2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)

Friend or Foe: Studying user trustworthiness for friend recommendation in the era of misinformation

Antonela Tommasel ISISTAN, CONICET-UNICEN. Argentina
antonela.tommasel@isistan.unicen.edu.ar

***Abstract*—The social Web, represented mainly by social media sites, is characterized by enriching the life and activities of its users, thus giving rise to new forms of communication and interaction. The unlimited possibilities offered by social media sites generate new problems related to information overload, the quality of published information and the formation of new social relationships. This opens the possibility to the contamination of social media with unwanted and unreliable content (false news, rumours, spam, hoaxes), which influences the perception and understanding of events, exposing users to risks. Motivated by the large amount of heterogeneous information available on the social Web and considering the consequences of the exposure to unwanted and unreliable content on social media, the existence of accounts dedicated to sharing said content, and the rapid dispersion of both phenomena, the goal of this work is is to define a profile to describe and estimate the trustworthiness or reputation of users, to avoid making "bad" recommendations that could favour the propagation of unreliable content and polluting users. The contribution of this work lies in the provision of reliable recommendation systems based on the integration of techniques that automatically allow the detection of unreliable content and the users publishing it. The final aim is to reduce the negative effects of the existence and propagation of such content, and thus improving the quality of the recommendations.**

I. Introduction and Motivation

The social web changed the ways in which individuals consume and produce information. Just as the real world can be a dangerous place, social media is no exception. Besides fostering social connections, social media also represent the ideal environment for undesirable phenomena, such as the dissemination of unwanted or unreliable content, and misinformation. This situation threatens the access to reliable and trustworthy information and, at the same time, the establishment of reliable social relations due to the proliferation of false or malicious accounts devoted to the dissemination of such information. Hence, although social media provides a great opportunity to learn about events and news, it also produces scepticism amongst users as relevant and accurate information coexist with unreliable and undesired information.

All the problems related to information overload and the quality of published information have led to the study of the causes of the viral distribution of information. In this way, the growing tendency to publish rumours and false information has motivated the development of systems to assess the reliability of information. However, the vulnerability of individuals and society to the manipulations made by malicious actors is still unknown [1, 5]. The increasing availability and popularity of social media, combined with the potential for automation and the low cost of producing fraudulent sites, allows the rapid creation and dissemination of misinformation, which overflows legitimate users with unreliable information [1], and influences public opinion, thus diminishing the value and quality of the social Web in the future and the user experience.

In this context, the recommendation of information similar to the already know, which is the basis of traditional recommender systems, may not be sufficient to ensure the reliability of recommendations. Thereby, new techniques are required for recommending information that is not only relevant to the users, but also comes from trustworthy sources. Trust analysis could help to mitigate those problems as it allows determining from whom to receive information, with whom to share it and which information to trust. New challenges arise regarding how to determine the trustworthiness of friends, content or locations to recommend, and, more importantly, how such trustworthiness can be integrated into existing recommender systems.

Motivated by the rapid dispersion and consequences of exposure to unwanted content on social networks, the existence of accounts dedicated to sharing such content, and the rapid dispersion of both phenomena, the **hypothesis** guiding this work is that the analysis of user and content credibility and trustworthiness can allow the development of more precise recommendation techniques that avoid the problems associated with social media pollution. Particularly, the **goal** is to define a profile to describe and estimate the trustworthiness or reputation of users, to avoid favouring the propagation of unreliable content and polluting users to be integrated into a recommender system, aiming at balancing both the relevance and reliability of recommendations. As a result, emphasis is also placed on defining techniques for detecting unreliable content and the users publishing it. Such techniques will serve as a basis for the construction of the trustworthiness user profile. The profile must dynamically detect mutations of user behaviour for adapting to new patterns of unreliable activities. Thereby, the **contribution** of this work lies *in the provision of reliable recommendation systems based on the integration of techniques that automatically allow the detection of unreliable content and the users who publish it with the aim of reducing the negative effects of the existence and dissemination of unreliable content in the social Web and, consequently, in the quality of the recommendations*.

