Received: 25 August 2023 | Revised: 31 January 2024 | Accepted: 1 February 2024

DOI: 10.1002/rhc3.12296

ORIGINAL ARTICLE

RHCPP Risk, Hazards & Crisis in Public Policy WILEY

Navigating misinformation in voice messages: Identification of user-centered features for digital interventions

Katrin Hartwig | Ruslan Sandler | Christian Reuter

Science and Technology for Peace and Security (PEASEC), Technische Universität Darmstadt, Darmstadt, Germany

Correspondence
Katrin Hartwig, Science and Technology for Peace and Security (PEASEC), Technische Universität Darmstadt, Pankratiusstraße 2, Darmstadt 64285, Germany.
Email: hartwig@peasec.tu-darmstadt.de

Funding information
Bundesministerium für Bildung und Forschung, Grant/Award Number: 13N16361; German Federal Ministry for Education and Research (BMBF), Grant/Award Numbers: 13N16361, 13N16636; German Federal Ministry of Education and Research; Hessian Ministry of Higher Education, Research, Science and the Arts; National Research Center for Applied Cybersecurity ATHENE

Abstract

Misinformation presents a challenge to democracies, particularly in times of crisis. One way in which misinformation is spread is through voice messages sent via messenger groups, which enable members to share information on a larger scale. Gaining user perspectives on digital misinformation interventions as countermeasure after detection is crucial. In this paper, we extract potential features of misinformation in voice messages from literature, implement them within a program that automatically processes voice messages, and evaluate their perceived usefulness and comprehensibility as user-centered indicators. We propose 35 features extracted from audio files at the character, word, sentence, audio, and creator levels to assist (1) private individuals in conducting credibility assessments, (2) government agencies faced with data overload during crises, and (3) researchers seeking to gather features for automatic detection approaches. We conducted a think-aloud study with laypersons ($N = 20$) to provide initial insight into how individuals autonomously assess the credibility of voice messages, as well as which automatically extracted features they find to be clear and convincing indicators of misinformation. Our study provides qualitative and quantitative insights into valuable indicators, particularly when they relate directly to the content or its creator, and uncovers challenges in user interface design.

KEYWORDS

countermeasure, disinformation, fake news, misinformation, user intervention, voice messages

INTRODUCTION

Recent crises, such as the Russian–Ukrainian conflict and the COVID-19 pandemic, have highlighted the impact of an excess of both reliable and incorrect information, particularly on social media and messaging platforms such as Telegram. Managing this substantial volume of information poses a challenge for governmental policies and societies that are resilient in times of crisis. Indeed, misinformation has the potential to deceive, lead to polarization (Levendusky, 2013; Osmundsen et al., 2021), manipulate elections (Kalsnes, 2018), and thus pose a challenge to democracy. In alignment with previous studies (Almaliki, 2019; Chen et al., 2022; Li et al., 2022; Wang et al., 2019), the term misinformation will be employed as a comprehensive umbrella term that encompasses both intentionally fabricated deceptive information ("disinformation" or "fake news") and unintentionally generated misleading information ("misinformation").

Much research has been conducted on the benefits of social media in disruptive situations, as it functions as a vital information channel (Gui et al., 2017). However, content posted on social media platforms (such as TikTok videos or posts on X [formerly Twitter]), and voice messages sent through messaging applications like Telegram or WhatsApp, often contribute to the dissemination of misinformation both during crises and in day-to-day life (Ng & Loke, 2021). Indeed, misinformation in voice messages, particularly within popular messenger apps, such as Telegram and WhatsApp, has emerged as prevalent and concerning phenomena (El-Masri et al., 2022). With messengers implementing the feature of sending and receiving voice messages (e.g., WhatsApp, 2013), it is now an extensively used feature for a majority of users (El-Masri et al., 2022; WhatsApp, 2022). As text, images, and videos, voice messages are a common format to spread misinformation (Resende et al., 2019). For example, in April 2020, a WhatsApp voice message sharing misleading information about the anticipated death tolls due to COVID-19 and subsequent medical response in the UK circulated.[1] Indeed, the audible nature of voice messages with its opportunity to demonstrate and evoke emotions through comes with unique implications for misinformation spread (El-Masri et al., 2022). Given the widespread use of the specific format of voice messages and the potential impact on public perception, it is crucial to develop effective countermeasures for maintaining information integrity and fostering a more informed digital society, mitigating the consequences of misleading and false information. The appropriate handling of misinformation in voice messages is significant from various standpoints. First, as messenger users, individuals are presented with a multitude of information and face the task of discerning accurate information from misinformation. Especially in uncertain periods, such as during crisis situations, for example, floods, affected persons, and helpers seek answers in public messenger groups, thus increasing their susceptibility to fake news. On the other hand, official authorities and decision makers encounter the challenge of identifying and rectifying rumors and misinformation rapidly, particularly in times of crisis. Technical approaches can assist with managing the abundance of information. The user-centered development and evaluation of technical interventions is noteworthy from both a public policy and Human–Computer Interaction (HCI) research viewpoint.

The ability to share information across messenger groups enhances the pace of information dissemination (Davies, 2020). The use of speech can convey emotions that may unintentionally cause individuals to propagate false information. Moreover, Wallbridge et al. (2021) have illustrated that individuals tend to pay more attention to the method of delivery rather than the content itself. This work adds to existing findings on misinformation in text by emphasizing and addressing the significance of misinformation in voice messages as a prevalent phenomenon, facilitating the education of individuals on characteristics to evaluate the trustworthiness of voice messages. The goal of this study is to expand the present research in public policy by creating a proof-of-concept digital intervention as a technological solution after (manual or automatic) detection with the potential to enhance users' media literacy, and obtaining qualitative and quantitative user-centered feedback on its comprehensibility and usefulness. Addressing users' need for transparency in interventions (Kirchner & Reuter, 2020), we concentrate on using comprehensible indicators as cues for misinformation in the evaluation of voice messages. Indicator-based interventions have already been developed and discussed for other modalities, such as text (Martino et al., 2020). Hence, we will extend these interventions to voice messages, taking into account their unique attributes such as tonality and speech rate. With voice messages as a common way of communication for users of all ages, we look at it from the perspective of both government agencies and private individuals. We researched the existing literature on the characteristics of misinformation in both written and spoken language to gain insight into the attributes of voice messages that could indicate misinformation. Additionally, we created a digital intervention as a proof-of-concept that automatically identifies multiple features in voice messages as possible indicators of misinformation. To gain insight into the potential usefulness of various features as indicators of voice message credibility for users, a user study was conducted using the think-aloud method.

We enhance public policy and HCI research by utilizing the promising existing knowledge of indicator-based interventions, which link media literacy enhancement and technological solutions after automatic or manual detection of misinformation, and applying them to the modality of voice messages as a modality with unique misleading potential. This approach is relevant for the design principles of policy frameworks and can be crucial for public authorities and private users, particularly in relation to public messenger groups. Our study provides initial insights into the perception of the approach by a diverse group of private users. There is potential for future research to extend these findings to practitioners from public authorities. Our main contributions (C) and findings (F) consist of (C1) a systematic identification of voice message features that may serve as potential indicators of misinformation. We (F1) identified 35 specific features on levels of characters, words, sentences, audio, and creators. Additionally, (C2) we developed a proof-of-concept misinformation intervention based on these indicators as a potential media literacy-enhancing technological solution and public policy response to the misinformation challenge, and (C3) assessed the perceived comprehensibility and usefulness of numerous features for laypersons listening to voice messages, which revealed that (F2) features directly linked to the content or creator are deemed particularly useful, as opposed to more abstract attributes of the voice message. The study also highlights (F3) the challenges and limitations of the indicator-based approach for voice messages, specifically regarding its presentation to users.

RELATED WORK

Our research adds to the body of research on digital, user-centered interventions that combat misinformation by promoting transparency and comprehensibility, thereby enhancing media literacy as a public policy response to the misinformation challenge. Related work has already demonstrated the significance of misinformation in social media, particularly during crises. Initial technical support approaches for dealing with this phenomenon have been developed and evaluated. Indicator-based approaches are a particularly promising option for enhancing media literacy. In accordance with other researchers, in this work we define media literacy as the ability to decode, evaluate, analyze, and produce both print and electronic media, that is, to have internalized a sense of "critical autonomy" in dealing with all media (Aufderheide, 1993). In the following, we discuss the relevance of misinformation in social media for public policy and how it motivates our work (see Section 2.1). We further review related work that addresses a demand of current research for technological solutions that facilitate media literacy (see Section 2.2), and literature on features of misinformation regarding both text and audio, serving as a basis for our indicator-based approach (see Section 2.3). We add to related findings by developing a proof-of-concept misinformation intervention for the assessment of voice messages based on the identified features and evaluating its user-centeredness as a technological solution.

Public policy perspective on misinformation

Social media is indispensable for information exchange during daily life. It comprises providers like Facebook, X (formerly Twitter), Instagram, and TikTok, among others, but also messaging apps like Telegram with 700 million and WhatsApp with 2 billion active users monthly in 2023 (Statista, 2023). Not only in everyday life but also in crises such as floods or pandemics, social media is crucial for both individuals and authorities (Clark et al., 2024), facilitating attempts to improve safety and security (Reuter & Kaufhold, 2018). The research field addressing social media during crises is often referred to as crisis informatics (Hagar, 2006; Palen et al., 2007) and combines findings from various disciplines such as computer science, HCI, social science, and political science. As an important subdiscipline, public policy and administration (Hildreth et al., 2007) investigates implementations of policies for public services like emergency services (Reuter, 2022). Indeed, governments and authorities are now increasingly confronted with the usage of information technologies including social media, for example, when aiming to engage with citizens during crises. While there are different theories on the relationship between technology and public policy and administration, the sociotechnological theory is most commonly used (Reddick, 2012). It denotes the interrelation of social and technical aspects, meaning that technological progress, organizational needs, and individuals' attitudes shape the utilization of information technology within an organization, thereby impacting social change (Reddick, 2012; Reuter, 2022).

While crisis informatics and related research fields have pointed out the advantages and potentials of social media, it comes with several challenges and risks as well (Kaufhold et al., 2019). The rapid information access, large amount of data, and (supposedly) anonymity pose a challenge to assess credibility and recognize misinformation. This applies to individuals as well as officials like crisis managers who create and analyze content (Imran et al., 2017). For instance, messaging apps with public channels do not only facilitate aid networking during crises but also face criticism for permitting unrestricted spread of conspiracy theories and propaganda (Gunz & Schaller, 2022; Herasimenka et al., 2023). Governments around the world have identified misinformation as a serious threat to national security, prompting calls for regulation of internet technologies (Jackson, 2021; Schulte & Pickard, 2020). Disinformation campaigns, for instance, those launched by Russia and China in relation to the COVID-19 pandemic (Brovdiy, 2020), impact international diplomatic relations, foreign policy, but also other sectors such as security policy and health policy. As a consequence, governments aim to strengthen confidence in their statements and empower citizens to recognize accurate information (Connolly et al., 2019). Komendantova et al. (2023) use the example of the migration discourse to explain the connection between misinformation and policymaking in that misinformation leads to a false public representation of an issue. The general public, who can participate in and influence legislation in various ways, is then prevented from adopting positions grounded in evidence-based decisions due to the inadequate information landscape. Synthesizing knowledge from misinformation research and public policy can be considered a crucial step towards "informed recommendations to [...] public administrators" (Connolly et al., 2019, p. 1). There are three main demands for policy responses to the misinformation challenge in social media identified and discussed in literature, to which our work strongly picks up on: (1) enhance users' media literacy skills, (2) promote technological solutions and self-regulation by platforms, and (3) impose liability laws on platform operators to remove false or misleading content (Medeiros & Singh, 2020; Schulte & Pickard, 2020; Tambini, 2017). Our work aims to specifically offer support to public policy approaches targeting media literacy and platform-specific technological solutions by investigating an integration of both. To achieve this, we apply an indicator-based approach that uses characteristics of misinformation in voice messages to assist users in making informed decisions about credibility. In the following, we give an overview on related research on indicators of misinformation that highly motivated our work.

