Empowering Users and Mitigating Harm: Leveraging Nudging Principles to Enhance Social Media Safety

Gregor Donabauer[1], **Emily Theophilou**[2], **Francesco Lomonaco**[3],
Sathya Bursic[3], **Davide Taibi**[4], **Davinia Hernández-Leo**[2]
Udo Kruschwitz[1], **Dimitri Ognibene**[3]
[1]Information Science, University of Regensburg, Regensburg, Germany
[2]Information and Communication Technologies, Pompeu Fabra University, Barcelona, Spain
[3]Department of Psychology, University of Milano-Bicocca, Milan, Italy
[4]Institute for Education Technology, National Research Council of Italy, Palermo, Italy
gregor.donabauer@ur.de, dimitri.ognibene@unimib.it

Abstract

Social media have become an integral part of our daily lives, yet they have also resulted in various negative effects on users, ranging from offensive or hateful content to the spread of misinformation. In recent years, numerous automated approaches have been proposed to identify and combat such harmful content. However, it is crucial to recognize the human aspect of users who engage with this content in designing efforts to mitigate these threats. We propose to incorporate principles of behavioral science, specifically the concept of nudging into social media platforms. Our approach involves augmenting social media feeds with informative diagrams, which provide insights into the content that users are presented. The goal of our work is to empower social media users to make well-informed decisions for themselves and for others within these platforms. Nudges serve as a means to gently draw users' attention to content in an unintrusive manner, a crucial consideration in the context of social media. To evaluate the effectiveness of our approach, we conducted a user study involving 120 Italian-speaking participants who interacted with a social media interface augmented with these nudging diagrams. Participants who had used the augmented interface were able to outperform those using the plain interface in a successive harmful content detection test where nudging diagrams were not visible anymore. Our findings demonstrate that our approach significantly improves users' awareness of potentially harmful content with effects lasting beyond the duration of the interaction. In this work, we provide a comprehensive overview of our experimental materials and setup, present our findings, and refer to the limitations identified during our study.

Keywords: Social Media, Fake News Detection, Hate Speech Detection, Nudging

1. Introduction

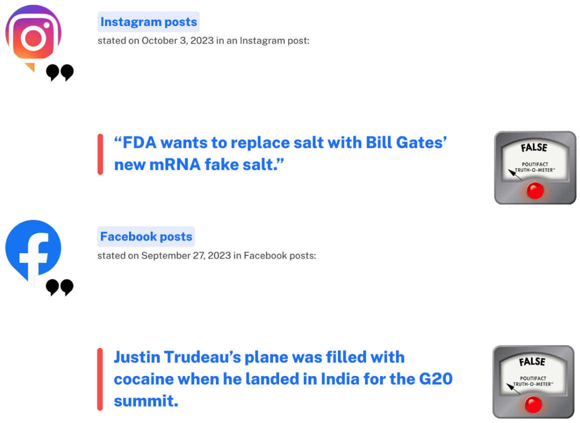
Several negative implications and threats of social media platforms have been highlighted in recent years, e.g. (Ognibene et al., 2023b). The platforms' goal of maximizing user engagement is exploiting human weaknesses with persuasive technology, resulting in extraordinarily profitable outcomes for the companies operating them (Church et al., 2023).

Two examples for serious types of harmful content spreading online are hate speech and misinformation. Hate speech posted on social media can trigger negative emotions among users and it has a low detection rate across various age and user demographics with both, younger and more experienced social media users, tending to identify hate speech content less effectively (Schmid et al., 2022). On the other side, disinformation is growing at unprecedented volumes, leading to an urgent need to tackle digital disinformation for social good, given the numerous negative implications associated with it (Shu, 2023). These problems could even get worse in the next years, as recent research has shown that large language models (LLMs) have the potential to be misused for generating misinformation that can be more challenging to identify than content written by humans (Chen and Shu, 2023; Pan et al., 2023), pointing out the urgency for proactive interventions.

Examples of fake news that have been debunked as false by the fact-checking organization Politifact[1] and are currently spreading online can be seen in Figure 1.

As it gets increasingly hard for people to recognize such harmful content, they would like to have warning labels related to posts (Kirchner and Reuter, 2020). Recent work has demonstrated that interacting with a social media feed that contains warning labels, as sometimes employed by these platforms, can have a positive effect on recognizing misinformation (Koch et al., 2023). Similarly, other studies not limited to social media platforms have shown that providing labels for news texts can improve people's ability to assess their credibility, e.g. (Kirchner and Reuter, 2020; Lu et al., 2022; Tafur and Sarkar, 2023). However, it

[1] https://www.politifact.com/



Instagram posts

stated on October 3, 2023 in an Instagram post:

“FDA wants to replace salt with Bill Gates’ new mRNA fake salt.”

Facebook posts

stated on September 27, 2023 in Facebook posts:

Justin Trudeau’s plane was filled with cocaine when he landed in India for the G20 summit.

Figure 1: Examples for fake news spreading on social media as fact checked by Politifact.

has also turned out that feeds that only partly provide warning signals, can lead to increases in the perceived credibility, even if posts are fake (Pennycook et al., 2020).

The objective of assisting users in their interaction with social media is to support them to make informed decisions for themselves and other people using such platforms. At the same time, it is especially important to not restrict their freedom of choice and assist them in a way that is as unintrusive as possible. Two principles from behavioral science that can be useful in that context are nudging (Thaler and Sunstein, 2009) and boosting (Hertwig and Grüne-Yanoff, 2017). For example, including warning lights is a nudging strategy that has be demonstrated to be effective in reducing harm in other contexts (Zimmerman et al., 2019, 2020).

We propose to make use of these strategies for supporting users in detecting potential threats on social media, while at the same time taking into account limitations of recent studies where such concepts are either not applied in more general settings (Kirchner and Reuter, 2020; Lu et al., 2022; Tafur and Sarkar, 2023) or are only applied to a limited extent when focusing on social media (Kirchner and Reuter, 2020).