II. Research Questions and Goals

This proposed work is based on the following research questions. First, considering the heterogeneous nature of social networks and the multiple sources of heterogeneous information available, what is their utility for the detection of unreliable content and malicious accounts. Second, how the detection of unreliable content and accounts can be integrated for the definition of a trustworthiness user profile in relation to their social and publication patterns. This analysis could also be useful to determine the susceptibility of users to unreliable content. Also, how to adapt said level of trustworthiness to changes in the undesired behaviours of said users or accounts

978-1-7281-1488-0/19/$31.00 ©2019 IEEE
DOI 10.1109/AIKE.2019.00053

(for example, changes in the patterns of publication or social interaction). Fourth, how to integrate the trustworthiness profile in a recommendation system that leverages on the characteristics and behaviour of users for personalising the recommendations. Similarly, how the information provided by different social platforms can complement the analysis. Fifth, to what extent the quality of recommendations can be improved if the evolution of the interests and behavioural patterns of users is also considered. In this context, three specific goals are derived from these questions:

- Analyse the interaction between the multiple and heterogeneous information sources in social media in relation to the behaviour and interests of users for detecting unreliable content, false accounts, bots or promoters of said content. This integration must cope with changes in user behaviour.
- Define a user profile to estimate the trustworthiness of users based on their behaviour and the identification of unreliable content and accounts.
- Integrate the developed trustworthiness profile in user recommendation algorithms.

III. Related Works

The rise of the social Web changed the ways in which individuals consume and produce information, giving rise to the uncontrolled and massive dissemination of misinformation, which threatens the access to reliable information and the establishment of trustworthy social relationships. Even though various misinformation detection techniques have been proposed in the literature, there are still some challenges to be solved. First, the identification of misinformation requires more than text analysis, hence multiple sources of information must be integrated. Second, the exploitation of such additional information can be challenging given the incompleteness, noise and volume of it. Second, bots and accounts spreading misinformation modify their behaviour patterns in an attempt to go unnoticed. In this sense, detection will be less effective if training data is not periodically updated. Third, techniques may be over-trained for a specific type of misinformation or spam campaign, limiting its applicability in broader domains. Likewise, each type of misinformation can present particular characteristics, which must be taken into account for detection. Fourth, since misinformation does not appear spontaneously, it is vital to analyse who published it, its intentions and processes, instead of analysing each piece of disinformation individually [5]. This analysis is also affected by the lack of integration of multiple sources of information, the updating of techniques and the disregard of the interrelation between different social platforms.

A. Unreliable Content and Users in Social Media

The ecology of false information and the dissemination of unwanted and unreliable content has evolved from the times of the written press, radio, television, to online social media. To mitigate the negative effects of unreliable content, the development of methods for automatically detecting it is essential. However, such detection is not simple. The existing detection approaches have been based mainly on one of three aspects: the textual content, the responses received and the identification of the content promoters, which might all be ambiguous. As regards content, techniques have focused on the analysis of linguistic patterns (e.g., the use of pronouns, conjunctions) in combination with traditional classifiers. Nonetheless, each type of unwanted content may have different textual indicators. In terms of responses, techniques have focused on content propagation in social media, thus requiring access to large amounts of data, which may not be feasible.

While many users on social networks are legitimate, other users may be malicious, and in some cases, not even human. The low cost of creating accounts in social media encourages the existence of unwanted accounts, such as bots, cyborgs, spammers or trolls. Then, the detection of the true identities of social media users is an essential step for the construction of safe and efficient communication channels in social networks. Traditionally, the detection of unreliable or malicious users is based on the same characteristics as the detection of unreliable content. For example, in [6], linguistic and metric characteristics related to the size of the social network of users and the diffusion of their publications were used for the detection of spammers.Note that existing techniques attempt to determine only if an account is a certain type of unwanted user, as opposed to estimating her/his level of trustworthiness.

Some works have addressed the analysis of the reliability of users in the context of the dissemination of unwanted content and promotion of spam. In [3], a user reliability metric was defined as a Page Rank variation that assesses the excessive consumption of information in terms of the production of information. Finally, in [4], a profile was created for each user as a means of authentication and identification of real user accounts. The profile includes information on social interactions, friendships and the topological similarity with these friendships, ignoring the information related to the content that these accounts may publish. The approach is based on the premise that accounts that promote the distribution of unwanted content have different behaviours and patterns of interaction than accounts that do not. The authors concluded that the computed level of reliability allowed them to correctly detect spam on social media, except in those cases where the spam came from apparently legitimate accounts. Note that these works do not perform an analysis of the particular characteristics of the content, nor consider the dynamism in behaviour or changes in the social environment.