Indicator-based misinformation interventions

Developing effective measures to combat misinformation is challenging, both from a political perspective as well as from an HCI perspective. Nonpunitive policy interventions that enhance users' media literacy skills as technological solutions that may be applied as self-regulation by platforms (Medeiros & Singh, 2020; Schulte & Pickard, 2020; Tambini, 2017) are deemed more appropriate than policy approaches that are mainly limited to sanctions and carry the risk of collateral censorship because platform operators are incentivized to excessively delete content to avoid penalties (Balkin, 2017). This is relevant also from a user perspective as users strongly prefer explainable handling of misinformation over simply deleting or flagging content without explanations (Kirchner & Reuter, 2020). Due to its user-centered and potentially media literacy-enhancing nature, this approach may be incentivized through policy in a liberal democracy and has been emphasized as a necessary step to "prevent overzealous regulation of speech platforms by the government" (Medeiros & Singh, 2020, p. 288), for instance, in India. While it is insufficient if taken separately as a focus on purely technical solutions neglects societal factors (Medeiros & Singh, 2020), digital misinformation interventions are developed to complement critical journalism and media literacy education in schools. Technological solutions can take various forms (Hartwig et al., 2023) and differ in their suitability to address the demands for policies encouraging media literacy and laws on platforms, including digital interventions that display a warning or correction next to disputed content without hiding the misinformation itself (Bhargava et al., 2023), default nudges to promote critical thinking for all social media posts (Bhuiyan et al., 2018), or machine-learning-based approaches that automatically identify misinformation and delete or hide it (Shu et al., 2017b). In this work, we pick up on the idea of postdetection decisions from a user perspective or an approach detached from detection (e.g., via default nudge to encourage reflection). Research suggests that transparency is crucial when aiming to establish trust among users in digital interventions (Kirchner & Reuter, 2020) and minimize reactance or other backfire effects (Nyhan & Reifler, 2010)—this strongly motivated our work.

Addressing policies to encourage media literacy, efforts have been made to develop digital interventions encompassing the user-centered display of indicators as comprehensible characteristics of misinformation to offer a guidance in autonomous assessment strategies while promoting critical thinking and trust (Bhuiyan et al., 2021). For instance, Bhuiyan et al. (2021) take up the idea of indicators by investigating their utilization like information about the author of social media posts from the perspective of journalists and news consumers. Indicators for the credibility of social media posts have been investigated regarding utility and user preferences for visualization, revealing positive feedback as long as the display of indicators was rather simple (Grandhi et al., 2021). Indeed, research has demonstrated that indicators as user feedback can enable the development of own assessment skills (Schmid et al., 2022), reduce uncertainty (Grandhi et al., 2021), and fit pre-existing mental models and practices of users (Sherman et al., 2021). The vast majority of research has focused on textual content like posts on X or Facebook (Hartwig et al., 2023), and some extend to images and videos (Bhargava et al., 2023; Hartwig et al., 2024; Sherman et al., 2021). For example, Sherman et al. (2021) identified the source of information as key indicator when assessing different types of content, including fake videos. Hartwig et al. (2024) similarly examine the perceived usefulness and comprehensibility of misinformation indicators in TikTok videos within a simulated smartphone app. Messaging apps like WhatsApp and Telegram offer the possibility to share voice messages in private conversations and (public) channels, encompassing an additional prevalent modality of information that nonetheless is confronted with misinformation. Research in (indicator-based) interventions has examined that context less exhaustively, however providing some related studies that guided our approach. For instance, Burgoon et al. (2003) use the audio modality to identify characteristics of deceptive speech from discourse, examining both asynchronous (e.g., text chat) and synchronous (e.g., face-to-face) forms of communication, and Maros et al. (2021) analyzed misinformation in voice messages and identified indicators we build upon. In the following, we will present research that tackles features of misinformation in various modalities and led to the categorization of characteristics of voice messages guiding our approach (see Appendix A).

Features of misinformation in text and audio

To develop indicator-based misinformation interventions aiming to encourage media literacy within a technical solution after manual or automatic prefiltering, it is necessary to form a knowledge base on typical characteristics of misinformation. Our work focuses on the context of misinformation in voice messages—thus in the following we discuss related research on features of the audio modality that informed our work. As the research landscape of textual content is by far more exhaustive and may be applicable to voice messages when using their transcript, we will further consider insights on textual features. Characteristics of misinformation discussed in related work helped us derive an overview of those potentially applicable to voice messages (see Appendix A) as a foundation for our technical implementation.

Text features

Considering text features for misinformation detection also involves exploring related research areas (Shu et al., 2017b). These include authorship recognition (Abbasi & Chen, 2008; Afroz et al., 2012; Zheng et al., 2006), deception detection (Afroz et al., 2012; Feng et al., 2012; Rubin & Lukoianova, 2015), clickbait detection (Chakraborty et al., 2016; Potthast et al., 2016; Shu et al., 2017b), and hyperpartisanship detection (Potthast et al., 2017).

For instance, people spreading misinformation might alter their writing style to avoid attribution. Authorship recognition and deception detection analyze style changes to uncover the original author (Zheng et al., 2006). Zheng et al. (2006) classify writing-style features into lexical, syntactic, structural, and contentspecific categories for authorship recognition. They seek features that remain consistent across texts from the same writer. Building on that and along with insights from deception and authorship studies (Brennan & Greenstadt, 2009; Burgoon et al., 2003; Hancock et al., 2008), Afroz et al. (2012) investigate deceptive writing style. The study demonstrates that deceit detection benefits from contentrelated and noncontentrelated aspects, including authorship mimicry and masking. Feng et al. (2012) concentrate on deception detection, particularly in product reviews. Their study employs features like "term frequency—inverse document frequency" for word uni- and bigrams, as well as tf-idf of shallow and deep syntax (utilizing Part-of-Speech [PoS] tags and Probabilistic Context Free Grammar). Both word features and deep syntax aid deception detection, with optimal performance when combined. The connection between authorship recognition and misinformation detection stems from the notion that misinformation spreaders possess distinct writing styles from those sharing genuine information. This suggests treating misinformation detection as an authorship recognition task to differentiate these types of spreaders. Hyperpartisan detection, closely linked to misinformation research, deals with one-sided messages omitting pertinent information (Potthast et al., 2017). Potthast et al. (2017) examine whether stylometry can reliably distinguish hyperpartisan from mainstream news, and explore its potential for misinformation detection. They find hyperpartisan articles are shorter on average, with a similar style between left- and right-wing documents. However, stylometry alone struggles to consistently identify misinformation.

Clickbait manipulates readers into clicking on links and reading articles, constituting a form of deception (Chen et al., 2015, p. 15). Chakraborty et al. (2016) analyze features distinguishing clickbait titles from nonclickbait ones. They note that clickbait titles generally exhibit longer syntactic dependencies, attributed to higher grammatical complexity. Additionally, clickbaits commonly contain words with "very positive" sentiments, like "awe-inspiring." In contrast, Potthast et al. (2016) assess clickbaits in tweets across three feature categories: the teaser message, linked web page, and associated meta information. They identify meaningful features such as tweet-wide sentiment polarity, stop-words-to-words ratio, and whether the tweet starts with a number.

Audio features

The comprehensive research on features of misinformation in text stands in contrast to a less examined research field of audio misinformation. While there is research that includes the audio track of videos in the detection of misinformation (Shang et al., 2021), this research uses black-box approaches, and is therefore not applicable to extract potentially explainable features. Others that use audio modality to identify features in deceptive speech, focus on an interrogation scenario (e.g., Burgoon et al., 2003; Levitan et al., 2016). Burgoon et al. (2003) aim to detect deception from discourse, where they examine both asynchronous (e.g., text chat) and synchronous (e.g., face-to-face) forms of communication. The researchers divided indicators into four types: *quantity*, *vocabulary complexity*, *grammatical complexity*, and *specificity and expressiveness*. They concluded that "deceivers do utilize language differently than truth tellers" (Burgoon et al., 2003, p. 91). Furthermore, it was observed that indicators may vary according to the mode of communication: participants employed fewer conjunctions when using text chat than when using audio chat. Nevertheless, the research team remarked that the participants had limited time to prepare for synchronous communication. As a result, it is suggested that given ample preparation time, the participants may have used language differently. This observation highlights a significant contrast between synchronous discourse and the dissemination of misinformation through voice messages. Voice messages are unidirectional, providing the spreaders with sufficient time to prepare and rehearse the delivery before dispersal.

Furthermore, this observation is supported by the research of Maros et al. (2021)—who analyzed misinformation in voice messages—as they found indicators that contradict the findings of Burgoon et al. (2003). One such example is the number of words spoken by deceivers: While Burgoon et al. (2003) found that deceivers said "less than truth tellers" (Burgoon et al., 2003, p. 94), Maros et al. (2021) found that voice messages containing misinformation were longer on average (Maros et al., 2021, p. 9). Another difference is that, while Burgoon et al. (2003) found that deceivers use less emotional language (Burgoon et al., 2003, pp. 94–95), Maros et al. (2021) found that voice messages with misinformation tend to contain more negative emotions (Maros et al., 2021, p. 9).

While many features can be derived from text as transcribed voice messages, others are based on the audio modality. Research has found that speech rate and sound volume may be used as an indicator when assessing the credibility of audio material (El-Masri et al., 2022)—an insight we build on in our user study.

We have collated the characteristics highlighted in previous research pertaining to the identification of misinformation in text and speech tasks, which are outlined in the appendix (see Appendix A). The properties that can be deduced from voice messages are categorized into five levels: character, word, sentence, audio, and creator level. In this study, we facilitate research into the suitability of these characteristics for voice messages by providing a digital intervention that automatically extracts them from audio material. We then present initial findings on their comprehensibility and perceived usefulness.

Research gap

Governments around the world have identified misinformation as a serious threat to national security, prompting calls for regulation of internet technologies (Jackson, 2021;

Schulte & Pickard, 2020). This study advances research in public policy regarding a user-centered perspective on misinformation interventions. It accomplishes this by identifying potential indicators of misinformation in voice messages and integrating them into an automated program, aiding further studies on indicator-based interventions. We address two gaps:

First, while misinformation investigations often focus on written content, the relevance of voice messages in messaging apps (e.g., Telegram and WhatsApp) during recent crises highlights the need to consider them as substantial sources of information. For instance, Telegram groups have been significant in propagating conspiracy content (Herasimenka et al., 2023), and misinformation propagated via mass-messages on WhatsApp is perceived to have functioned as a catalysator for lynchings in India (Medeiros & Singh, 2020). This emphasizes the necessity to expand the user-centered technological perspective (*first gap*).