To address these issues, we propose a series of experiments aimed at assessing how individuals perform in recognizing potentially harmful content after engaging with a social media interface, where all posts are labeled with information about hate speech and misinformation. Our objective is to determine whether such assistance can yield positive outcomes. To investigate this, we conduct a controlled study involving the implementation of a social media interface and comparing various experimental conditions to validate our approach. While we acknowledge that our work may have a different emphasis compared to traditional NLP contributions, our intention is to bridge the gap between algorithmic advancements in NLP and real-world user behavior. We believe that understanding the practical implications of algorithms is crucial for the holistic evaluation of NLP techniques.

In the spirit of TRAC@LREC-COLING we release all our resources, including the annotated posts, our questionnaires, and code to run the interface[2].

2. Related Work

We will start by offering a comprehensive review of various threats that can appear on social media. Furthermore, we will summarize educational strategies, with a particular focus on non-invasive methods such as nudging and boosting. Lastly, we will showcase ongoing efforts regarding the incorporation of warning labels as part of social media threat education.

2.1. Social Media Threats

Due to the diversity of content on social media and the underlying mechanisms of these platforms there is a broad range of threats occurring on such platforms that can negatively affect their users. Threat categories are spanning from content-based concerns to algorithmic issues, dynamics, cognitive challenges, and socio-emotional risks (Ognibene et al., 2023b). For our contextualization of these threats we will focus on content-based risks, as these are the ones we intend to address primarily by displaying information about posts via diagrams.

Content-based threats are not unique to classical media but manifest in distinct ways, often thriving on the web and social media. These threats include various problematic aspects, such as toxic content (Sheth et al., 2022), fake news/misinformation (Shu et al., 2017; Aïmeur et al., 2023), beauty stereotypes (Aparicio-Martinez et al., 2019), and bullying (Craig et al., 2020).

As a result, this can for example lead to body dissatisfaction and eating disorders in the case of beauty stereotypes (Aparicio-Martinez et al., 2019), increase mental distress and suicidality among youth (Abi-Jaoude et al., 2020), or threaten democracy, justice, public trust, freedom of expression, journalism, and economic growth in the case of misinformation (Shu, 2023).

Given the importance of these threats, various research directions focus on the development of dedicated detection systems. Examples include fake news (Bhattarai et al., 2022; Hartl and Kruschwitz, 2022; Guo et al., 2022; Donabauer and Kruschwitz, 2023), hate speech (Zampieri et al.,

[2] https://github.com/DimNeuroLab/COURAGE_api

2022; Jahan et al., 2022; Ababu and Woldeyohannis, 2022) or offensive language detection (Ajvazi and Hardmeier, 2022; Hoefels et al., 2022).

2.2. Education About Threats: Nudging and Boosting

In response to the negative impact of social media use on its users, educators and researchers have been actively engaged in developing and delivering interventions aimed at promoting social media literacy and responsible online behaviors (Guess et al., 2020; Gordon et al., 2021; Sánchez-Reina et al., 2021; Theophilou et al., 2023). These interventions encompass a wide range of educational materials and tools, such as workshops, online courses, games, and awareness campaigns, often delivered in schools. Their goal is to empower individuals with the knowledge and skills necessary to critically assess the information they encounter online. Despite these efforts, not all segments of the population can take advantage of these educational opportunities (Lee, 2018). This is due to a significant portion of the social media population being over 18 and no longer enrolled in educational institutions[3].

To bridge this gap and further support social media users in their daily interactions, there is a growing consensus on the importance of integrating unobtrusive features directly into these platforms to raise awareness regarding potentially negative aspects (Morrow et al., 2022). These features can enhance the transparency of social media platforms by providing valuable information on a range of topics, including misinformation (Saltz et al., 2021), image editing (Rodríguez-Rementería et al., 2022), and the hidden engineering of social media (Ognibene et al., 2023a).

Integrating unobtrusive features directly into social media platforms can raise awareness about potential negative aspects and discourage belief in misinformation. Seamlessly embedding tools within the platforms that users already engage with can have an important immediate impact, from self-reflection (Purohit et al., 2020) to misinformation identification (Grady et al., 2021; Epstein et al., 2022).

Behavioral and cognitive science strategies offer a well-founded framework for subtly influencing people's behavior, which is especially important in settings such as social media. Two such paradigms are nudging (Thaler and Sunstein, 2009) and boosting (Hertwig and Grüne-Yanoff, 2017), both of which leverage behavioral patterns to subtly influence people's behavior without restricting their freedom of choice.

Nudging (Thaler and Sunstein, 2009) represents a behavioral public policy approach designed to support individuals in making better choices through the "choice architecture" of their environment, which includes aspects such as default settings.

However, their inherent limitation lies in their inability to teach new skills or competencies. Consequently, when a nudge is removed, users tend to revert to their previous behavior without having acquired any lasting knowledge.

This is where the concept of boosting offers an alternative approach. Unlike nudges, boosts prioritize interventions that enhance individuals' competence in making independent decisions (Hertwig and Grüne-Yanoff, 2017).

An example of a tool integrated in social media leveraging the boosting mechanism is the one proposed by (Aprin et al., 2022). This work integrates a virtual learning companion that guides users through a process to identify the credibility of images. The companion does not simply label content as credible or non-credible; instead, it provides educational materials and critical thinking exercises to help users learn how to assess the credibility of images on their own.

On the contrary, an approach utilizing the nudging strategy in the form of a web-browser plugin is the one proposed by (Kyza et al., 2021). This plugin evaluates the credibility of tweets and uses a nudging mechanism to allow users to blur out low-credibility tweets by customizing their preferences. This nudging mechanism directly blurs out content, but other forms of nudging, such as warning lights and information nutrition labels, also have the potential to reduce harm and risks in web searches (e.g. Zimmerman et al. (2020)).

Nudges are particularly suitable for integration into social media interfaces, as they generally impose minimal additional cognitive burden on users. In addition, the objective of assisting users on social media is to support them to make informed decisions for themselves and other people using such platforms. Nudges offer a way to push content to users, making them aware of it in a way as unintrusive as possible, something particularly important in contexts like social media.