B. Trustworthiness in Recommender Systems

Trust analysis can help to mitigate the proliferation of false accounts, spam and cyberaggression, and the problems of information overload, as it allows to determine who receives information and with whom is shared [8]. Consequently, the recommendation of users or items purely based on the homophily amongst them (either according to one or multiple factors) may be insufficient [7], giving rise to new challenges respect not only to the determination of the reliability of recommendations but also to the integration of said reliability in existing recommendation systems.

Deciding how reliable or trustworthy a user is involves the analysis of various factors, such as personal relationships, past experiences of a user with their friends and actions and opinions made in the past, amongst others. In the context of online social media, the analysis has traditionally only focused on behavioural aspects exhibited by active users [3]. These behaviours are generally expressed in the way information is produced and shared, as it is assumed that trustworthy users will generally publish useful content, while untrustworthy users will publish unwanted and even malicious content [3].

In this sense, little attention has been paid to the principle of unequal participation, which establishes that the largest proportion of content is created by a minority of users, while the rest of the individuals only observe the publications and discussions.

Although it is important to consider the trustworthiness of the elements to be recommended, it has been generally studied in the context of collaborative filtering to determine the reliability of users' ratings [7, 2], or to mitigate the cold-start problem [8]. Conversely, few studies have incorporated this concept in relation to the problems affecting social media, such as the proliferation of unwanted content or in the selection or ranking of users to recommend. Amongst them, in [2] they focused on assisting members of a community in making decisions regarding other members of the same community by defining an indicator of the reliability of said user. Reliability was computed as an expression of their explicit connections, the opinions that had been expressed about them (e.g., comments, positive votes "likes", negative votes "dislikes"), users' interests and the propagation of said reliability. That is, aspects specifically related to unwanted content were not considered and it requires explicit reputation indicators, which may not always be available, and even when available using the positive votes may not be effective considering the mimicking tactics of spammers. Moreover, the existing definitions of reliability applied do not consider the dynamism of the social environment.

IV. Research Method

The defined goals are materialised into two milestones. The first milestone marks the design of the trustworthiness profile, whilst the second marks the integration of the developed profile in a prototyped recommender system. Each planned activity will be evaluated individually considering data collections publicly available from *Instagram*, *Twitter*, *Flickr* and *Facebook*. Such collections include information regarding the social relations, and the published posts, tags and comments. Some of these collections were already used in the context of recommendation tasks in [14, 15, 10], community detection [12, 11] and the detection of aggressive content and users [17, 16]. Then, the developed techniques will be refined before their integration in a real social media environment based on comparisons with state-of-the-art techniques made in offline settings. In conjunction with the defined activities, periodic reviews of the state-of-the-art will be carried out. The partial results obtained during the execution of the described activities will be reported in conferences and journals.

A. Milestone #1: Design of Trustworthiness profile

This Milestone is divided into two main activities:

- Definition of the trustworthiness profile from the fusion of multiple information sources applying mining techniques and data inference about the set of activities carried out by the users (friendships, published content, interaction with other users, topics, etc.).
- Definition and empirical evaluation of the strategy for adapting the trustworthiness profile to changes in the environment and user behavioural patterns.

The measurement of trustworthiness has been an important topic of psychology and social sciences [3]. Trustworthiness depends a set of factors, such as the personal relationship and past experiences with said user or his friends, and opinions about actions that the user has taken in the past, amongst others [3]. In the context of social media, most studies have focused on behavioural aspects to distinguish between "good" and reliable (and even influential) and "bad" or unreliable users. In general, existing studies have considered the detection of unwanted and unreliable content and the users publishing it as two independent activities. However, the detection of unreliability content is closely linked to the detection of users promoting such content.

