Personal and Ubiquitous Computing
https://doi.org/10.1007/s00779-022-01682-0

GUEST EDITORIAL

Special issue on intelligent systems for tackling online harms

Daniela Godoy[1] · Antonela Tommasel[1] · Arkaitz Zubiaga[2]

1 Introduction

Social media platforms have become an integral part of everyday life and activities of most people, providing new forms of communication and interaction. These sites allow users to freely share information and opinions (in the form of photos, short texts, and comments) while promoting the formation of links and social relationships (friendships, follower/followee relations). One of the most valuable features of social platforms is the potential to disseminate information widely and rapidly. However, the adoption of social media also exposes users to risks, giving rise to what has been referred to as online harms.

Online harms are also widespread on social media and can have serious damaging effects on individuals and society at large. Different forms of online harms include the distribution of false and misleading content (such as hoaxes, conspiracy theories, fake news, and even satirical content), harmful content such as abusive, discriminatory, offensive, and violence-inciting comments, or the augmentation of societal biases and inequalities online, among other things. The proliferation of harms online has become a serious problem with several negative consequences, ranging from public health issues to the disruption of democratic systems [1]. However, the identification of harmful content online has proven difficult, with the scientific community, social media platforms, and governments worldwide calling for support to develop effective methods.

✉ Daniela Godoy
daniela.godoy@isistan.unicen.edu.ar

Antonela Tommasel
antonela.tommasel@isistan.unicen.edu.ar

Arkaitz Zubiaga
a.zubiaga@qmul.ac.uk

[1] ISISTAN Research Institute, CONICET/UNCPBA, Campus Universitario, Tandil, Bs. As, Argentina

[2] Queen Mary University of London, London, UK

Online harm-aware mechanisms based on intelligent methods have become essential to mitigate the adverse effects of this unwanted content, preventing it from reaching large audiences and its amplification by social media. Although intelligent techniques in the intersection of machine learning, natural language processing, and social computing have made substantial advances in detecting and modelling the propagation of harmful content in social networks, there are several open problems in this area. Among others, concerns have been raised about the potential social biases and the lack of fairness of intelligent systems tackling the detection of online harms, which also stems from the lack of explainability and transparency of models.

The aim of this Special Issue was to bring together a community of researchers interested in tackling online harms and mitigating their impact on social media. The guest editorship of a special issue on this topic was largely motivated by the success and interest of two workshops on Online Misinformation- and Harm-Aware Recommender Systems (OHARS) [2, 3], co-located with ACM RecSys in 2020 and 2021. We called for novel research contributions on misinformation and harm-aware intelligent systems assisting users in making informed decisions in the context of online misinformation, hate speech, and other forms of online harms. We encouraged original contributions on intelligent systems that could circumvent the adverse effects of online harms in social media by improving detection methods, modelling their diffusion, and addressing concerns such as social biases, fairness, and explainability.

The present special issue attracted 19 submissions, of which seven were ultimately accepted for publication in the journal. A summary of the accepted submissions is given in the following section.

2 Accepted contributions

The first two contributions we describe in this section deal with abusive language and its detection.

Published online: 17 March 2022

Pamungkas et al. observe that abusive language in social media is not limited to specific languages, even though they are frequently addressed in a single language (mostly English). In addition, abusive language features multiple phenomena, such as hate speech, sexism, racism, among others, which are usually treated separately, thus ignoring their interconnections. In this context, the paper presents an extensive survey on abusive language detection covering the multi-lingual and multi-domain perspectives and the related challenges. First, the paper collects and analyzes the available datasets from existing studies on abusive language detection across different domains. Here, domain refers to both aim (e.g. hate speech, cyberbullying, and offensiveness) and platforms (e.g. Twitter, Facebook, and YouTube). Available datasets are compared based on their aim, sources, availability, annotation scheme, and data distribution. Then, the authors discuss the state-of-the-art in abusive language detection focusing on building robust models across different domains. Several works are compared based on the trained models (traditional machine learning, neural, or transformer-based), selected features, and adaptations to deal with domain shifts. Multilingual abusive language detection is also addressed. Again, an analysis of abusive language datasets is performed, in this case, for languages other than English, followed by a discussion on existing detection approaches focusing on methods to transfer knowledge between languages. Finally, challenges and opportunities in cross-domain and cross-lingual abusive language detection are identified.

Sarracén and Rosso propose an automatic keyword extraction (AKE) technique for extracting the most relevant words in texts from datasets composed of offensive and non-offensive tweets. To this end, keywords are defined as words that are relevant to identify offensive content. An unsupervised hybrid approach combining BERT multi-head self-attention and reasoning over a word graph is presented. The rationale behind the approach is that the attention mechanism allows capturing contextual relationships between words. Then, relationships are used to generate a graph where the most relevant words are identified using eigenvector centrality. The experimental evaluation was carried out using the datasets released for the OffensEval shared task for the 2019 and 2020 editions, focusing on identifying offensive language in tweets. The method's suitability to extract keywords for the particular purpose of identifying offensive content over other general-purpose AKE techniques was ultimately analyzed.