2.3. Warning Labels and Social Media

Social media platforms have introduced features to warn users about potentially misleading content, for example on Facebook[4] as well as Twitter/X[5]. These warnings are valuable signals that can help users assess the credibility of the information they are about to access. Such in-platform measures could play a significant role in curbing the spread of misinformation and improving the overall user

[3] `https://backlinko.com/social-media-users`

[4] `https://about.fb.com/news/tag/misinformation/`

[5] `https://communitynotes.x.com/guide/en/about/introduction`

experience which is why assessing the impact of such flagging is important to determine the usefulness of their functionality.

Research has been conducted to investigate the impact of warning labels, specifically those related to misinformation, on users' perception of news articles. Typically, participants are presented with articles that could be shared on social media, accompanied by warning labels and then give them the task to assess the authenticity of the content (Clayton et al., 2020; Kirchner and Reuter, 2020; Pennycook et al., 2020). These experiments have shown that people perform better in identifying misinformation when they have access to ground truth labels during the annotation process. However, it is important to note that these experiments do not replicate the real-world dynamics of using social media platforms as these studies only present the news articles as screenshots and the labeled information is visible during evaluation of user awareness.

In a more realistic setting, Seo et al. (2019) show screenshots to participants that simulate Facebook posts, rather than presenting plain text, while Koch et al. (2023) provide an interface mimicking a social media platform. However, in the case of Koch et al. (2023), only one post in the feed is labeled, leaving the remaining posts unlabeled. Pennycook et al. (2020) have shown that such partial labeling can negatively impact the perceived credibility of other posts in the feed.

Other studies have introduced variations in the experimental setup by including partially incorrect annotations, simulating results of machine learning classifiers (Lu et al., 2022; Tafur and Sarkar, 2023). When the classifier performance is too low in such settings, participants' annotation performance also suffers, as observed by Snijders et al. (2023); Theophilou et al. (2023).

Seo et al. (2019) argue that providing participants with training that demonstrates the positive effects of labels on identifying potentially harmful content can lead to improvements. This approach has not been widely adopted in related work, presenting a gap that we try to fill by evaluating the impact of a training phase in our experiments.

It is worth noting that forms of threats appearing on social media are multifaceted and not only limited to fake news. The studies presented so far have solely focused on misinformation detection, e.g. (Kirchner and Reuter, 2020; Snijders et al., 2023; Koch et al., 2023). We extend these evaluations by including additional warning labels for hate speech.

While most studies show news items along with labels during the annotation process, this approach may encourage participants to only rely on the provided labels and does not allow to measure whether or not the labels provide a lasting effect independently of the explicit task they are involved in during the experiment that can bias the results. In contrast, Lu et al. (2022) and Seo et al. (2019) present the only two studies (to the best of our knowledge) where article labels are shown before the annotation phase (Lu et al., 2022) or where participants first annotate labeled articles, then re-annotate the same articles without labels (Seo et al., 2019).

Our research aims to evaluate a more realistic process when encountering such labels in a feed by subsequently requiring them to annotate content without the benefit of ground truth during annotation while in addition considering multiple posts for reflection.

3. Materials

3.1. Interface

In general, the interface developed for the experiments mimics the well-known social media platform X (formerly Twitter) in its appearance. This includes a navigation menu on the left side, the actual feed in the center as well as some topic and page recommendations on the right side.

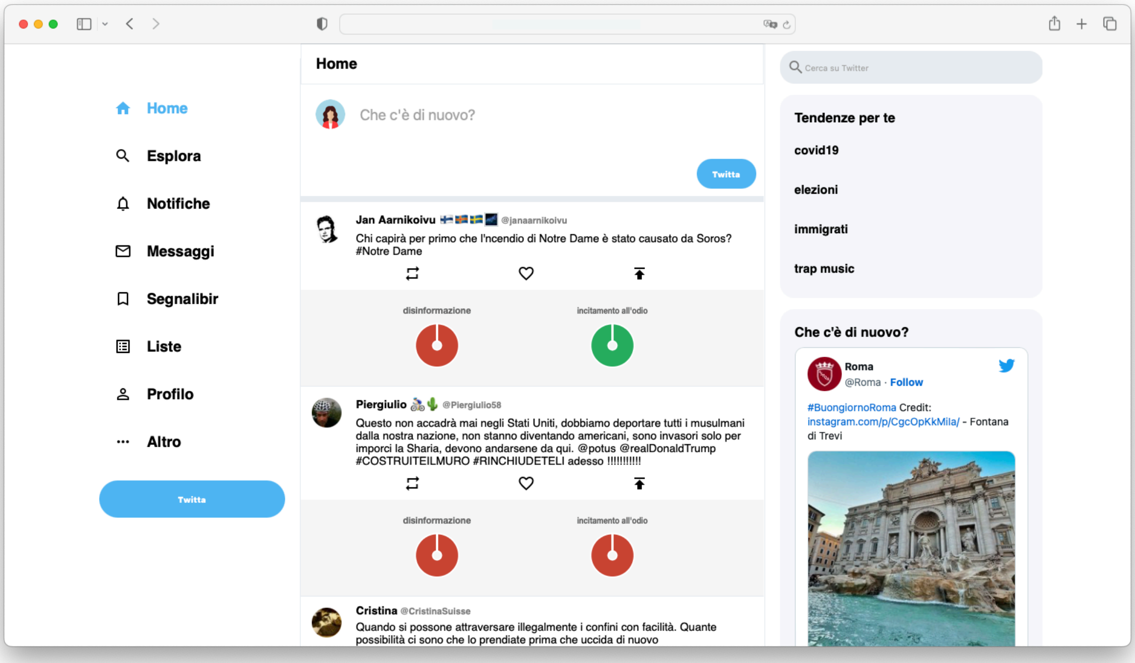
For our investigations we have developed two versions of the interface:

- a plain social media feed without any additional information regarding hatefulness or fakeness of posts;
- the same interface with additional, interactive diagrams that provide information about the checked characteristics (see Figure 2).

The diagrams are titled with the respective information they hold (misinformation and hate speech) and are colored either fully in green (i.e. no misinformation and hate speech) or fully in red (i.e. contains misinformation or hate speech). When hovering over the diagram the same label as indicated by the color appears. Colors and shape of these diagrams are inspired by stoplights which have proven to be effective in reducing harm in search (Zimmerman et al., 2019).