A common phenomenon is that accounts distributing unreliable content (that is, unreliable or malicious accounts) tend to publish more unreliable content than other accounts, and unreliable content is more likely to be shared by unreliable accounts. Consequently, the joint detection of unreliable content and the users publishing it can improve the results of independently performing such tasks. Also, due to the increase in the number of unreliable accounts that are daily created, the impact of malicious activities has drastically increased [5]. In this way, to assess the prevalence of unreliable content, it is important to take into account the intentions and the processes of the publishers, rather than the individual publications [5]. In this sense, the detection of unreliable users can take as a basis the characteristics associated with published content, the social network of users, and the social processes of content propagation. Existing works have been mainly oriented to the individual identification of one type of unreliable users (e.g. spammer, aggressive). However, no comprehensive approaches have not been yet proposed to simultaneously differentiate or classify users in any of the categories of the unwanted user taxonomy. In this way, it is necessary to develop alternatives that integrate the different detection models for estimating the degree of integral reliability of each of the users. In addition to seeking the combination of the various aspects already mentioned, it is necessary to monitor and update the selected features and models to changes in users and particularly, to the attempts to go unnoticed by unreliable users.

B. Milestone #2: Prototyped Recommender System

This Milestone is divided into three main activities:

- Definition of techniques for merging the trustworthiness profile with other sources of information available in social media. The contribution of each information source to the final recommendations should be tailored to the specific user needs and characteristics.
- Definition, empirical evaluation and refinement of a recommendation technique from the fusion techniques previously developed to obtain performance measures and parameter adjustment.
- Integration and evaluation of the trustworthy technique in user recommendation systems at the prototype level to collect data on its use by real users in controlled environments, following an A/B test technique.

The presented trustworthiness profile corresponds to the behavioural profile of users regarding their social and publication patterns. However, this characterization does not reflect the interests of users in terms of their friends or the information they consume. In this context, it is necessary to define a profile of interests of users. These interests could then allow the adaptation of the relevance of each of the characteristics of the trustworthiness profile, as previously studied in [10]. Note

that the integration of the profile can be useful not only for the selection of users to recommend, but also to the problem of "friends spam" to guide the user regarding what requests should or should not accept, according to the profile of the users requesting friendship.

The trustworthiness profile will be inserted in a user recommendation system (such as those in [10, 15]) to enrich other possible sources of information, such as personality or communities. In this sense, alternatives for the integration of the different sources of information should be studied that allow not only the adjustment of the relevance of each one of said factors, but also to the changes and evolution of the users over time. Finally, the process of recommending new users should try to find matches between the trustworthiness and interests profiles.

V. Expected Contributions

The contributions expected from this work are as follows:

- The provision of techniques for estimating the level of trustworthiness of users based on their behaviour and the detection of the unreliable content they published.
- The provision of novel recommendation techniques based not only on recommending relevant but also trustworthy items.
- The prevention of the creation or diffusion of such content and accounts due to the early detection of unreliable or malicious account or content, helping to reduce the contamination of social media.

VI. Previous Work and Research Agenda

The proposed work is founded on previous research that explored the dynamics of social networks in terms of their users, their particular interests and motivations as represented by their behaviour. First, it was studied the importance of personality and user behaviour in friend recommendation [14, 15]. These studies were motivated by the fact that existing approaches solely focus on either topological or content information, disregarding how the interests and decisions of users are influenced by their particular psychological characteristics. Second, a technique was proposed for adapting the friend selection criteria to the characteristics of previously selected friends [10]. The technique (based on both the relevance and novelty of recommendations) adapts the relevance of such criteria over time.

Previous research lead to an integral analysis of how to exploit the linked nature of social media by detecting explicit and implicit user relations in multiple social media sites, and how to integrate them for community detection [12, 11]. Then, a user recommendation technique was developed based on community discovery and textual feature selection. Such features were used for training learning models describing the particularities of each community. The developed technique provided capabilities to not only identify diverse reasons for choosing friends (represented by the different communities), but also to personalise and dynamically adapt the learned models to changes in users' interests.

Current research focuses on studying the needs and challenges of integrating multiple information sources for user recommendation. In this context, the challenges of detecting fake content, spam and threatening behaviours (such as aggression or cyberbullying) on social networks began to be addressed. Particularly, we have started to explore the detection of aggressive content (complemented with the detection of aggressors) in multiple social media [16, 17]. These works focused only on the analysis of the content published from the definition of characteristics at the level of character, word, information of feelings, irony identifiers and word embeddings, omitting the analysis of related social aspects. These studies can be complemented with a previous study in which users' style of writing was studied in relation to their personality and gender [13], and a study of the factors that would allow estimating the influence of users [9]. Then, present research aims detecting unreliable users to complement the studies on unreliable and unwanted content detection and, thus, build the proposed trustworthiness profile.