Second, interventions combating digital misinformation have shown promise when providing transparent indicators or explanations (Kirchner & Reuter, 2020). Overcoming reactance and meeting users' need for comprehensibility have been underscored (Kirchner & Reuter, 2020; Müller & Denner, 2019; Nyhan & Reifler, 2010). While indicators for text-based content have been partially explored (Ayoub et al., 2021; Furuta & Suzuki, 2021; Martino et al., 2020; Papadopoulou et al., 2022), their application to voice messages and audio material is incomplete (*second gap*). While existing research examines misinformation in speech, mostly in impromptu discourse (Section 2.3.2), we expand on studies analyzing misinformation characteristics in voice messages (Maros et al., 2020, 2021) from a technical and user-centered standpoint, including a proof-of-concept. Our prototype solution is extended by qualitative findings (and supplementary quantitative insights) to demonstrate technology's potential in managing the overabundance of information, researchers, crisis managers, and messenger users face.

We address the research gaps by answering the following research questions:

RQ1: *How can an automated tool be designed to extract features of misinformation in voice messages, supporting public policy efforts for technical solutions that enhance media literacy?*

RQ2: *How do users evaluate the comprehensibility and usefulness of the automatically computed indicators when assessing the credibility of voice messages?*

Our contribution comprises two main aspects. In Step 1, we created an extendable automated program that calculates potential indicators of misinformation from voice messages using features identified in relevant literature (see Section 3). Subsequently in Step 2, we offer preliminary insights into user-perceived comprehensibility and usefulness of these indicators through a think-aloud study (see Section 4).

STEP 1: AUTOMATICALLY EXTRACTING FEATURES OF MISINFORMATION IN VOICE MESSAGES

Technical approaches that support crises authorities, crisis management organizations, laypersons, and researchers in dealing with misinformation can contribute to media literacy development, complementing training and professional journalistic work and greater crisis resilience within society. This is especially the case if the solutions use transparent and understandable explanations or indicators. We refer to *features* as specific characteristics of voice messages that have a value during the analysis. For instance, one can analyze and compare features like "number of unique words" between multiple voice messages. In contrast, we refer to *indicators* as specific features of voice messages that are explainable to end users and indicate that a voice message contains misinformation (e.g.,"The voice message contains common buzzwords related to conspiracy theories").

We designed and implemented a program using *Python* as the backend for such a technical approach. In the future, a user-centered frontend and, for example, an integration within a smartphone app or desktop application could expand the backend. The program receives voice messages and processes them to output features and explanations, as potential indicators for misinformation. For now, we provide Command Line Interface functionality. In total, the envisioned usage scenarios are as follows:

(1) A layperson wants to analyze a voice message in terms of credibility. This requires functionality that generates indicators for authentic information and for misinformation when given an audio file.
(2) A crisis management organization or authority needs to analyze a large data set (e.g., to find indicators for misinformation and identify trends during a crisis). This requires human-readable output, and functionality that supports the analysis.
(3) A researcher needs to incorporate features for a detection model. This requires functionality that, given audio files, generates features that may be relevant to distinguish between authentic information and misinformation. Furthermore, the functionality should be easily extendable and maintainable to accommodate any future application upgrades by researchers.

The program accepts the following types of input data:

- a single audio file to extract features for a downstream detection model from voice messages or other audio material,
- a single text file (as a previously transcribed audio file) to extract features for a downstream detection model from transcribed voice messages or other text material,
- a directory that contains a data set of text files (as previously transcribed audio files) to allow researchers to analyze its features for detection or indicators for end users,
- previously extracted features to save time and compute during the analysis of data sets.

Given the input data, the program uses rule-based and traditional statistical approaches to compute indicators of misinformation to facilitate explainability. This allows users (e.g., laypersons or crises authorities) to check the credibility of voice messages and understand whether—and how—they might be designed to deceive. We generate an explanation for each feature on the server to make them easier to understand for end users.

Data processing

When audio input is being processed, the program first performs speech-to-text by using *Whisper* (Radford et al., 2022). This step is skipped when text input is being processed Then, given (extracted) text, *spaCy* (Honnibal et al., 2020) performs preprocessing, including tokenization, lemmatization, PoS tagging, and dependency parsing, which is required to compute word- and sentence-level features. In contrast, character-level features are computed directly on plain text. Thereafter, the computed features are stored in a human-readable .JSON file. Lastly, if features are regarded as comprehensible indicators of misinformation and explanations are necessary, the program extracts these indicators from the generated instance. In case of data set input in contrast to single files, the processing elaborated above is executed for all files in the data set to create one instance per file. After all instances are created, we store the instances in a *Data set* instance and compute averaged features. The outputs as human-readable .JSON files may be opened (e.g., in a browser) for analysis by professional members of crises authorities or researchers. They can then be sent to a detection model for training or testing, or to a digital application as a whole (e.g., a smartphone app) to be displayed to end users. You can find the data flow of the program in Figure 1.

Computed features

The choice of features we compute in our program is motivated by previous research we summarized in Section 2. You can find the list of features in Table A1, Appendix A. While we included a broad variety of features on character, word, and sentence level, this selection is not exhaustive and may be extended based on a systematic literature review, especially on the explicit speech modality that may imbue emotions. Existing research indicates that speech rate and sound volume may compose a relevant attribute of voice messages when assessing its credibility (El-Masri et al., 2022). While these features are not yet included in the prototype implementation, they were part of our user study. For the implemented feature of "keywords for a specific topic," we specified this for our user study based on existing research. Specifically, we included whether there was a "claim to be an expert" (e.g.,"I have researched that!") (El-Masri et al., 2022) or a "call for action" (e.g.,"Share this voice message!") (El-Masri et al., 2022; Maros et al., 2021). Additionally, we considered the existing research on the relevance of creator's profiles and thus included whether a voice message was sent by a "suspicious profile (e.g., name, description of the creator)" which might be detectable via keywords or the excessive usage of emojis.

Like Potthast et al. (2017, 2016), we use the *General Inquirer* dictionary to compute features that depend on linguistic word categories (e.g. words that express uncertainty). The categories that are used to analyse a file or data set can be specified in the configuration file. For instance, words related to certainty or uncertainty were derived from the categories "SureLw" and "If," words related to affect were derived from the categories "AffTot," "NegAff," and "PosAff," and pronouns were derived from "Self," "Our," and "You."[2]

STEP 2: A USER PERSPECTIVE ON THE INDICATOR-BASED APPROACH

Having implemented a program that allows to automatically extract a plethora of features from voice messages as potential comprehensible indicators to assist users in dealing with misleading information, as a next step we gained initial qualitative and quantitative insights into the comprehensibility and perceived usefulness of indicator-based approach. In the following, we describe the study design and derived results.

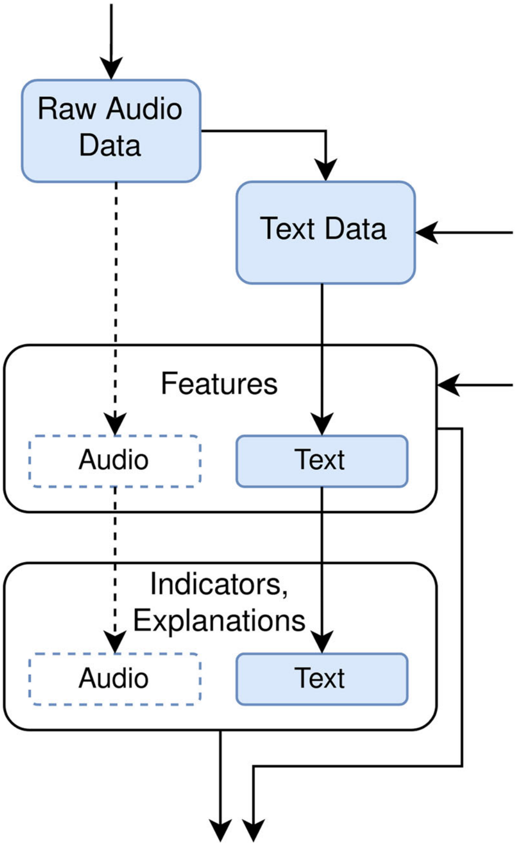


FIGURE 1 The program's data flow. While the solid lines represent current data flow, the dotted lines and boxes represent data flow the program may be extended by in follow-up research.

Study design

We conducted a user study (prestudy $N = 5$, main study $N = 15$) using the think-aloud method, where participants were asked to express their thoughts aloud while listening to audio material and answering questions or performing tasks (Fonteyn et al., 1993). Participants were assured that there is no right or wrong, we are rather interested in their subjective perceptions. In said study we aimed to gain insights into (1) which aspects participants autonomously pay attention to when assessing the credibility of voice messages without any assistance, (2) if the participants' perception towards the information's credibility changes when we additionally display automatically computed features from our developed Python program, and (3) which features the participants perceive as comprehensible or convincing indicators for misinformation when assessing the credibility of voice messages.

Selection of stimuli

The stimuli in our user study consist of four voice messages and corresponding features. See Table A2 for a detailed description.

Voice messages: A systematic approach was used to identify stimuli. In the absence of accessible databases of voice messages, official fact-checking websites were first checked to determine which topics were currently circulating and had been officially debunked. Then, open telegram groups were searched for voice messages on these topics. Two real-world voice messages could be identified this way. We made sure to include different political perspectives, however only politically right-wing misinformation was found in the exploratory search in messenger groups. Thus, we supplemented our stimuli with two artificially generated voice messages. For this purpose, corresponding misinformation in the form of other modalities (Twitter post and Youtube video) was identified on fact-checking sites and, based on this, voice messages were generated by a researcher. Typical characteristics of voice messages were taken into account, such as a length of at least 1:45 min (Maros et al., 2020).

Features: We generated the features as potential indicators for misinformation and their explanations as follows: (1) We used *Whisper* (Radford et al., 2022) to transcribe the four audio files we collected or generated. (2) Given the transcribed text for each of the audio files, we used our Python program̀s *get-features* command to extract features, such as "number of words," and "number of words that express uncertainty." (3) This creates four *.JSON* files that contain features grouped by character, word, and sentence level. Additionally, we included features on audio and creator level that are known from literature and are not (yet) included in our Python program. These features were extracted manually. We did not include all computable features in our user study, as especially for those on character level it is foreseeable that this is not suitable for reporting to private individuals but is more suitable for practitioners who deal with the subject in depth (e.g. researchers developing a detection method). It is crucial to emphasize that the features are not defining for misinformation but can appear and be computed for reliable information as well. As in other indicator-based misinformation interventions, the approach relies on a preceding manual or automatic detection of misinformation as prefiltering (see Section 2.2). For a full list of features we presented in the study, see Table A1, Appendix A.

Participants

In total, we recruited 20 participants (prestudy $N = 5$, main study $N = 15$). As we only used the data of our main study for explicit evaluation, we present the participant's characteristics of the main study with $N = 15$ in the following. The participants' age ranges from 19 to 63 years (*Median* = 32) and cover a diverse set of educational levels (e.g., high school diploma, professional training, and university degree). We explicitly decided against recruiting solely college students as these are already overrepresented in misinformation studies due to their easy accessibility and they do not represent the average user. Of the 15 participants, seven were male, six were female, and two were of diverse gender. All but two participants reported German as their native language. Twelve participants reported to receive voice messages at least once a month, and four participants report they might have received misinformation per voice message before. Nine participants stated to currently use at least one messenger group with more than 50 members. As two of our stimuli address misinformation and reliable information on the topic of the Russian war against Ukraine, we further evaluated central general attitudes towards these topics and a general political orientation of our participants. All but one participants fully agree with Russia waging a war of aggression against Ukraine. We asked our participants to place themselves on a scale from 1 (extreme left) to 11 (extreme right) regarding politics. Three participants placed themselves on the far left between 1 and 2, three rather left between 3 and 4, nine in the center between 5 and 7, and none further right than this. Participants were acquired through the panel provider *Prolific* and got an expense allowance of €9 for an average duration of 45 min.