The interaction opportunities are limited to the feeds at the center of the interface to put the participants' focus on that area and prevent them from unintended behavior not related to the actual experiment. We have added these restrictions to maintain control over the experimental setting, a practice commonly employed in experiments involving web pages to ensure a greater degree of control over the overall interactions, e.g. (Pogacar et al., 2017).

Furthermore, we eliminated any form of social endorsement cues, given their potential influence on the perception of posts (Ali et al., 2022; Shin et al.,



Jan Aarnikoivu @janaarnikoivu
Chi capirà per primo che l'ncendio di Notre Dame è stato causato da Soros? #Notre Dame

Piergiulio @Piergiulio58
Questo non accadrà mai negli Stati Uniti, dobbiamo deportare tutti i musulmani dalla nostra nazione, non stanno diventando americani, sono invasori solo per imporci la Sharia, devono andarsene da qui. @potus @realDonaldTrump #COSTRUITEILMURO #RINCHIUDETELI adesso !!!!!!!!!!

Cristina @CristinaSuisse
Quando si possone attraversare illegalmente i confini con facilità. Quante possibilità ci sono che lo prendiate prima che uccida di nuovo

Roma
@Roma · Follow
#BuongiornoRoma Credit: instagram.com/p/CgcOpKkMila/ - Fontana di Trevi

Figure 2: Interface as used in conditions 2 and 3 with diagram augmented feed.

2022) while our objective is to assess the effects of diagrams containing information about the posts.

3.2. Posts

Overall, we set the number of posts in the feed to eight to avoid information overload (Edson C Tandoc and Kim, 2023). All posts are actual tweets from Twitter, sometimes with slight modifications in their wording as we translated most of them from English into Italian. To ensure that the translation are of high quality we did not rely on automated approaches but performed it manually. We used posts from a diverse set of topics: (1) Notre Dame fire; (2) Charlie Hebdo attacks and (3) Immigration. We selected these topics as we sourced a large proportion of them from annotated datasets in the domain which partly provides us with ground truth information for the diagrams and later evaluation. As posts with images/links draw more attention of the user (Vraga et al., 2016), we only include posts that solely contain textual content to prevent adding bias. The datasets we used to select the tweets from are Zubiaga et al. (2016) for misinformation and Basile et al. (2019) for hate speech. We decided for these resources as we required datasets that contain labels corresponding, at least in part, to the categories in our diagrams. The dataset should also contain tweets, and the tweet-IDs allowing us to recrawl profile-related meta-information, which we presented within the feed. Additionally, we made efforts to ensure that the content of the posts did not solely represent obvious misinformation but rather included inaccurate details about events, for instance. As we provide two labels for each of the posts but most of the time only have parts of the information available we annotate the remaining characteristics on our own. One of the authors served as annotator following the guidelines provided for the original dataset when annotating hate speech (Basile et al., 2019), and proceeded to label the previously unlabeled posts. For each post in the feed we use the annotations obtained through this process as ground truth annotations for fake news and hate speech.

4. Experiments

4.1. Procedure

The study begins by informing the participants about its relation to a research project. During this initial phase, participants provide informed consent for their participation. Additionally, we offer an explanation of certain aspects of the interface, particularly those that are different from their familiarity with conventional social media platforms, such as the inclusion of supplementary diagrams. After this step, the actual interaction with the interface begins. To ensure their active involvement, and to prevent them from skipping after a few seconds (reducing the participation time results in higher payment per hour), we included a hidden timer in the interface. After two minutes, an alert is triggered, displaying a code, and we expect them to copy this code into the first field of the subsequent questionnaire. Participants are informed of this process before they are directed to the interface. However, it remained possible for participants to continue spending additional time on the feed, as we did not impose any restrictions on their interactions with the interface after displaying the code. We note that adding a timer might also have the opposite effect, potentially leading individuals to pay less attention, as observed in previous NLP annotation tasks (Chamberlain, 2015).

After the participants are done spending time on interacting with the interface, they are forwarded to the annotation phase of the questionnaire. In this phase, participants are presented with the previously viewed posts one after the other. Their objective is to identify whether each post contains either misinformation or hateful content. Apart from that they have to submit a confidence value, representing how sure they are about their annotations. To enhance the complexity of the task, the presentation order of the posts during annotation differs from their order within the interface. Additionally, we remove visual cues such as profile images and usernames (and of course diagrams). To check whether the participants are paying attention during this phase we include an attention check (Abbey and Meloy, 2017). The check is done by adding an additional artificial post text that advises the participant to mark both misinformation and hate speech as *false*. Thus, random or inattentive annotations are likely to fail this check.

Lastly, participants are required to provide demographic information and respond to questions about their typical social media usage behaviour.

4.2. Conditions

To compare how labeling of social media content influences users' awareness and understanding of social media threats in an realistic environment, we compared different conditions with each other:

1. **No Training and no Diagrams:** For the baseline condition we presented a plain feed without any further information on hate speech and fake news to the participants. This setup reflects the standard interaction of users with a social media platform.

2. **No Training but Diagrams:** The second condition introduces diagrams to the feed which hold information about the posts that are displayed. These diagrams represent the ground truth labels. As this style of adding information to posts is new to the participants we also introduce a third condition that includes a training phase to make the participants familiar with the concept.

3. **Training and Diagrams:** During the training phase the participants get presented two post and their associated annotation diagrams. They are asked to annotate whether the posts contain misinformation or hate speech. After submitting their annotations they get immediate feedback in form of point scores (correct annotations lead to better scores). The information displayed in the diagrams again represents the ground truth (same as in condition 2) which means that relying on these labels leads to higher scores and teaches their usefulness to the participants.

