

[25] 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, pp. 336–359, 2013.

[26] R.L. Glass and I. Vessey, “Contemporary application-domain taxonomies”, IEEE Software 12(4), pp. 63–76, 1995.

[27] J. Iivari, “A paradigmatic analysis of information systems as a design science”, Scandinavian Journal of Information Systems 19(2), pp. 39–64, 2007.

[28] J. Delcker, Z. Wanat, and M. Scott, „The coronavirus fake news pandemic sweeping WhatsApp”, Politico, Retrieved May 2022 from https://www.politico.eu/article/the-coronavirus-covid19-fake-news-pandemic-sweeping-whatsapp- misinformation/, 2020.

[29] S. Yu and D. Lo, “Disinformation detection using passive aggressive algorithms”, ACM Southeast Conference, Session 4, p. 324f, 2020.

[30] 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), pp. 881–893, 2021.

[31] M. Mahyoob, J. Al-Garaady, and M. Alrahaili, “Linguistic-based detection of fake news in social media.” Forthcoming, International Journal of English Linguistics, 11(1), pp. 99-109, 2020.

[32] 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), pp. 49-59, 2022.

[33] W. H. Bangyal et al., “Detection of Fake News Text Classification on COVID-19 Using Deep Learning Approaches”, Computational and Mathematical Methods in Medicine, pp. 1-13, 2021.

[34] L. Bozarth and C. Budak, “Toward a better performance evaluation framework for fake news classification”, Proceedings of the international AAAI conference on web and social media, 14, pp. 60–71, 2020.

[35] C. Lai et al., “Fake news classification based on content level features”, Applied Sciences, 12(3), p. 1116, 2022.

[36] A. M. Guess et al., “A digital media literacy intervention increases discernment between mainsitream and false news in the United States and India”, PNAS, 117(27), pp. 15536–15545, 2020.

[37] G. Pennycook and D. G. Rand, “Lazy, not biased: Suceptibility to partisan news is better explained by lack of reasinong than by motivated reasoning”, Cognition, pp. 1–12, 2018.

[38] D. Ribes Lemay et al., “Trust indicators and explainable AI: A study on user perceptions”, IFIP Conference on Human-Computer Interaction - INTERACT 2021, pp. 662–671, 2021.

[39] K. Siau and W. Wang, “Building trust in artificial intelligence, machine learning, and robotics”, Cutter Business Journal, 31(2), pp. 47-53, 2018.

[40] P. Dahlgren, “Media and political engagement: Citizens, communication, and democracy”, Cambridge: Cambridge University Press, 2009.

[41] S. M. Jang and J. K. Kim, “Third person effects of fake news: Fake news regulation and media literacy interventions”, Computers in Human Behavior, 80, pp. 295–302, 2018.

[42] D. Rieger et al., “Propaganda und Alternativen im Internet - Medienpädagogische Implikationen. Propaganda and Alternatives on the Internet - Media Pedagogical Implications“, merz / medien + erziehung, (3), pp. 27-35, 2017.

[43] D. Kellner and J. Share ”Critical media literacy, democracy, and the reconstruction of education”, In D. Macedo & S. R. Steinberg, Media Literacy: A Reader, Peter Lang Publishing, pp. 3-23, 2007.

[44] D. Kellner and J. Share, “Toward critical media literacy: Core concepts, debates, organizations, and policy”, Discourse: studies in the cultural politics of education, 26(3), pp. 369–386, 2005.

[45] J. B. Schmitt, D. Rieger, J. Ernst, H. J. Roth, „Critical media literacy and Islamist online propaganda: The feasibility, applicability and impact of three learning arrangements”, International Journal of Conflict and Violence, 12, pp. 1–19, 2018.

[46] N. Luhmann, „Vertrauen“, Trust (5), UVK, 2014.

[47] F. Schwerter and F. Zimmermann, „Determinants of trust: The role of personal experiences”, Games and Economic Behavior, 122, pp. 413-425, 2020.

[48] D. Castelvecchi, “Can we open the black box of AI?”, Nature, 538, pp. 20–23, 2016.

[49] M. Ter Hoeve et al., „Do news consumers want explanations for personalized news rankings?” FARTEC, pp. 1–6, 2017.

[50] D. Shin, B. Zhong, F. A. Biocca, “Beyond user experience: What constitutes algorithmic experiences?” International Journal of Information Management, 52, pp. 1–11, 2020.

[51] D. Shin, “The effects of explainability and causability on perception, trust and acceptance: Implications for explainable AI”, International Journal of Human-Computer Studies, 146, pp. 1-11, 2021.

[52] T. Schmidt, F. Biessmann, T. Teubner, „Transparency and trust in artificial intelligence systems”, Journal of Decision Systems, 29(4), pp. 260–278, 2020.

[53] K. Peffers, T. Tuunanen, M. A. Rothenbergre, S. Chatterjee, “A design science research methodology for information systems research”, Journal of Management Information Systems, 24(3), pp. 45–77, 2007.

[54] H.A. Simon, “The Sciences of the Artificial”, The MIT Press, Cambridge, MA, 1969.

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

[56] E. Kapantai, A. Christopoulou, C. Berberidis, and V. Peristeras, “A Systematic Literature Review on

Disinformation: Toward a Unified Taxonomical Framework", New Media & Society, 23(5), pp. 1301–1326, 2021.

[57] W.Y. Wang, "liar, liar pants on fire: A new benchmark dataset for fake news detection", arXiv preprint arXiv:1705.00648, 2017.

[58] M. Szpakowski, "Fake News Corpus Dataset", https://github.com/several27/FakeNewsCorpus, 2020.

[59] A.R. Hevner, S.T. March, J. Park, and S. Ram, "Design science in information systems research", MIS Quarterly 28(1), pp. 75–105, 2004.