Study procedure

The one-on-one think-aloud sessions were conducted online via *Zoom* in summer 2023. (1) After a brief introduction and obtaining written consent, the participants were (2) forwarded to an online survey for collecting demographics and general usage behavior of messengers. Afterwards, (3) the researcher subsequently displayed four voice messages in random order while sharing the screen. Participants were able to see the original interface of a Telegram group where the voice messages are displayed. After each voice message, the researcher began audio recording the session and the participants were verbally asked to rate the credibility of the voice message. Participants were instructed to keep thinking aloud and were told that there was no wrong or right, and that we were only interested in their individual assessment (Fonteyn et al., 1993). (4) The researcher then subsequently displayed the calculated features and asked the participants to rate its comprehensibility and usefulness the assess the voice message's credibility giving explanations. Then, participants were asked to (5) rate the credibility of the voice message again. The researcher clarified the origin and truthfulness of the voice message. Finally, (6) participants were asked to give an overall evaluation of the features and the general approach.

Analysis

Step 2 generated rich qualitative data and complementing quantitative survey ratings on a scale from one to five for comprehensibility and perceived usefulness. This allows for a triangulation of data, combining quantitative survey ratings with qualitative explanations. We transcribed the audio records using Whisper (Radford et al., 2022), followed by a manual revision and anonymization of participants' responses. We analyzed the quantitative survey items calculating descriptive statistics (*mean, median, and frequencies*) and clustered the verbally given explanations according to the referenced features thematically using *MaxQDA* where a qualitative content analysis was conducted. Codes refer to the individual features and were created before the analysis. Additional assessment strategies of participants were added iteratively as novel codes when they emerged during coding.

Results

We report how a diverse set of laypersons as messenger users assess the credibility of voice messages autonomously, and how they assess the comprehensibility and perceived usefulness of a plethora of features.

Autonomous assessments

While users have many assessment strategies regarding layout, interactions, and reactions for other modalities (e.g., textual social media posts), they seem to rely mostly on their knowledge when assessing voice messages and only partly relying on characteristics like the usage of specific words or tonality. Specifically, all 15 participants stated to at least partially use their *existing knowledge or gut feeling* when assessing the voice messages in our study. Two participants explicitly stated to consider if the *speaker uses swear words or other negative expressions*. Particularly when *conspiracy theories* were addressed in the voice message, participants were very capable of emphasizing how those keywords (e.g., “BRD GmbH” as a common conspiracy theory in Germany) influenced their assessment of the voice message. Some described how the voice message seems *contrived and disjointed* and, thus, indicated to be misinformation. Others stated how they differentiated between *personal opinions and facts* when listening to a voice message, for instance, via wordings like “I think.” Only one participant stated to consider the *tonality* of a voice message when assessing its credibility, evaluating if it sounded angry or very emotional in another way. Overall, the majority of participants were very capable of correctly assessing the voice messages credibility for voice messages 1 and 2 where the content was clearly misinformation. Many were unsure when assessing voice messages 3 and 4, which were containing both genuine and wrong aspects of information. We found that three participants corrected their credibility assessment for these messages for the better after being confronted with our indicators, as these nudged a more thorough critical reflection of the content. This effect might have appeared as well, when autonomously reflecting on the content without the indicators for a similar amount of time. Thus, we cannot derive reliable implications regarding a learning effect at this point of the study.

Perceived comprehensibility and usefulness of features

Our participants rated the comprehensibility and usefulness on a scale from 1 (not comprehensible/useful at all), over 2 (rather not comprehensible/useful), and 3 (rather comprehensible/useful) to 4 (very comprehensible/useful) and gave additional qualitative explanations and reasons.

Word level: We included several features at the word level that can be automatically detected using our prototype. While all features at the word level were generally perceived as comprehensible (*mean* > 3.0 for all features) and participants were mostly able to correctly describe what they meant, the perceived usefulness varied among more abstract and more contentrelated features. On the abstract part, the *number of words* was perceived as rather not useful (*mean* = 2.0) for several reasons. Most prominently, participants criticized that the number of words or corresponding length of a voice message has nothing to do with the content of the message and can be considered a personal trait. While some tried to make sense of it (“I guess [...] then it's probably to indicate that the lady doesn't have too much information.” [P05, over 56 years, male]), the majority was confident that this feature was not very suitable at first sight.

When taking a closer look at word level features that tend to be less abstract, the perceived usefulness is overall higher.

Some features at word level can be directly related to the content which increased the perceived usefulness. This applies to *verbs that refer to a belief (e.g., I think)* (*mean* = 2.9; “Everyone can believe anything.” [P03, 26–30 years, female]) and *words that express uncertainty or certainty* (*mean* = 2.5):

> Because this claim to be in sole possession of the truth is already an important characteristic of conspiracy theorists. In that respect, it's also a good unique indicator almost. So if you don't notice all the other things, for

> me that would definitely be something where I would say: No, I don't think so. (P09, 36–45 years, diverse)

On an even more concrete level, *suspicious words referring to conspiracy theories* were very well received (*mean* = 3.8) in terms of usefulness to assess a voice message's credibility. This is not surprising, as many are very familiar with common conspiracy theories due to (social) media and state it that voice message 2 (see Table A2, Appendix A) is like playing Bingo due to the many keywords that occur referring to suspicious content. Similarly, many perceived a *claim to be an expert* (*mean* = 3.1) and *calls to action* (*mean* = 3.0) as rather useful: "Because that is also related to the claim of truth and that, well, is always a sure sign for me." (P09, 36–45 years, diverse), however always with a need to differentiate between rightful expert claims and calls to action.

Sentence level: Again, for features on a sentence level, the usefulness was dependent on how abstract or concrete a feature was received. A high *number of adjectives and adverbs* is perceived as a way of figuratively embellishing a situation to give the impression that the author experienced it personally. This can be used to mislead intentionally, however, it was perceived as rather not useful (*mean* = 2.4) due to its abstract nature. The participants struggled with the feature *number of first-person pronoun usage* and spent some effort and time to make sense of it: "Somehow he may have experienced something that practically just happened, but […] may not objectively describe some situation" (P01, 30–35 years, male). However, the feature was overall perceives as rather not useful as well (*mean* = 2.2). Similarly, the *length of sentences* (*mean* = 2.2) and on a broader level, the *vocabulary richness* was perceived as rather not useful (*mean* = 2.1). However, that was very much related to the expression of features. The vocabulary richness of all four voice messages was moderate, which prevents a meaningful comparison within the study. Some expressed it might be useful when there were more extreme expressions of vocabulary richness:

> I think the vocabulary used, if it is particularly rich, could perhaps indicate that the person is very well educated, and thus might also provide an indication of truthfulness, but with an average vocabulary, I wouldn't say that's indicative of truthfulness. (P10, 19–25 years, female)

Similarly, when assessing the usefulness of *number of words with a positive or negative sentiment*, it depended if the expression of the feature was clearly positive or negative, or if it tended to be balanced (*mean* = 2.6). Some participants were very aware of the potential to intentionally use emotional language to mislead and influence: "Because, sure, using fear you can convince people." (P03, 26–30 years, female).

Audio level: Both *speech rate* (*mean* = 2.2) and *sound volume* (*mean* = 1.9) were perceived as rather not useful when assessing voice messages. Participants stated that this is a very individual factor, where some tend to speak fast or low and others do the opposite. Nevertheless, some emphasized the misleading potential of intentionally controlling how fast or loud you speak: "because I think that the, how shall I say, that with the slow speaking speed is also supposed to get across in the message that there is a deep emotional touch somehow." (P05, over 56 years, male; referring to voice message 1).

Creator level: As expected due to existing research on other modalities or social media platforms, the feature of a *suspicious profile (e.g., name, description of creator)* was perceived as useful (*mean* = 3.7). Indeed, several participants autonomously referred to the suspicious name of the creator when listening to the voice message (and seeing the name of the creator in voice message 2): “Because of the many emojis in the name and those red punctuation marks and so on. That is almost common knowledge.” (P09, 36–45 years, diverse).

Overall assessment of the indicator-based approach

Our findings give insights into the overall applicability of showing features of voice messages as indicators for misinformation. Our participants emphasized that it is important to consider the features as a whole rather than as stand-alone components. They noted that these features only make sense when perceived together, to obtain a full picture: “It is comprehensible, but I think it is useful only in combination with all the other indicators” (P09, 36–45 years, diverse) In contrast, this participant also highlighted specific features as a unique indicator (see *words that express uncertainty or certainty*). Furthermore, upon closer examination of all levels, perceived usefulness is closely related to its connection to the content or creator. Participants found features more useful when they could easily recognize its relation to the content. Conversely, more abstract features received less positive feedback.

In contrast to digital misinformation interventions for text, image or even video, voice messages come with the unique attribute of not having any visual indicators. This leads to a particular challenge for interventions which also was highlighted in our study: Receivers of voice messages listen to the message (often only once), and afterwards there are no visual reference points to remember when confronted with a list of indicators. Our study highlights that after listening to the message and gaining feedback (e.g., long sentences), people often do not remember and cannot verify the indicators without the effort to listen to the message again: “I thought I had understood it very well, but now I realize how quickly you can forget what you have just listened to.” (P03, 26–30 years, female) This is very different from other modalities, where intervention designers can easily place the indicator display directly at the place of occurrence.

DISCUSSION

In this study we derived a structured overview on features of voice messages as potential indicators for misinformation in voice messages to develop a prototype of a digital intervention and evaluate the perceived comprehensibility and usefulness of computed features for layperson when listening to voice messages. Thereby, we extend existing research on indicator-based misinformation interventions for text (Martino et al., 2020) and research in misinformation in voice messages (El-Masri et al., 2022; Maros et al., 2021). This allows for in-depth findings on the suitability of features on the character, word, sentence, audio, and creator levels as potential user-centered indicators.

RQ1: How can an automated tool be designed to extract features of misinformation in voice messages, supporting public policy efforts for technical solutions that enhance media literacy?

Governments, authorities, and individuals are confronted with an overabundance of both true information and misinformation on social media, during crises and in everyday life. This includes not only tweets on Twitter/X, TikTok videos, or Facebook posts, but also content on messaging apps like Telegram and WhatsApp with public channels, where the distribution of misleading voice messages has shown to be a common problem in recent years (Resende et al., 2019). Research has highlighted the serious threat of misinformation and disinformation governments are confronted with, emphasized needs for regulation of online content (Jackson, 2021; Schulte & Pickard, 2020), and the necessity to find ways to empower citizens to distinguish misleading from credible information (Connolly et al., 2019) (see Section 2.1). Indeed, public policy research aims to define implications for policymakers when investigating the effects of misinformation and its mitigation, including suggestions to maintain fact-checking and flagging efforts (Diaz Ruiz & Nilsson, 2023) and applying noncensoring techniques like labeling and limiting algorithmic promotion (Di Domenico et al., 2022). When looking into demands for policy responses to misinformation on social media, there are multiple ways to contribute to these necessities via user-centered technology. As Medeiros and Singh (2020), Tambini (2017) among others emphasize, encouraging media literacy can be considered central to counteract the effects of misinformation online. We pick up on that idea by proposing a technological media literacy approach for the user-centered presentation of misinformation indicators in voice messages. Research to combat misinformation on social media while encouraging learning effects has mostly focused on text-based content, especially on Twitter/X, Facebook, and more generic news platforms. Our study addresses the context of voice messages with its unique potentials to mislead (e.g., via emotional tonality) (El-Masri et al., 2022) and may offer support to public policy efforts targeting technological solutions for social media platforms and the empowerment of citizens (Connolly et al., 2019).