5. Results

5.1. Participants

During spring/summer 2023 we recruited 40 participants for each of the three conditions on Prolific, employing a between-groups design. This approach resulted in an overall sample size of $N = 120$ (which is a similar number compared to the ones as reported in related studies, e.g. Tafur and Sarkar (2023): 40 participants; Snijders et al. (2023): 110 participants; and Theophilou et al. (2023): 144 participants). We chose this experimental design to prevent information leakage during the study, as we utilized the same set of posts for all conditions to increase comparability. Presenting diagrams in one phase might influence the subsequent annotation phases in another condition. All participants are native Italian speakers. To make sure that the data collected are of high quality, we excluded participants who did not pass an attention check. Interestingly, this did not apply to any of the people taking part in the final study. On average they were 31.12 years old ($std = 10.67$), 54% were male ($n = 65$), 42% female ($n = 50$) and 4% of other gender ($n = 5$). In terms of highest degree obtained the participants were rather highly educated: middle school or lower ($n = 2$, 1.7%); high school diploma ($n = 62$, 52%); Bachelor degree ($n = 28$, 23%); Master degree ($n = 24$, 20%); PhD ($n = 2$, 1.7%); other ($n = 2$, 1.7%).

We also asked them about their social media routines. 12% spend less than one hour a day ($n = 14$), 30% between one and two hours a day ($n = 36$), 19% between two and three hours a day ($n = 23$), another 19% between three and four hours a day ($n = 23$) and 20% even more than four hours a day ($n = 24$) on social media platforms. 43% ($n = 52$) replied that checking social media is the first thing they do in the morning, compared to 57% ($n = 68$) who do not do so.

5.2. Detection Performance

For each post in the feed we have ground truth information for fake news and hate speech. We use the annotations submitted by each participant to calculate a metric for their performance in detecting fake/hateful posts. We use the accuracy and macro F1 metrics. As a result, we get a list of values for each condition, representing the performance of participants in this group. For simplicity we will only report detailed results for macro F1 in this section. However, we note that the accuracy scores are highly similar and we provide detailed statistics for both metrics in our GitHub repository.

Condition	F1 Hate Speech	F1 Fake News
nT-nD	0.799	0.763
nT-D	**0.886**	0.869
T-D	0.877	**0.890**

Table 1: Average macro F1 scores for detection performance of hate speech and fake news between different experimental conditions. nT-nD = no training and no diagrams; nT-D = no training but diagrams; T-D = training and diagrams.

Table 1 shows the mean detection performance (F1 scores) of participants within each group. Additionally, for a more comprehensive perspective, we have included a detailed overview in Figure 3 for fake news detection and Figure 4 for hate speech detection using boxplots.

In order to evaluate the differences, we conduct tests to determine their significance. First, we test for normal distribution within each group. Since some of the values are not normally distributed we apply a Kruskal-Wallis test for independent samples. As the results are significant at $p < 0.01$ for all conditions we apply a post hoc pairwise test for multiple comparisons with Bonferroni correction to adjust the p-values.

In terms of hate speech detection performance, we observe a statistically significant difference with a p-value of slightly smaller than 0.01 between conditions nt-nD and nT-D, as well as a p-value of 0.038 between conditions nT-D and T-D. However, no statistically significant difference is evident between conditions nT-D and T-D.

Similar trends can be observed in the performance of fake news detection, with p-values that are much smaller than 0.01 for comparisons between conditions nT-nD and nT-D, as well as between conditions nT-nD and T-D. Once again, there is no statistical distinction between conditions nT-D and T-D.

We additionally conduct Cohen's d tests between the groups. Consistent with the findings from Kruskal-Wallis tests and Bonferroni correction, the effect size between groups nt-nD and nT-D is calculated at 0.59, and for groups nT-nD and T-D, it is 0.58. Moreover, the effect size between conditions nT-D and T-D is negligible, with Cohen's d amounting to only 0.06.

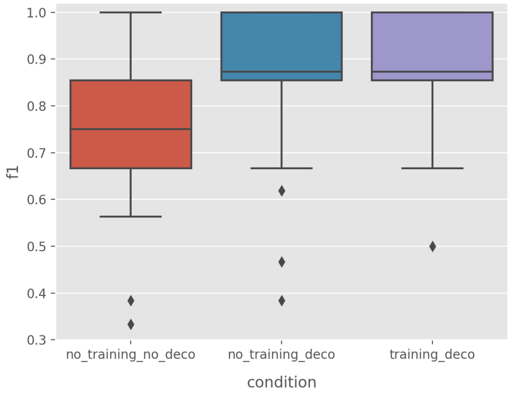


Figure 3: Boxplot for macro F1 fake news detection performance scores between conditions.

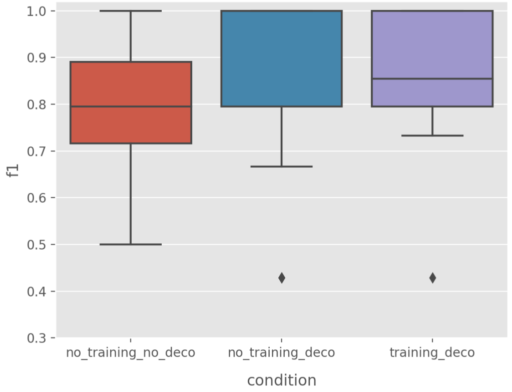


Figure 4: Boxplot for macro F1 hate speech detection performance scores between conditions.

In summary, the results indicate that groups receiving annotated posts during their interaction with the interface perform significantly better in labeling these afterwards in terms of fake news and hate speech. However, adding a training phase to demonstrate the usefulness of the decorations does not yield additional significant benefits. While these differences are not significant, it is worth mentioning that the incorporation of a training phase resulted in slightly better performance in detecting fake news.

6. Discussion and Limitations

Below, we will discuss findings and potential limitations of our study. One important finding is that incorporating diagrams consistently results in significantly higher performance when identifying potentially harmful content, compared to viewing a plain social media feed. This suggests that our evaluated approach appears to have the desired effects, even when users no longer see the annotations when assessing posts. This shows a lasting and unbiased effect of the approach. However, it is important to note that our experiments involved a limited number of posts. It would be interesting to explore whether similar effects can be observed when users are presented with a larger number of posts, as this could potentially lead to information overload or habituation effects.