The following two contributions deal with misinformation and filter bubbles.

Roitero et al. present a study of crowdsourcing and its effectiveness for assessing the truthfulness of recent health statements related to the COVID-19 pandemic. Motivated by the lack of studies on crowd assessment of COVID-19-related misinformation and the recency of the statements as judged by the crowd, the authors focused on statements about COVID-19 to understand whether crowdsourcing might be adequate to a sensitive domain as health. Crowd workers were asked to assess the truthfulness of statements and provide evidence for the assessments. In addition to the accuracy of judging the truthfulness of statements, the authors also reported results on workers' behaviour, agreement between workers, effect of aggregation functions, scale transformations, and workers' background and bias. Results showed that workers could detect and objectively categorize online (mis)information related to COVID-19 and that both crowdsourced and expert judgments can be transformed and aggregated to improve quality. They also showed that workers' backgrounds can impact data quality. The longitudinal study performed by re-launching the task multiple times with both novice and experienced workers demonstrates that the time span significantly affects the judgment quality for both novice and experienced workers.

The filter bubble phenomenon in the context of news recommendation systems is addressed by Gharahighehi and Vens by proposing scenarios to make session-based recommender systems (SBRSs) diversity-aware. The authors present diversification approaches for four state-of-the-art neighbourhood-based SBRSs using news article metadata. More diverse recommendation lists are obtained with two methods that manipulate the candidate item or neighbour selection. The paper's main goal is to investigate the accuracy/diversity trade-off, i.e. quantifying the accuracy loss as a cost for the introduced diversity. The diversity boosting approaches were evaluated on four news datasets, showing that diversification can widen the filter bubbles around the users by recommending news from more diverse topics.

Finally, the remaining three contributions address the identification of bots and fake social media profiles as potential sources of malicious behaviours, as well as the protection of sensitive information.

Gilmary et al. addressed the problem of analyzing and categorizing Twitter accounts as bots or humans. Aiming to identify bots that deceive genuine users by sharing spurious content or distributing malware and improving safety, the article provides a novel approach for identifying bots by computing entropy on temporal information. Thus, bots are identified by determining the randomness and regularity present in the temporal tweet attribute of the user. In this approach, the real-time tweets posted by individual Twitter profiles are collected. The number of tweets posted by the user over a sampling period is extracted as an activity signal. Later, the degree of regularity present in the activity signals is measured through the lens of entropy. Approximate entropy and sample entropy are used to quantify the global degree of regularity in the signal. Finally, the paper investigates how strong is the association between entropy and the

class of Twitter profile (bot or human) through point-biserial correlation.

Elhussein considers that, while the research in fake account detection is mainly focused on automatic and semi-automatic accounts, the detection of fake accounts run by humans is often neglected. The paper aims to reveal the characteristics of accounts with fake identities on Facebook within the Sudanese community and contribute to their detection using machine learning and human assertion. As part of this research, 250 Sudanese people on Facebook who fell victim to eight of these accounts were interviewed to collect data, including both confirmed fake and real accounts. Machine learning classifiers were tested for the task of classifying fake identities. Interviewees also identified human-based methods, such as profile image verification and history, as ways to assert the legitimacy of an account.

Naseem et al. proposed a cognitive information protection agent using artificial general intelligence, Artificial General Intelligence-based Rational Behavior Detection Agent (AGI-RBDA), for protecting sensitive information. The proposed agent possesses human-like rationality for protecting sensitive information before sharing it with the other agents using cognitive correlates like a human mind. The authors claim that the human mind does not apply any encryption technique. Instead, it uses various cognitive factors such as intention, perception, motivation, emotions, implicit and explicit knowledge to secrecy sensitive information. Various cognitive factors are combined in the proposed agent to attain human-like information protection and manipulation capabilities. Experimental evaluation relied on a simulation of rational behaviour based on fuzzy logic.

Funding The organizers are partly supported by the CONICET–Royal Society International Exchange (IEC\R2\192019).

References

1. UK (2019) House of commons, digital, culture, media and sport committee, disinformation and 'fake news': final report (Eighth Report of Session 2017–19). Accessed 19 Feb 2020
2. Tommasel A, Godoy D, Zubiaga A (2020) Workshop on online misinformation- and harm-aware recommender systems. fourteenth acm conference on recommender systems. Association for Computing Machinery, New York, pp 638–639. https://doi.org/10.1145/3383313.3411537
3. Tommasel A, Godoy D, Zubiaga A (2021) OHARS: second workshop on online misinformation- and harm-aware recommender systems. Fifteenth ACM Conference on Recommender Systems. Association for Computing Machinery, New York, pp 789–791. https://doi.org/10.1145/3460231.3470941