With the development of a backend solution for a prototype that automates the detection and extraction of features of voice messages as potential indicators for misinformation, this work contributes to informed information handling, particularly in messenger channels like Telegram (El-Masri et al., 2022; Herasimenka et al., 2023). The prototype receives voice messages or other audio files, transcribes them, and analyzes them for character-, word-, and sentence-level features. Additionally, the prototype facilitates an extension for audio and creator-level features for future work. By doing so, we add to the public policy response to the misinformation challenge, presenting a potential technological solution to enhance users' media literacy and encourage informed decisions on voice messages' credibility. We envision usage scenarios at both the individual and government levels: (1) Private users of messenger services can have their voice messages (e.g., in large public channels) analyzed and receive a comprehensive output on attributes of the message. This allows them to reflect on the content and make a more informed assessment independently. (2) The intervention encompasses a potential part of a policy toolkit that places a greater emphasis on self-efficacy in dealing with misinformation as opposed to highly regulatory approaches by governments that amount to censorship. Additionally, on a microlevel perspective, people working in crisis management or government agencies may utilize the tool to automatically analyze large amounts of data and quickly identify problematic content, such as conspiracy theories, by specifying specific keywords. This support in monitoring and analysis can supplement manual procedures and extend existing dashboard approaches to classic social media platforms like Facebook and X (formerly Twitter). (3) Researchers in detection methods can selectively extract features from large amounts of data to expand and optimize their detection techniques.

The study examined the extracted indicators from the user's perspective with respect to Scenario 1. As a next step, we recommend conducting more comprehensive investigations from the perspective of government agencies in future work. It is essential to collect user interface requirements to expand our backend component, as our study identified unique challenges (e.g., on the placement of indicators) specific to voice messaging in comparison to other modalities, including text and images, where indicator-based interventions have already been addressed (Ayoub et al., 2021; Furuta & Suzuki, 2021; Martino et al., 2020).

RQ2: How do users evaluate the comprehensibility and usefulness of the automatically computed indicators when assessing the credibility of voice messages?

When assessing the comprehensibility and perceived usefulness of features to assess a voice message's credibility, we received a mostly positive feedback especially on features that refer to the content itself. Indeed, focusing mostly on the content instead of tonality was the main autonomous assessment strategy even before being confronted with our features, sometimes resulting in a certain gut feeling. However, this was very much dependent on the expression of this characteristic. If it was strongly pronounced and deviated more from the average, it was rated all the more useful.

All features received a comprehensibility rating of *mean* > 3.0 which indicates that the wordings of our features as potential indicators were perceived as rather suitable. Indeed, when taking a closer look at our participants' verbal answers and attempts to explain what each feature means, they were overall very capable of doing so. We found that features closely related to the content of the voice message (e.g., suspicious conspiracy theory keywords, emotionally loaded words, and words referring to the speaker's belief like "I think that …") were particularly rated as comprehensible. Also, the feature related to a suspicious profile (e.g. name, description of the creator) was considered particularly comprehensible among the participants. Regarding features on a more abstract level (e.g., number of words, length of sentences, number of first-person pronoun usage), some participants struggled with comprehending what the feature means.

We can observe a similar tendency regarding perceived usefulness. Features that relate directly to the content of the voice message or the creator were rated particularly useful. The suspicious profile (e.g., name, description of creator) and suspicious keywords about conspiracy theories are rated as the most useful feature. Furthermore, other keyword-based features like the claim to be an expert or the call to action (El-Masri et al., 2022) were very positively received. On the other hand, the abstract features like number of words, sentence length, number of first-person pronoun usage, and the percentage of adjectives and adverbs were perceive as rather not useful when assessing a voice message's credibility. While other research on voice messages (El-Masri et al., 2022) and short-videos (Hassoun et al., 2023) have emphasized the importance of affectively charged content as misinformation attribute, the features used in our study were too abstract to apply that message in a user-centered way. Similarly, this applies for ascertaining credibility of misinformation in voice messages as "eyewitness, expert, or insider" (El-Masri et al., 2022, p. 5) which was strongly the case in voice message 1 (see Table A2, Appendix A) where we expected the number of first-person pronouns to be a comprehensible indicator for that context—an assumption that was fulfilled only for a few participants. Regarding the audio-based features like speech rate and volume, participants again did not perceive them as useful features, extending findings of related research (El-Masri et al., 2022).

Overall, our findings emphasize the importance to combine multiple features from different levels to receive a suitable and useful full picture. However, the indicator-based approach comes with a challenge when applied to voice messages in contrast to other content like text (Martino et al., 2020), image, or video (Hassoun et al., 2023), as there are no visual anchors users can relate to and verify when indicators are displayed. This is a challenge that may be further addressed in future HCI research, having a direct influence on how indicators can be presented to end-users.

Limitations and future work

Our contribution comes with some limitations and potentials for future work.

First, the proposed digital intervention is yet to be extended by a frontend part that displays the extracted features as potential indicators for misinformation in a user-centered manner. Regarding voice messages, this is not a trivial task, as the modality comes with limited possibilities for visual anchoring of indicators where they occur in the message. This is unique to audio material, as in text, images, and even video content there are more intuitive ways of highlighting indicators visually immediately at their place of occurrence. Further, user needs for displaying the extracted features may significantly vary between the different usage scenarios of private messenger users, practitioners in crisis management, and researchers dealing with detection approaches themselves—a challenge to be addressed in future work.

Second, our indicator-based approach relies on preceding manual or automatic detection of misinformation. User-centered misinformation interventions typically rely on prior successful determination of whether a given piece of information is misleading. This is by no means a trivial task and has been a focal subject of research in recent years (Shu et al., 2017a; Wu & Liu, 2018), particularly within the field of machine learning. Default display of indicators on all voice messages would result in a not acceptable number of false positives and, thus, is not the aim of this approach. Future work could evaluate the fully implemented approach for voice messages including (1) an automatic or manual detection of misinformation and (2) extracting and displaying indicators for content that has been successfully identified as misleading.

Third, evaluating our approach with a larger real-world sample of voice messages is crucial to make valid assumption about its applicability. While for other content like Tweets or news articles there are publicly available research data sets in multiple languages, this is to the best of our knowledge not the case for voice messages. We therefore applied a systematic and convenient approach of looking for officially debunked misinformation and actively identifying corresponding voice messages in Telegram groups. Due to an overabundance of misinformation in one specific political direction, we chose extend our real-world sample with two artificially generated voice messages addressing misinformation from an opposite political perspective. This might have influenced the results of our user study, however we did not find any differences in our results when comparing the participants' answers on real-world and artificial messages in comparison.

Fourth, the appropriateness of indicator-based approaches as media literacy interventions is debated, as it depends on rationality (Boyd, 2017) and might create false confidence (Bulger & Davison, 2018). Further, media literacy has been criticized for only being one of many components within the complex information space (Bulger & Davison, 2018; Hassoun et al., 2023). Therefore, there are calls to "rethinking media literacy in the age of platforms" (Bulger & Davison, 2018, p. 17). However, studies also show positive outcomes of media literacy interventions on critical thinking and behavior change regarding misinformation behavior (Spezzano, 2021; Winiecki et al., 2023)—an optimistic view we draw on in this study. Our approach is only one of many possible solutions and (like other types of interventions that have already been researched) can only be one of many contributions. For example, the approach has the disadvantage that the features of misinformation in voice messages are characteristics that change over time and with the (technical) progress of disinformation campaigns and require adaptation. With the ever-changing information landscape, the skills and strategies of disinformation creators also change. A combination of different components, areas of responsibility and types of intervention in the interplay of critical journalism, media literacy training and the use of the latest technologies (including AI-based technologies) is therefore essential.

CONCLUSION

We developed a prototype as proof-of-concept of an indicator-based digital misinformation intervention for voice messages and conducted a user-centered evaluation of the perceived comprehensibility and usefulness of computational features for laypersons as users of messengers. Thereby, we gained insights into how indicator-based interventions that have been shown to be promising for text-based content can be applied to voice messages as a modality highly relevant in the context of misinformation. We adopt the user perspective of diverse users of messengers like Telegram and Whatsapp. The approach encompasses a public policy response (Tambini, 2017) providing a foundation for technical solutions to enhance users' media literacy skills, placing greater emphasis on self-efficacy in dealing with misinformation as opposed to regulatory approaches by governments that amount to censorship. While we identified three usage scenarios of the proposed indicator-based intervention ([1] laypersons that want to assess the credibility of voice messages in crises or in everyday life, [2] practitioners in crisis management organizations or authorities that need to analyze large data sets, for example, during emerging crises, and [3] researchers requiring features for a detection model), we exemplarily conducted the user study regarding the first scenario. We generated novel insights into the opportunities and limitations of features on character, word, sentence, audio, and creator level as potential indicators for misinformation in voice messages. Our work is relevant to the public policy domain as it addresses demands for public policy responses to the misinformation challenge promoting technological media literacy solutions and preventing overzealous censoring regulations by governments (Medeiros & Singh, 2020). The challenge lies in controlling the deluge of credible and misleading information. It is crucial to consider the subtleties of voice messaging being a significant yet problematic information channel. We propose an indicator-based misinformation intervention for voice messages to partly address the need. This provides decision makers with a tool to navigate the complex landscape of information dissemination, particularly during crises. Our main contributions are (C1) giving a structured overview on features of voice messages as potential indicators for misinformation, (C2) developing an indicator-based misinformation intervention as proof-of-concept, and (C3) evaluating the perceived comprehensibility and usefulness of a plethora of features to assess the credibility regarding the scenario of layperson listening to voice messages.

ACKNOWLEDGMENTS

This work was supported by the German Federal Ministry for Education and Research (BMBF) in the project NEBULA (13N16361) and in the project CYLENCE (13N16636), and by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE. Open Access funding enabled and organized by Projekt DEAL.

ORCID

Katrin Hartwig http://orcid.org/0000-0003-4875-0110
Christian Reuter http://orcid.org/0000-0003-1920-038X

ENDNOTES

[1] https://www.manchestereveningnews.co.uk/news/greater-manchester-news/warning-over-hoax-ambulance-voice-18067450

[2] for further information regarding the categories, see. https://web.archive.org/web/20171007184158/http://www.wjh.harvard.edu/~inquirer/homecat

REFERENCES

Abbasi, Ahmed, and Hsinchun Chen. 2008. "Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace." *ACM Transactions on Information Systems* 26(2): 7:1–7:29. https://doi.org/10.1145/1344411.1344413

Afroz, Sadia, Michael Brennan, and Rachel Greenstadt. 2012. "Detecting Hoaxes, Frauds, and Deception in Writing Style Online." In *2012 IEEE Symposium on Security and Privacy*, 461–75. San Francisco: IEEE. https://doi.org/10.1109/SP.2012.34

Almaliki, Malik. 2019. "Online Misinformation Spread: A Systematic Literature Map." In *Proceedings of the 2019 3rd International Conference on Information System and Data Mining (ICISDM 2019)*, 171–8. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3325917.3325938

Anderson, Jonathan. 1983. "Lix and Rix: Variations on a Little-known Readability Index." *Journal of Reading* 26(6): 490–496. jstor:40031755.