Another observation is that there is no statistically significant difference between the two conditions involving diagrams. The training phase does not yield significantly positive effects on participants' performance and, in the case of hate speech detection, even results in a slightly worse result compared to the group that did not have a training phase. However, these differences are very small and the opposite trend is observed for fake news detection. In summary, this suggests that the diagrams are self-explaining and do not necessarily require a training phase before. However, further investigation is needed to understand why the training phase did not yield more substantial benefits.

In general, across all three conditions we can observe relatively good performance, with the lowest F1 scores starting at 0.799 for hate speech detection and 0.763 for fake news detection within the group that did not see any additional decorations. One reason for this might be that the attributes we evaluated are relatively easy to identify in the posts we used in our experiments. To obtain more generalizable results, it would be beneficial to repeat the experiments using a different set of more challenging posts, diverse topics, or other attributes to check than hate speech and fake news. Our results are also limited by the fact that we only looked at posts in a single language (Italian). In any case, we consider the experiments we conducted as a stepping stone for others to explore these different dimensions so that we get a clear picture what approaches are most effective in addressing threats on social media without imposing any restrictions on the user's autonomy.

One aspect that we did not consider is the possibility of incorrectly labeled posts (i.e. inaccurate diagrams). Given that assessing content on social media often is based on automated approaches, such as machine learning detectors, it would be interesting to explore whether users follow wrong annotations or show enough critical thinking to notice inaccurate labels. Educational activity aimed at counterbalancing AI failure and AI overdependence would be crucial in this setting (Theophilou et al., 2023).

Lastly, it is important to note that our study was conducted on desktop computers rather than handheld devices. Existing research suggests that significant differences exist when compared to mobile devices. For example, higher engagement on desktop computers than on mobile devices when it comes to news consumption time (Dunaway et al., 2018) and user attention to social media posts (Keib et al., 2022).

7. Conclusion

Threats faced by social media users in relation to the content they encounter on these platforms have become an increasing problem. We proposed an unintrusive approach to support users in making informed decisions for both themselves and others when using such platforms. Our approach makes use of principles from behavioral science, such as nudging. We demonstrated that enhancing the social media feed with diagrams that contain information about the posts significantly improves users' ability to identify potentially harmful content even when not explicitly asked to do so (as the task is presented when the diagrams are not visible anymore). We show that these diagrams are intuitively understandable and do not require additional participant training.

An interesting finding is also the observation that the **nudges** we deploy actually demonstrate properties that more resemble the idea of **boosts** in that they appear to teach some practical skill. For future it might be worthwhile to explore a range of different nudging and boosting techniques as each one might, for example, be effective for different audiences (Lorenz-Spreen et al., 2020).

In conclusion, our findings present promising directions in reducing content-related threats on social media platforms. To foster reproducibility we will make all our resources available. We hope that our results can serve as a benchmark for future experimental work.

8. Ethical Considerations

It is important to balance support of users in making informed decisions about potentially harmful content on social media while at the same time maintaining principles like transparency, free expression, and privacy. Below we will summarize several ethical considerations related to our study: One central point is freedom of expression. We recognize that the line between harmful content and legitimate discourse can be blurred, resulting in a need for clear guidelines. This also means that the accuracy of our evaluated diagrams is crucial. If they are inaccurate or misleading, they may worsen the problem by spreading false information. Augmenting posts might also be considered as censorship if content is wrongly categorized as harmful. Therefore, potential effects on free expression should be minimized. One way of doing so is to acknowledge that the augmentation should be optional, allowing users to choose whether or not to view the diagrams. We do not intend to force or intrusively augment content resulting in a violation of users' autonomy and privacy.

It is also worth noting that it varies across cultures and countries what is considered harmful content. Implementing such a system on a global scale requires sensitivity to these differences and respecting local laws and norms. In addition, algorithms that could be used to automate the analysis of the posts can be biased, leading to false positives or negatives. This again could affect certain groups and restrict free expression. Thus, in such a case ensuring fairness and minimizing bias is crucial.

We acknowledge that the impact of augmented posts on user behavior, perceptions, and the overall information ecosystem should also be monitored over time to be able to draw more detailed conclusions about the effects of the diagrams.

9. Acknowledgements

We would like to thank the anonymous reviewers for their constructive feedback which has helped us improve the paper.

This work was supported by the project COURAGE: A Social Media Companion Safeguarding and Educating Students funded by the Volkswagen Foundation, grant number 95564.

10. References

Teshome Mulugeta Ababu and Michael Melese Woldeyohannis. 2022. Afaan Oromo Hate Speech Detection and Classification on Social Media. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6612–6619, Marseille, France. European Language Resources Association.

James D. Abbey and Margaret G. Meloy. 2017. Attention by design: Using attention checks to detect inattentive respondents and improve data quality. *Journal of Operations Management*, 53-56:63–70.

Elia Abi-Jaoude, Karline Treurnicht Naylor, and Antonio Pignatiello. 2020. Smartphones, social media use and youth mental health. *CMAJ*, 192(6):E136–E141.

Adem Ajvazi and Christian Hardmeier. 2022. A Dataset of Offensive Language in Kosovo Social Media. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1860–1869, Marseille, France. European Language Resources Association.

Khudejah Ali, Cong Li, Khawaja Zain ul abdin, and Syed Ali Muqtadir. 2022. The effects of emotions, individual attitudes towards vaccination, and social endorsements on perceived fake news credibility and sharing motivations. *Computers in Human Behavior*, 134:107307.

Pilar Aparicio-Martinez, Alberto-Jesus Perea-Moreno, María Pilar Martinez-Jimenez, María Dolores Redel-Macías, Claudia Pagliari, and Manuel Vaquero-Abellan. 2019. Social Media, Thin-Ideal, Body Dissatisfaction and Disordered Eating Attitudes: An Exploratory Analysis. *International Journal of Environmental Research and Public Health*, 16(21).

Farbod Aprin, Irene Angelica Chounta, and H. Ulrich Hoppe. 2022. "See the Image in Different Contexts": Using Reverse Image Search to Support the Identification of Fake News in Instagram-Like Social Media. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13284 LNCS:264–275.

Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(30):1869–5469.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Bimal Bhattarai, Ole-Christoffer Granmo, and Lei Jiao. 2022. Explainable Tsetlin Machine Framework for Fake News Detection with Credibility Score Assessment. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4894–4903, Marseille, France. European Language Resources Association.

Jon Chamberlain. 2015. *Harnessing Collective Intelligence on Social Networks*. University of Essex. PhD Thesis.

Canyu Chen and Kai Shu. 2023. Can LLM-Generated Misinformation Be Detected? Https://arxiv.org/abs/2309.13788.

Kenneth Church, Annika Schoene, John E. Ortega, Raman Chandrasekar, and Valia Kordoni. 2023. Emerging trends: Unfair, biased, addictive, dangerous, deadly, and insanely profitable. *Natural Language Engineering*, 29(2):483–508.

Katherine Clayton, Spencer Blair, Jonathan A Busam, Samuel Forstner, John Glance, Guy

Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, et al. 2020. Real solutions for fake news? measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political behavior*, 42:1073–1095.

Wendy Craig, Meyran Boniel-Nissim, Nathan King, Sophie D. Walsh, Maartje Boer, Peter D. Donnelly, Yossi Harel-Fisch, Marta Malinowska-Cieślik, Margarida Gaspar de Matos, Alina Cosma, Regina Van den Eijnden, Alessio Vieno, Frank J. Elgar, Michal Molcho, Ylva Bjereld, and William Pickett. 2020. Social Media Use and Cyber-Bullying: A Cross-National Analysis of Young People in 42 Countries. *Journal of Adolescent Health*, 66(6, Supplement):S100–S108. Understanding Adolescent Health and Wellbeing in Context: Cross-National Findings from the Health Behaviour in School-aged Children Study.

Gregor Donabauer and Udo Kruschwitz. 2023. Exploring fake news detection with heterogeneous social media context graphs. In *Advances in Information Retrieval*, pages 396–405, Cham. Springer Nature Switzerland.

Johanna Dunaway, Kathleen Searles, Mingxiao Sui, and Newly Paul. 2018. News Attention in a Mobile Era. *Journal of Computer-Mediated Communication*, 23(2):107–124.

Jr Edson C Tandoc and Hye Kyung Kim. 2023. Avoiding real news, believing in fake news? investigating pathways from information overload to misbelief. *Journalism*, 24(6):1174–1192.

Ziv Epstein, Nicolo Foppiani, Sophie Hilgard, Sanjana Sharma, Elena Glassman, and David Rand. 2022. Do Explanations Increase the Effectiveness of AI-Crowd Generated Fake News Warnings? *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):183–193.

Chloe S. Gordon, Hannah K. Jarman, Rachel F. Rodgers, Siân A. McLean, Amy Slater, Matthew Fuller-Tyszkiewicz, and Susan J. Paxton. 2021. Outcomes of a Cluster Randomized Controlled Trial of the SoMe Social Media Literacy Program for Improving Body Image-Related Outcomes in Adolescent Boys and Girls. *Nutrients*, 13(11).

Rebecca Grady, Peter Ditto, and Elizabeth Loftus. 2021. Nevertheless, partisanship persisted: fake news warnings help briefly, but bias returns with time. *Cognitive Research: Principles and Implications*, 6.

Andrew M. Guess, Michael Lerner, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. 2020. A digital media literacy intervention increases discernment between mainstream and false news in the united states and india. *Proceedings of the National Academy of Sciences*, 117(27):15536–15545.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Philipp Hartl and Udo Kruschwitz. 2022. Applying Automatic Text Summarization for Fake News Detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2702–2713, Marseille, France. European Language Resources Association.

Ralph Hertwig and Till Grüne-Yanoff. 2017. Nudging and boosting: Steering or empowering good decisions. *Perspectives on Psychological Science*, 12(6):973–986.

Diana Constantina Hoefels, Çağrı Çöltekin, and Irina Diana Mădroane. 2022. CoRoSeOf - An Annotated Corpus of Romanian Sexist and Offensive Tweets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2269–2281, Marseille, France. European Language Resources Association.

Md Saroar Jahan, Mourad Oussalah, and Nabil Arhab. 2022. Finnish Hate-Speech Detection on Social Media Using CNN and FinBERT. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 876–882, Marseille, France. European Language Resources Association.

Kate Keib, Bartosz W. Wojdynski, Camila Espina, Jennifer Malson, Brittany Jefferson, and Yen-I Lee. 2022. Living at the Speed of Mobile: How Users Evaluate Social Media News Posts on Smartphones. *Communication Research*, 49(7):1016–1032.

Jan Kirchner and Christian Reuter. 2020. Countering Fake News: A Comparison of Possible Solutions Regarding User Acceptance and Effectiveness. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2).

Timo K. Koch, Lena Frischlich, and Eva Lermer. 2023. Effects of fact-checking warning labels and social endorsement cues on climate change fake news credibility and engagement on social media. *Journal of Applied Social Psychology*, 53(6):495–507.

Eleni Kyza, Christiana Varda, Loukas Konstantinou, Evangelos Karapanos, Serena Coppolino Perfumi, Mattias Svahn, and Yiannis Georgiou. 2021. Social media use, trust and technology acceptance: Investigating the effectiveness of a co-created browser plugin in mitigating the spread of misinformation on social media. *AoIR Selected Papers of Internet Research*.

Nicole M. Lee. 2018. Fake news, phishing, and fraud: a call for research on digital media literacy education beyond the classroom. *Communication Education*, 67(4):460–466.

Philipp Lorenz-Spreen, Stephan Lewandowsky, Cass R. Sunstein, and Ralph Hertwig. 2020. How behavioural sciences can promote truth, autonomy and democratic discourse online. *Nature Human Behaviour*, 4(11):1102–1109.

Zhuoran Lu, Patrick Li, Weilong Wang, and Ming Yin. 2022. The Effects of AI-Based Credibility Indicators on the Detection and Spread of Misinformation under Social Influence. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Garrett Morrow, Briony Swire-Thompson, Jessica Montgomery Polny, Matthew Kopec, and John P. Wihbey. 2022. The emerging science of content labeling: Contextualizing social media content moderation. *Journal of the Association for Information Science and Technology*, 73(10):1365–1386.