References

[1] K. S. Adewole, N. B. Anuar, A. Kamsin, K. D. Varathan, and S. A. Razak. Malicious accounts: Dark of the social networks. *J NETW COMPUT APPL*, 79:41 – 67, 2017. ISSN 1084-8045.

[2] M. Eirinaki, M. D. Louta, and I. Varlamis. A trust-aware system for personalized user recommendations in social networks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 44(4):409–421, April 2014. ISSN 2168-2216. doi: 10.1109/TSMC.2013.2263128.

[3] R. Interdonato and A. Tagarelli. To trust or not to trust lurkers?: Evaluation of lurking and trustworthiness in ranking problems. In A. Wierzbicki, U. Brandes, F. Schweitzer, and D. Pedreschi, editors, *Advances in Network Science*, pages 43–56, Cham, 2016. Springer International Publishing. ISBN 978-3-319-28361-6.

[4] S. Jeong, J. Lee, J. Park, and C. Kim. The social relation key: A new paradigm for security. *Information Systems*, 71:68 – 77, 2017. ISSN 0306-4379.

[5] D. Lazer, M. Baum, Y. Benkler, A. Berinsky, K. Greenhill, F. Menczer, M. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. Sloman, C. Sunstein, E. Thorson, D. Watts, and J. Zittrain. The science of fake news. *Science*, 359(6380):1094–1096, 2018. ISSN 0036-8075. doi: 10.1126/science.aao2998.

[6] S. Nilizadeh, F. Labréche, A. Sedighian, A. Zand, J. Fernandez, C. Kruegel, G. Stringhini, and G. Vigna. POISED: Spotting Twitter Spam Off the Beaten Paths. In *Proceedings of the 2017 ACM Conference on Computer and Communications Security*, NY, USA, Nov. 2017. ACM.

[7] J. O'Donovan and B. Smyth. Trust in recommender systems. In *Proceedings of the 10th International Conference on Intelligent User Interfaces*, IUI '05, pages 167–174, NY, USA, 2005. ACM. ISBN 1-58113-894-6.

[8] J. Tang and H. Liu. Trust in social media. *Synthesis Lectures on Information Security, Privacy, & Trust*, 10(1):1–129, 2015.

[9] A. Tommasel and D. Godoy. A novel metric for assessing user influence based on user behaviour. In *SocInf@ IJCAI*, pages 15–21, 2015.

[10] A. Tommasel and D. Godoy. Learning and adapting user criteria for recommending followees in social networks. *Journal of the Association for Information Science and Technology*, 2017.

[11] A. Tommasel and D. Godoy. Consensus community detection for multi-dimensional networks. In *Proceedings of SLIOIA in the XLIII CLEI*, Córdoba, Argentina, 2017.

[12] A. Tommasel and D. Godoy. Multi-view community detection with heterogeneous information from social media data. *Neurocomputing*, 289:195 – 219, 2018. ISSN 0925-2312.

[13] A. Tommasel, J. Balmaceda, D. Godoy, and S. Schiaffino. On the relationship of writing style, gender and personality in social texts: Myspace case study. In *Proceedings of the V LAWCC in the CLEI*, Naiguatá, Venezuela, 2013.

[14] A. Tommasel, A. Corbellini, D. Godoy, and S. Schiaffino. Exploring the role of personality traits in followee recommendation. *Online Information Review*, 39(6), 2015.

[15] A. Tommasel, A. Corbellini, D. Godoy, and S. Schiaffino. Personality-aware followee recommendation algorithms: An empirical analysis. *Eng. Appl. of AI*, 51:24–36, 2016.

[16] A. Tommasel, J. M. Rodriguez, and D. Godoy. Textual aggression detection through deep learning. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 177–187, Santa Fe, New Mexico, USA, Aug. 2018. ACL.

[17] A. Tommasel, J. Rodriguez, and D. Godoy. An experimental study on feature engineering and learning approaches for aggression detection in social media. *Inteligencia Artificial*, 22(63):81–100, Feb. 2019.