Aufderheide, Patricia. 1993. A Report of the National Leadership Conference on Media Literacy. Technical Report. Washington, DC: Aspen Institute.

Ayoub, Jackie, X. Jessie Yang, and Feng Zhou 2021. "Combat COVID-19 Infodemic Using Explainable Natural Language Processing Models." *Information Processing & Management* 58(4): 102569. https://doi.org/10.1016/j.ipm.2021.102569

Balkin, Jack M. 2017. "Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation." *SSRN Electronic Journal* 615: 1–68. https://doi.org/10.2139/ssrn.3038939

Bhargava, Puneet, Katie MacDonald, Christie Newton, Hause Lin, and Gordon Pennycook 2023. "How Effective Are TikTok Misinformation Debunking Videos?" *Harvard Kennedy School Misinformation Review* 4(2): 1–17. https://doi.org/10.37016/mr-2020-114

Bhuiyan, Md. Momen, Hayden Whitley, Michael Horning, Sang Won Lee, and Tanushree Mitra. 2021. "Designing Transparency Cues in Online News Platforms to Promote Trust: Journalists' & Consumers' Perspectives." *Proceedings of the ACM on Human–Computer Interaction* 5(CSCW2): 1–31. https://doi.org/10.1145/3479539

Bhuiyan, Md. Momen, Kexin Zhang, Kelsey Vick, Michael A. Horning, and Tanushree Mitra. 2018. "FeedReflect: A Tool for Nudging Users to Assess News Credibility on Twitter." In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '18)*, 205–8. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3272973.3274056

Björnsson, Carl-Hugo. 1971. *Læsbarhed*. Copenhagen: Gad.

Boyd, Danah. 2017. "Did Media Literacy Backfire?" *Journal of Applied Youth Studies* 1(4): 83–9.

Brennan, Michael, and Rachel Greenstadt. 2009. "Practical Attacks Against Authorship Recognition Techniques." In *Proceedings of the Twenty-First Conference on Innovative Applications of Artificial Intelligence*, 60–5. Pasadena, CA: AAAI.

Brovdiy, Yana. 2020. Disinformation in Times of COVID-19: Reinforcing the Responses of the European Union and the United States. Technical Report. Bruges, Belgium: College of Europe.

Brunet, Étienne. 1978. *Le Vocabulaire de Jean Giraudoux: Structure et Évolution: Statistique et Informatique Appliquées à l'étude Des Textes à Partir Des Données Du Trésor de La Langue Française*. Genève: Slatkine.

Bulger, Monica, and Patrick Davison. 2018. "The Promises, Challenges, and Futures of Media Literacy." *Journal of Media Literacy Education* 10(1): 1–21. https://doi.org/10.23860/JMLE-2018-10-1-1

Burgoon, Judee K., J. P. Blair, Tiantian Qin, and Jay F. Nunamaker. 2003. "Detecting Deception through Linguistic Analysis." In *Proceedings of Intelligence and Security Informatics (Lecture Notes in Computer Science)*, edited by Hsinchun Chen, Richard Miranda, Daniel D. Zeng, Chris Demchak, Jenny Schroeder, and Therani Madhusudan, 91–101. Tucson, AZ: Springer. https://doi.org/10.1007/3-540-44853-5_7

Chakraborty, Abhijnan, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. 2016. *Stop Clickbait: Detecting and Preventing Clickbaits in Online News Media*. New York City, USA: IEEE.

Chen, Jiangjie, Qiaoben Bao, Changzhi Sun, Xinbo Zhang, Jiaze Chen, Hao Zhou, Yanghua Xiao, and Lei Li. 2022. "LOREN: Logic-Regularized Reasoning for Interpretable Fact Verification." *Proceedings of the AAAI Conference on Artificial Intelligence* 36(10): 10482–91. https://doi.org/10.1609/aaai.v36i10.21291

Chen, Yimin, Niall J. Conroy, and Victoria L. Rubin. 2015. "Misleading Online Content: Recognizing Clickbait as "False News"." In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection (WMDD '15)*, 15–9. Seattle, WA: Association for Computing Machinery. https://doi.org/10.1145/2823465.2823467

Clark, Nathan Edward, Kees Boersma, Sara Bonati, Chiara Fonio, Simon Gehlhar, Therese Habig, Robert Larruina, et al. 2024. "Exploring the Impacts of Social Media and Crowdsourcing on Disaster Resilience." *Open Research Europe* 1: 1–17. https://doi.org/10.12688/openreseurope.13721.3

Coleman, Meri, and T. L. Liau. 1975. "A Computer Readability Formula Designed for Machine Scoring." *Journal of Applied Psychology* 60: 283–4. https://doi.org/10.1037/h0076540

Connolly, Jennifer M., Joseph E. Uscinski, Casey A. Klofstad, and Jonathan P. West. 2019. "Communicating to the Public in the Era of Conspiracy Theory." *Public Integrity* 21(5): 469–76. https://doi.org/10.1080/10999922.2019.1603045

Davies, William. 2020. *What's Wrong with WhatsApp*. Minnesota, USA: University of Minnesota. https://www.theguardian.com/technology/2020/jul/02/whatsapp-groups-conspiracy-theories-disinformation-democracy

Diaz Ruiz, Carlos, and Tomas Nilsson. 2023. "Disinformation and Echo Chambers: How Disinformation Circulates on Social Media Through Identity-Driven Controversies." *Journal of Public Policy & Marketing* 42: 18–35. https://doi.org/10.1177/07439156221103852

Di Domenico, Giandomenico, Daniel Nunan, and Valentina Pitardi. 2022. "Marketplaces of Misinformation: A Study of How Vaccine Misinformation Is Legitimized on Social Media." *Journal of Public Policy & Marketing* 41(4): 319–35. https://doi.org/10.1177/07439156221103860

El-Masri, Azza, Martin J. Riedl, and Samuel Woolley. 2022. "Audio Misinformation on WhatsApp: A Case Study from Lebanon." *Harvard Kennedy School Misinformation Review* 3(4): 1–13. https://doi.org/10.37016/mr-2020-102

Feng, Song, Ritwik Banerjee, and Yejin Choi. 2012. "Syntactic Stylometry for Deception Detection." In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 171–5. Jeju Island, Korea: Association for Computational Linguistics.

Fonteyn, Marsha E., Benjamin Kuipers, and Susan J. Grobe. 1993. "A Description of Think Aloud Method and Protocol Analysis." *Qualitative Health Research* 3(4): 430–41. https://doi.org/10.1177/104973239300300403

Furuta, Tomoya, and Yu Suzuki. 2021. "A Fact-checking Assistant System for Textual Documents." In *2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, 243–6. Tokyo, Japan: IEEE. https://doi.org/10.1109/MIPR51284.2021.00046

Grandhi, Sukeshini, Linda Plotnick, and Starr Roxanne Hiltz. 2021. "By the Crowd and for the Crowd: Perceived Utility and Willingness to Contribute to Trustworthiness Indicators on Social Media." *Proceedings of the ACM on Human–Computer Interaction* 5(GROUP): 1–24. https://doi.org/10.1145/3463930

Gui, Xinning, Yubo Kou, Kathleen H. Pine, and Yunan Chen. 2017. "Managing Uncertainty: Using Social Media for Risk Assessment during a Public Health Crisis." In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 4520–33. Denver, CO: ACM. https://doi.org/10.1145/3025453.3025891

Gunning, Robert. 1952. *The Technique of Clear Writing*. New York, NY: McGraw-Hill.

Gunz, Hendrik, and Isa Schaller. 2022. "Antisemitic Narratives on YouTube and Telegram as Part of Conspiracy Beliefs about COVID-19." In *Antisemitism on Social Media* (1st ed.). 129–50. London: Routledge. https://doi.org/10.4324/9781003200499-9

Hagar, Christine. 2006. "Using Research to Aid the Design of a Crisis Information Management Course." In *Association of Library & Information Science Educators SIG*, 6–10. San Antonio, TX: ALISE.

Hancock, Jeffrey, Lauren Curry, Saurabh Goorha, and Michael Woodworth. 2008. "On Lying and Being Lied To: A Linguistic Analysis of Deception in Computer-Mediated Communication." *Discourse Processes* 45: 1–23. https://doi.org/10.1080/01638530701739181

Hartwig, Katrin, Frederic Doell, and Christian Reuter. 2023. *The Landscape of User-centered Misinformation Interventions—A Systematic Literature Review*. arXiv:2301.06517 [cs]. New York, USA: Cornell University.

Hartwig, Katrin, Tom Biselli, Franziska Schneider, and Christian Reuter. 2024. "From Adolescents' Eyes: Assessing an Indicator-Based Intervention to Combat Misinformation on TikTok." In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. Hawaii: ACM. https://doi.org/10.1145/3613904.3642264

Hassoun, Amelia, Ian Beacock, Sunny Consolvo, Beth Goldberg, Patrick Gage Kelley, and Daniel M. Russell. 2023. "Practicing Information Sensibility: How Gen Z Engages with Online Information." In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, 1–17. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3544548.3581328

Herasimenka, Aliaksandr, Jonathan Bright, Aleksi Knuutila, and Philip N. Howard. 2023. "Misinformation and Professional News on Largely Unmoderated Platforms: The Case of Telegram." *Journal of Information Technology & Politics* 20: 198–212. https://doi.org/10.1080/19331681.2022.2076272

Hildreth, W. Bartley, Gerald Miller, Jack Rabin, and Gerald J. Miller. 2007. *Handbook of Public Administration*. Oxfordshire, United Kingdom: Routledge.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd 2020. *spaCy: Industrial-strength Natural Language Processing in Python*. New York, USA: Cornell University. https://doi.org/10.5281/zenodo.1212303

Honoré, Antony. 1979. "Some Simple Measures of Richness of Vocabulary." *Association for Literary and Linguistic Computing Bulletin* 7(2): 172–7.

Imran, Muhammad, Patrick Meier, and Kees Boersma. 2017. "The Use of Social Media for Crisis Management: A Privacy by Design Approach." In *Big Data, Surveillance and Crisis Management*, 19–37. London, United Kingdom: Routledge.

Jackson, Nicole J. 2021. "The Canadian Government's Response to Foreign Disinformation: Rhetoric, Stated Policy Intentions, and Practices." *International Journal* 76(4): 544–63. https://doi.org/10.1177/00207020221076402

Kalsnes, Bente. 2018. "Fake News." In *Oxford Research Encyclopedia of Communication*. Oxford, United Kingdom: Oxford University Press.

Kaufhold, Marc-André, Alexis Gizikis, Christian Reuter, Matthias Habdank, and Margarita Grinko. 2019. "Avoiding Chaotic Use of Social Media Before, During, and After Emergencies: Design and Evaluation of Citizens' Guidelines." *Journal of Contingencies and Crisis Management* 27(3): 198–213. https://doi.org/10.1111/1468-5973.12249

Kincaid, J. Peter, Robert P. Fishburne Jr., Richard L. Rogers, and Brad S. Chissom 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical Report 8-75. Florida: Institute for Simulation and Trainings.