Dimitri Ognibene, Gregor Donabauer, Emily Theophilou, Sathya Buršić, Francesco Lomonaco, Rodrigo Wilkens, Davinia Hernández-Leo, and Udo Kruschwitz. 2023a. Moving Beyond Benchmarks and Competitions: Towards Addressing Social Media Challenges in an Educational Context. *Datenbank-Spektrum*.

Dimitri Ognibene, Rodrigo Wilkens, Davide Taibi, Davinia Hernández-Leo, Udo Kruschwitz, Gregor Donabauer, Emily Theophilou, Francesco Lomonaco, Sathya Bursic, Rene Alejandro Lobo, J. Roberto Sánchez-Reina, Lidia Scifo, Veronica Schwarze, Johanna Börsting, Ulrich Hoppe, Farbod Aprin, Nils Malzahn, and Sabrina Eimler. 2023b. Challenging social media threats using collective well-being-aware recommendation algorithms and an educational virtual companion. *Frontiers in Artificial Intelligence*, 5.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023. On the Risk of Misinformation Pollution with Large Language Models. Https://arxiv.org/abs/2305.13661.

Gordon Pennycook, Adam Bear, Evan T. Collins, and David G. Rand. 2020. The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings. *Management Science*, 66(11):4944–4957.

Frances A. Pogacar, Amira Ghenai, Mark D. Smucker, and Charles L.A. Clarke. 2017. The Positive and Negative Influence of Search Results on People's Decisions about the Efficacy of Medical Treatments. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '17, page 209–216, New York, NY, USA. Association for Computing Machinery.

Aditya Kumar Purohit, Louis Barclay, and Adrian Holzer. 2020. Designing for Digital Detox: Making Social Media Less Addictive with Digital Nudges. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, page 1–9, New York, NY, USA. Association for Computing Machinery.

Amaia Rodríguez-Rementería, Roberto Sanchez-Reina, Emily Theophilou, and Davinia Hernández-Leo. 2022. Actitudes sobre la edición de imágenes en redes sociales y su etiquetado: un posible preventivo. In *EDUTEC 2022, XXV Congreso internacional*, pages 334–336, Palma, España. IRIE.

Emily Saltz, Claire R Leibowicz, and Claire Wardle. 2021. Encounters with Visual Misinformation and Labels Across Platforms: An Interview and Diary Study to Inform Ecosystem Approaches to Misinformation Interventions. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

Ursula Kristin Schmid, Anna Sophie Kümpel, and Diana Rieger. 2022. How social media users perceive different forms of online hate speech: A qualitative multi-method study. *New Media & Society*, page 14614448221091185.

Haeseung Seo, Aiping Xiong, and Dongwon Lee. 2019. Trust It or Not: Effects of Machine-Learning Warnings in Helping Individuals Mitigate Misinformation. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, page 265–274, New York, NY, USA. Association for Computing Machinery.

Amit Sheth, Valerie L. Shalin, and Ugur Kursuncu. 2022. Defining and detecting toxicity on social media: context and knowledge are key. *Neurocomputing*, 490:312–318.

Inyoung Shin, Luxuan Wang, and Yi-Ta Lu. 2022. Twitter and Endorsed (Fake) News: The Influence of Endorsement by Strong Ties, Celebrities, and a User Majority on Credibility of Fake News During the COVID-19 Pandemic. *International Journal of Communication*, 16(0).

Kai Shu. 2023. Combating Disinformation on Social Media and Its Challenges: A Computational Perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13):15454–15454.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36.

Chris Snijders, Rianne Conijn, Evie de Fouw, and Kilian van Berlo. 2023. Humans and algorithms detecting fake news: Effects of individual and contextual confidence on trust in algorithmic advice. *International Journal of Human–Computer Interaction*, 39(7):1483–1494.

J. R. Sánchez-Reina, E. Theophilou, D. Hernández-Leo, and P. Medina-Bravo. 2021. *The power of beauty or the tyranny of algorithms: How do teens understand body image on Instagram?*, pages 429–450. Editorial Dykinson S.L., Sevilla.

Bruno Tafur and Advait Sarkar. 2023. User Perceptions of Automatic Fake News Detection: Can Algorithms Fight Online Misinformation?

Richard H Thaler and Cass R Sunstein. 2009. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Penguin.

Emily Theophilou, Francesco Lomonaco, Gregor Donabauer, Dimitri Ognibene, Roberto J. Sánchez-Reina, and Davinia Hernàndez-Leo. 2023. AI and Narrative Scripts to Educate Adolescents About Social Media Algorithms: Insights About AI Overdependence, Trust and Awareness. In *Responsive and Sustainable Educational Futures*, pages 415–429, Cham. Springer Nature Switzerland.

Emily Vraga, Leticia Bode, and Sonya Troller-Renfree. 2016. Beyond Self-Reports: Using Eye Tracking to Measure Topic and Style Differences in Attention to Social Media Content. *Communication Methods and Measures*, 10(2-3):149–164.

Himanshu Zade, Megan Woodruff, Erika Johnson, Mariah Stanley, Zhennan Zhou, Minh Tu Huynh, Alissa Elizabeth Acheson, Gary Hsieh, and Kate Starbird. 2023. Tweet Trajectory and AMPS-Based Contextual Cues Can Help Users Identify Misinformation. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1).

Nicolas Zampieri, Carlos Ramisch, Irina Illina, and Dominique Fohr. 2022. Identification of Multiword Expressions in Tweets for Hate Speech Detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 202–210, Marseille, France. European Language Resources Association.

Steven Zimmerman, Alistair Thorpe, Jon Chamberlain, and Udo Kruschwitz. 2020. Towards Search Strategies for Better Privacy and Information. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, CHIIR '20, pages 124–134. Association for Computing Machinery.

Steven Zimmerman, Alistair Thorpe, Chris Fox, and Udo Kruschwitz. 2019. Investigating the Interplay Between Searchers' Privacy Concerns and Their Search Behavior. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 953–956, New York, NY, USA. Association for Computing Machinery.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads. *PLOS ONE*, 11(3):1–29.