Kirchner, Jan, and Christian Reuter. 2020. "Countering Fake News: A Comparison of Possible Solutions Regarding User Acceptance and Effectiveness." *Proceedings of the ACM on Human–Computer Interaction* 4(CSCW2): 1–27. https://doi.org/10.1145/3415211

Komendantova, Nadejda, Dmitry Erokhin, and Teresa Albano. 2023. "Misinformation and Its Impact on Contested Policy Issues: The Example of Migration Discourses." *Societies* 13(7): 168. https://doi.org/10.3390/soc13070168

Levendusky, Matthew S. 2013. "Why Do Partisan Media Polarize Viewers?" *American Journal of Political Science* 57(3): 611–23. https://doi.org/10.1111/ajps.12008

Levitan, Sarah, Guozhen An, Min Ma, Rivka Levitan, Andrew Rosenberg, and Julia Hirschberg. 2016. "Combining Acoustic-Prosodic, Lexical, and Phonotactic Features for Automatic Deception Detection." In *Proceedings of the Interspeech*, 2006–10. San Francisco: ISCA. https://doi.org/10.21437/Interspeech.2016-1519

Li, Yang-Jun, Jens Marga, Christy Cheung, Xiao-Liang Shen, and Matthew Lee. 2022. "Health Misinformation on Social Media: A Systematic Literature Review and Future Research Directions." *AIS Transactions on Human–Computer Interaction* 14(2): 116–49. https://doi.org/10.17705/1thci.00164

Maros, Alexandre, Jussara Almeida, Fabrício Benevenuto, and Marisa Vasconcelos. 2020. "Analyzing the Use of Audio Messages in WhatsApp Groups." In *Proceedings of the Web Conference 2020*, 3005–11. Taipei, Taiwan: ACM. https://doi.org/10.1145/3366423.3380070

Maros, Alexandre, Anastasia Giachanou, Viktoria Spaiser, Francesca Spezzano, Anna George, Alexandra Pavliuc, Alexandre Bright, Jonathan, Jussara M. Almeida, and Marisa Vasconcelos. 2021. "A Study of Misinformation in Audio Messages Shared in WhatsApp Groups." In *Disinformation in Open Online Media*, vol. 12887, 85–100. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-87031-7_6

Martino, Giovanni Da San, Shaden Shaar, Yifan Zhang, Seunghak Yu, Alberto Barrón-Cedeño, and Preslav Nakov. 2020. "Prta: A System to Support the Analysis of Propaganda Techniques in the News." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 287–93. Association for Computational Linguistics, Online. arXiv:2005.05854

McAlpine, Rachel. 1997. *Global English: For Global Business*. North Shore, NZ: Pearson Education New Zealand Limited.

Mclaughlin, G. Harry. 1969. "SMOG Grading—A New Readability Formula." *The Journal of Reading* 12(8): 639–46.

Medeiros, Ben, and Pawan Singh. 2020. "Addressing Misinformation on WhatsApp in India Through Intermediary Liability Policy, Platform Design Modification, and Media Literacy." *Journal of Information Policy* 10: 276–98.

Metzger, Miriam J., and Andrew J. Flanagin. 2013. "Credibility and Trust of Information in Online Environments: The Use of Cognitive Heuristics." *Journal of Pragmatics* 59: 210–20. https://doi.org/10.1016/j.pragma.2013.07.012

Müller, Philipp, and Nora Denner. 2019. Was tun gegen Fake News? Eine Analyse anhand der Entstehungsbedingungen und Wirkweisen gezielter Falschmeldungen im Internet: Kurzgutachten im Auftrag der Friedrich-Naumann-Stiftung für die Freiheit. Technical Report. Friedrich-Naumann-Stiftung für die Freiheit, 1–32.

Ng, Lynnette Hui Xian, and Jia Yuan Loke. 2021. "Analyzing Public Opinion and Misinformation in a COVID-19 Telegram Group Chat." *IEEE Internet Computing* 25(2): 84–91. https://doi.org/10.1109/MIC.2020.3040516

Nyhan, Brendan, and Jason Reifler. 2010. "When Corrections Fail: The Persistence of Political Misperceptions." *Political Behavior* 32(2): 303–30. https://doi.org/10.1007/s11109-010-9112-2

Osmundsen, Mathias, Alexander Bor, Peter Bjerregaard Vahlstrup, Anja Bechmann, and Michael Bang Petersen. 2021. "Partisan Polarization Is the Primary Psychological Motivation Behind Political Fake News Sharing on Twitter." *American Political Science Review* 115(3): 999–1015. https://doi.org/10.1017/S0003055421000290

Palen, Leysia, Sarah Vieweg, Sophia B. Liu, and Amanda Lee Hughes. 2009. "Crisis in a Networked World: Features of Computer-Mediated Communication in the April 16, 2007, Virginia Tech Event." *Social Science Computer Review* 27(4): 467–80. https://doi.org/10.1177/0894439309332302

Papadopoulou, Olga, Themistoklis Makedas, Lazaros Apostolidis, Francesco Poldi, Symeon Papadopoulos, and Ioannis Kompatsiaris. 2022. "MeVer NetworkX: Network Analysis and Visualization for Tracing Disinformation." *Future Internet* 14(5): 147. https://doi.org/10.3390/fi14050147

Potthast, Martin, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. *A Stylometric Inquiry into Hyperpartisan and Fake News*. arXiv:1702.05638 [cs]. New York, USA: Cornell University. https://doi.org/10.48550/arXiv.1702.05638.

Potthast, Martin, Sebastian Köpsel, Benno Stein, and Matthias Hagen. 2016. "Clickbait Detection." In *Proceedings of Advances in Information Retrieval (Lecture Notes in Computer Science)*, edited by Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello, 810–7. Padua, Italy: Springer International Publishing. https://doi.org/10.1007/978-3-319-30671-1_72

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. *Robust Speech Recognition via Large-Scale Weak Supervision*. arXiv:2212.04356 [cs, eess]. New York, USA: Cornell University. https://doi.org/10.48550/arXiv.2212.04356.

Reddick, Christopher G. 2012. *Public Administration and Information Technology*. Burlington: Jones & Bartlett Learning.

Resende, Gustavo, Philipe Melo, Hugo Sousa, Johnnatan Messias, Marisa Vasconcelos, Jussara Almeida, and Fabrício Benevenuto. 2019. "(Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures." In *The World Wide Web Conference*, 818–28. San Francisco, CA: ACM. https://doi.org/10.1145/3308558.3313688

Reuter, Christian. 2022. *A European Perspective on Crisis Informatics: Citizens' and Authorities' Attitudes Towards Social Media for Public Safety and Security*. Nijmegen, The Netherlands: The Radboud University Thesis Repository.

Reuter, Christian, and Marc-André Kaufhold. 2018. "Fifteen Years of Social Media in Emergencies: A Retrospective Review and Future Directions for Crisis Informatics." *Journal of Contingencies and Crisis Management* 26(1): 41–57. https://doi.org/10.1111/1468-5973.12196

Rubin, Victoria L., and Tatiana Lukoianova. 2015. "Truth and Deception at the Rhetorical Structure Level." *Journal of the Association for Information Science and Technology* 66(5): 905–17. https://doi.org/10.1002/asi.23216

Schmid, Stefka, Katrin Hartwig, Robert Cieslinski, and Christian Reuter. 2022. "Digital Resilience in Dealing with Misinformation on Social Media During COVID-19 A Web Application to Assist Users in Crises." *Information Systems Frontiers* 2022: 1–23. https://doi.org/10.1007/s10796-022-10347-5

Schulte, Stephanie Ricker, and Victor Pickard. 2020. "Fixing Fake News: Self-Regulation and Technological Solutionism." In *Fake News: Understanding Media and Misinformation in the Digital Age*. Cambridge, MA: MIT Press.

Shang, Lanyu, Ziyi Kou, Yang Zhang, and Dong Wang. 2021. "A Multimodal Misinformation Detector for COVID-19 Short Videos on TikTok." In *2021 IEEE International Conference on Big Data (Big Data)*, 899–908. IEEE. https://doi.org/10.1109/BigData52589.2021.9671928

Sherman, Imani N., Jack W. Stokes, and Elissa M. Redmiles. 2021. "Designing Media Provenance Indicators to Combat Fake Media." In *24th International Symposium on Research in Attacks, Intrusions and Defenses*, 324–39. San Sebastian, Spain: ACM. https://doi.org/10.1145/3471621.3471860

Shu, Kai, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017a. "Fake News Detection on Social Media." *ACM SIGKDD Explorations Newsletter* 19(1): 22–36.

Shu, Kai, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017b. *Fake News Detection on Social Media: A Data Mining Perspective*. New York, USA: Cornell University. arXiv:1708.01967 [cs]. https://doi.org/10.48550/arXiv.1708.01967

Sichel, H. S. 1975. "On a Distribution Law for Word Frequencies." *Journal of the American Statistical Association* 70(351): 542–547. jstor:2285930. https://doi.org/10.2307/2285930

Simpson, E. Hugh. 1949. "Measurement of Diversity." *Nature* 163(4148): 688. https://doi.org/10.1038/163688a0

Smith, Edgar A., and J. Peter Kincaid. 1970. "Derivation and Validation of the Automated Readability Index for Use with Technical Materials." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 12(5): 457–564. https://doi.org/10.1177/001872087001200505

Spezzano, Francesca. 2021. "Using Service-Learning in Graduate Curriculum to Address Teenagers' Vulnerability to Web Misinformation." In *Proceedings of the 26th ACM Conference on Innovation and Technology in Computer Science Education V. 2*, 637–8. Virtual Event, Germany: ACM. https://doi.org/10.1145/3456565.3460039

Statista. 2023. *Ranking of the Biggest Social Networks and Messengers Regarding Number of Users in January 2023* (in German). https://de.statista.com/statistik/daten/studie/181086/umfrage/die-weltweit-groessten-social-networks-nach-anzahl-der-user/

Tambini, Damian. 2017. Fake News: Public Policy Responses. Technical Report. London, United Kingdom: Media Policy Project, London School of Economics and Political Science.

Tweedie, Fiona J., and R. Harald Baayen. 1998. "How Variable May a Constant Be? Measures of Lexical Richness in Perspective." *Computers and the Humanities* 32(5): 323–352. jstor:30200474

Wallbridge, Sarenne, Peter Bell, and Catherine Lai. 2021. *It's Not What You Said, It's How You Said It: Discriminative Perception of Speech as a Multichannel Communication System*. arXiv:2105.00260 [cs]. New York, USA: Cornell University. https://doi.org/10.48550/arXiv.2105.00260.

Wang, Yuxi, Martin McKee, Aleksandra Torbica, and David Stuckler. 2019. "Systematic Literature Review on the Spread of Health-Related Misinformation on Social Media." *Social Science & Medicine* 240: 112552. https://doi.org/10.1016/j.socscimed.2019.112552

WhatsApp. 2013. *Introducing Voice Messages*. https://blog.whatsapp.com/introducing-voice-messages?lang=en

WhatsApp. 2022. *We're Making Voice Messages Even Better*. https://blog.whatsapp.com/making-voice-messages-better

Winiecki, Don, Francesca Spezzano, and Chandler Underwood. 2023. "Understanding Teenagers' Real and Fake News Sharing on Social Media." In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*, 598–602. Chicago, IL: ACM. https://doi.org/10.1145/3585088.3593864

Wu, Liang, and Huan Liu. 2018. "Tracing Fake-News Footprints: Characterizing Social Media Messages by How They Propagate." In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 637–45. Marina Del Rey, CA: ACM. https://doi.org/10.1145/3159652.3159677

Yule, G. Udny. 1944. *The Statistical Study of Literary Vocabulary*. Cambridge, UK: Cambridge University Press.

Zheng, Rong, Jiexun Li, Hsinchun Chen, and Zan Huang. 2006. "A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques." *Journal of the American Society for Information Science and Technology* 57(3): 378–93. https://doi.org/10.1002/asi.20316

How to cite this article: Hartwig, Katrin, Ruslan Sandler, and Christian Reuter. 2024. "Navigating misinformation in voice messages: Identification of user-centered features for digital interventions." *Risk, Hazards, & Crisis in Public Policy*. 1–33. https://doi.org/10.1002/rhc3.12296

APPENDIX A: SUMMARY OF FEATURES AS POTENTIAL INDICATORS FOR MISINFORMATION IN VOICE MESSAGES

See Tables A1 and A2.

TABLE A1

Feature	Related task	Concepts	References	User study
Textual features at the character level				
Number of characters[a]	Authorship recognition, deception detection[b]		Abbasi and Chen (2008), Afroz et al. (2012), Potthast et al. (2016), and Zheng et al. (2006)	–
Percentage of digits	Authorship recognition, deception detection		Afroz et al. (2012)	–
Frequency of character *n*-grams[c]	Authorship recognition, clickbait detection, deception detection		Abbasi and Chen (2008), Afroz et al. (2012), Potthast et al. (2017, 2016), and Zheng et al. (2006)	–
Whether the text starts with a number	Clickbait detection		Potthast et al. (2016)	–
Textual features at the word level				
Number of words	Authorship recognition, clickbait detection, deception detection		Abbasi and Chen (2008), Afroz et al. (2012), Burgoon et al. (2003), Chakraborty et al. (2016), Maros et al. (2021), and Zheng et al. (2006)	"The number of words is X"
Number of unique words	Authorship recognition, deception detection	Vocabulary richness	Afroz et al. (2012) and Zheng et al. (2006)	–
Number of hapax (dis) legomena[d]	Authorship recognition		Abbasi and Chen (2008) and Zheng et al. (2006)	–
Number of tentative words	Authorship recognition, deception detection	Uncertainty	Afroz et al. (2012)	"Contains X words expressing certainty and Y words expressing uncertainty"

(Continues)

TABLE A1 (Continued)

Feature	Related task	Concepts	References	User study
Number of words that express certainty	Authorship recognition, deception detection	Uncertainty	Afroz et al. (2012)	"Contains X words expressing certainty and Y words expressing uncertainty"
Number of syllables	Authorship recognition, deception detection		Afroz et al. (2012) and Burgoon et al. (2003)	–
Average number of characters per word	Authorship recognition, clickbait detection, deception detection		Afroz et al. (2012), Chakraborty et al. (2016), Potthast et al. (2016), and Zheng et al. (2006)	–
Average number of syllables per word	Authorship recognition, deception detection	Vocabulary complexity	Afroz et al. (2012) and Burgoon et al. (2003)	–
Frequency of *n*-grams[e]	Authorship recognition, clickbait detection, deception detection		Abbasi and Chen (2008), Afroz et al. (2012), Chakraborty et al. (2016), Feng et al. (2012), and Potthast et al. (2016)	–
Frequency of 1–20 letter words[f]	Authorship recognition, deception detection	Vocabulary complexity	Abbasi and Chen (2008), Burgoon et al. (2003), and Zheng et al. (2006)	–
Stop words-to-words ratio	Clickbait detection		Potthast et al. (2016)	–
Keywords for a specific topic[g]	Authorship recognition, deception detection		Afroz et al. (2012) and Zheng et al. (2006)	"Contains X words related to a belief (e.g., I think)"/"Contains common buzzwords related to conspiracy theories"/"Contains claim to be an expert"/"Contains call to action"
Modal verbs	Authorship recognition, deception detection	Uncertainty	Afroz et al. (2012)	–
Textual features at the sentence level				
Vocabulary richness measures[h,i]	Authorship recognition	Vocabulary richness	Zheng et al. (2006)	"The vocabulary richness is moderate"
Number of affective terms	Authorship recognition, deception detection	Specificity and expressiveness	Afroz et al. (2012) and Burgoon et al. (2003)	"There are X words with a negative sentiment and Y words with a positive sentiment"

TABLE A1 (Continued)

Feature	Related task	Concepts	References	User study
Number of (short/long) sentences	Authorship recognition, deception detection	Grammatical complexity	Afroz et al. (2012), Burgoon et al. (2003), and Zheng et al. (2006)	"The sentences are on average rather long with X words per sentence"
Number of first, second, third person pronoun usage	Authorship recognition, clickbait detection, deception detection	Verbal nonimmediacy	Afroz et al. (2012), Chakraborty et al. (2016), and Maros et al. (2021)	"There are X first person pronouns (e.g., I, my, me)."
Number of conjunctions	Authorship recognition, deception detection	Grammatical complexity	Afroz et al. (2012) and Burgoon et al. (2003)	–
Average number of words per sentence	Authorship recognition, deception detection	Grammatical complexity	Afroz et al. (2012), Burgoon et al. (2003), and Zheng et al. (2006)	"The sentences are on average rather long with X words per sentence"
Percentage of adjectives and adverbs	Authorship recognition, deception detection	Specificity and expressiveness	Afroz et al. (2012) and Burgoon et al. (2003) (emotional component)	"There are X adjectives or adverbs which is on average rather a lot"
Frequency of function words[j]	Authorship recognition, deception detection		Abbasi and Chen (2008), Afroz et al. (2012), and Zheng et al. (2006)	–
Frequency of Part-of-Speech (PoS) tag *n*-grams[k]	Authorship recognition, clickbait detection, deception detection		Abbasi and Chen (2008), Afroz et al. (2012), Chakraborty et al. (2016), Feng et al. (2012), and Potthast et al. (2017)	–
Frequency of deep syntax structures using a PCFG	Deception detection		Feng et al. (2012)	–
Frequency of syntactic - ngrams (SN-grams)	Clickbait detection		Chakraborty et al. (2016)	
Length of syntactic dependencies	Clickbait detection	Grammatical complexity	Chakraborty et al. (2016)	–
Readability scores[l]	Authorship recognition, deception detection	Grammatical complexity	Afroz et al. (2012), Burgoon et al. (2003), and Potthast et al. (2017, 2016)	–

(Continues)

TABLE A1 (Continued)

Feature	Related task	Concepts	References	User study
Frequency of sentiments in words[m]	Clickbait detection	Sentiment	Chakraborty et al. (2016)	"There are X words with a negative sentiment and Y words with a positive sentiment"
Overall sentiment	Clickbait detection	Sentiment	Hassoun et al. (2023), Maros et al. (2021), and Potthast et al. (2016)	"There are X words with a negative sentiment and Y words with a positive sentiment"
Audio-based features				
Speech rate	Misinformation	Length of message/ number of words	El-Masri et al. (2022)	"The speech rate is rather low"
Sound volume	Misinformation		El-Masri et al. (2022)	"The sound volume is rather high"
Creator features				
Suspicious profile	Misinformation		Metzger and Flanagin (2013)	"Suspicious profile (e.g., name, description of creator)"

Abbreviation: PCFG, Probabilistic Context Free Grammar.

[a] Includes letters and digits, computed in sum and separately.

[b] Some research focuses on lies rather than on deception. As lying is a form of deception, we use the overarching term "deception detection" to refer to both areas of research.

[c] $n \in \{1, 2, 3\}$.

[d] Words that occur exactly once or exactly twice.

[e] $n \in \{1, 2, 3\}$.

[f] This feature includes the number of big words, where we define "big" as having 10 or more characters.

[g] For example,"spammers use words like 'online banking' and 'paypal'" (Feature selection section Afroz et al., 2012).

[h] Yule's K (Yule, 1944), Simpson's D (Simpson, 1949), Sichel's S (Sichel, 1975), Brunet's W (Brunet, 1978), and Honor´s R (Honoré, 1979).

[i] The same as Zheng et al. (2006), we refer for the definitions to Tweedie and Baayen (1998).

[j] We use the same function words as Zheng et al. (2006, p. 393).

[k] $n \in \{1, 2, 3\}$.

[l] Such as the following readability scores used by Potthast et al. (2017): Automated Readability Index (Smith & Kincaid, 1970), Coleman Liau Index (Coleman & Liau, 1975), Flesch Kincaid Grade Level and Reading Ease (Kincaid et al., 1975), Gunning Fog Index (Gunning, 1952), Swedish läsbarhetsindex: "readability score" (LIX) (Björnsson, 1971), McAlpine portmanteau combination of "English as a Foreign Language" and "flow" (EFLAW) Score (McAlpine, 1997), Modification of LIX (RIX) (Anderson, 1983), and (readability formula; acronym for "Simple Measure of Gobbledygook (SMOG) Grade (Mclaughlin, 1969).

[m] For example, frequency of positive words, frequency of negative words, and so forth.

TABLE A2 Description of voice messages used in out user study and the corresponding features.

Features	Voice message 1	Voice message 2	Voice message 3	Voice message 4
	Pro-Russian Misinformation (original): Female speaker claims to be witness and reports Ukrainian attacks on civilians.	Conspiracy Misinformation (original): Female speaker talks about various conspiracy theories, for example, that Germany is not a legitimate state and has a criminal organization as its alleged government.	Partly wrong (artificially generated): Man talks about a Russian propaganda video and wrongly states it was produced by a German party. The video exists but it was produced by Russia Today.	Partly misleading (artificially generated): Man talks about a selfie of a Ukrainian official in front of a portrait of Stephan Bandera. The selfie exists but the message is misleading.
Number of words	170 Words	534	225	115
Sentence length	10 per sentence	42 per sentence	13 per sentence	8 per sentence
Number of first-person pronouns	10	9	5	3
Number of adjectives and adverbs	10	34	28	14
Vocabulary richness	Moderate	Moderate	Moderate	Moderate
Words expressing certainty or uncertainty	2 Uncertainty and 3 certainty	0 Uncertainty and 7 certainty	4 Uncertainty and 5 certainty	1 Uncertainty and 2 certainty
Verbs referring to a belief	3	–	–	–
Common buzzwords related to conspiracy theories	–	For example, BRD GmbH	–	–
Number of words with positive and	Mostly negative sentiment	Mostly negative sentiment	Between neutral and negative	Balanced

(Continues)

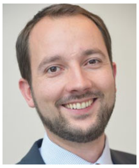
TABLE A2 (Continued)

Features	Voice message 1	Voice message 2	Voice message 3	Voice message 4
negative sentiment				
Speech rate	Slow	Moderate	Moderate and partly fast	Moderate
Speech volume	Low	Moderate to partly loud	Moderate	Moderate
Claim to be an expert	Yes	Yes	–	–
Call to action	–	Yes	–	–
Suspicious Profile	–	Yes (emojis in name)	–	–

AUTHOR BIOGRAPHIES

Katrin Hartwig is a research associate and PhD student at the Chair of Science and Technology for Peace and Security (PEASEC) in the Department of Computer Science at Technical University of Darmstadt. Her research interests lie in the intersection of computer science and psychology, particularly in the context of misinformation research and user-centered countermeasures.

Ruslan Sandler is a master's graduate from the Technical University of Darmstadt, where he studied computer science with focus on artificial intelligence. Currently, he is working as Machine Learning Engineer at Drooms.



Christian Reuter is a professor and dean of the Department of Computer Science at Technical University of Darmstadt. His chair of Science and Technology for Peace and Security (PEASEC) combines computer science with peace and security research. He holds a PhD in Information Systems (University of Siegen) and another PhD in the Politics of Safety and Security (Radboud University Nijmegen).