SMART: a Tool for Trust and Reputation Management in Social Media[*]

Nishant Saurabh[1][0000−0002−1926−4693][**], Manuel Herold[1], Hamid Mohammadi Fard[2][0000−0003−0562−3824],, and Radu Prodan[1][0000−0002−8247−5426]

[1] Institute of Information Technology, University of Klagenfurt, Austria
[2] Department of Computer Science, Technical University of Darmstadt, Germany

Abstract. Social media platforms are becoming increasingly popular and essential for next-generation connectivity. However, the emergence of social media also poses critical trust challenges due to the vast amount of created and propagated content. This paper proposes a data-driven tool called SMART for trust and reputation management based on community engagement and rescaled sigmoid model. SMART's integrated design adopts a set of expert systems with a unique inference logic for trust estimation to compute weighted trust ratings of social media content. SMART further utilizes the trust ratings to compute user reputation and represent them using a sigmoid curve that prevents infinite accumulation of reputation ratings by a user. We demonstrate the SMART tool prototype using a pilot social media application and highlight its user-friendly interfaces for trustworthy content exploration.

Keywords: Social media · trust · reputation · sigmoid model · community engagement.

1 Introduction

Social media platforms gained prominence as an essential technology for next-generation connectivity. Typically, social media are centralized platforms that allow users to create, publish, and share content across an interconnected network. This poses critical issues of trust [17] over the created content and the authentication of users who publish them. This is particularly problematic when fake news, trolling, and misinformation are a regular phenomenon across popular social media platforms such as Facebook and Twitter [8]. Moreover, the integration of privacy-by-design [18] features in social media platforms such as pseudonymized or anonymized identity systems that enable users to control their digital identity access aggravates this problem further. While such platforms improve upon privacy violations, they pose traceability challenges, for example, in identifying users publishing fake content.

[*] ARTICONF receives funding from the European Union's Horizon 2020 research and innovation program under grant agreement number 825134.

[**] Corresponding author: Nishant Saurabh, Distributed and Parallel Systems Group, Institute of Information Technology, University of Klagenfurt, Klagenfurt 9020, Austria. Email: `nishant.saurabh@aau.at`.

To prevent the propagation of malicious information in electronic networks requires innovative decision-making solutions at the user level (i.e., content creation, propagation, consumption) [1] and the underlying social media environment. The essential need is to explore trust and reputation management solutions [3] that involve social media users and allow them to be a part of decision-making. Such a process facilitates trustworthy and authenticated content creation and consumption and empowers users to tackle disinformation [4] and foster a positive engagement with fast-evolving technologies.

To achieve these goals, we propose a data-driven tool called SMART developed in the European ARTICONF [12] project, which provides a decision-making methodology engaging community experts [6] in computing weighted trust content ratings and classifying them as trustworthy or not. The trust ratings employ the rescaled sigmoid model [13] to compute the reputation ratings of a social media user who created them. Additionally, SMART associates each user with a contextualized local and global reputation, where the local rating reflects a user's trust for the created content within the same context. The global reputation, in contrast, provides the weighted trust ratings of a user across all contexts. Such a design allows SMART to provide fair and democratic decision-making for content trust management and prevents infinite accumulation of reputation by any user.

We developed a pilot social media application similar to Reddit[3] [2] called `SocialApp` to demonstrate SMART's trust and reputation management methodology. We highlight the current status of the SMART prototype and its interfaces, where a `SocialApp` user can perform trustworthy content exploration based on interesting topics, endorsements, and their time of creation.

The paper has five sections. Section 2 presents the SMART tool architecture and trust and reputation methodology. Section 3 demonstrates the SMART tool prototype and its interfaces using the sample social media application `SocialApp`. Section 4 briefly discusses the related works and industry-based trust and reputation management systems. Section 5 concludes the paper.

2 SMART Architecture

Figure 1 describes the SMART architectural workflow for trust and reputation management through a pseudonymized user who creates and publishes several posts in the `science and technology` community. Furthermore, SMART provides a list of trust oracles to the community members, representing expert systems with a unique knowledge base and an inference logic to compute the content trust ratings. The community members can choose one or more trust oracles by consensus to compute intermediary trust values for each content based on a particular inference logic. Afterward, SMART computes the weighted average of the trust ratings obtained from each oracle and labels the trustworthy or fake content. Finally, SMART aggregates the intermediary trust values of all posts created by the user and generates its reputation.

[3] `https://www.reddit.com/r/socialmedia/`

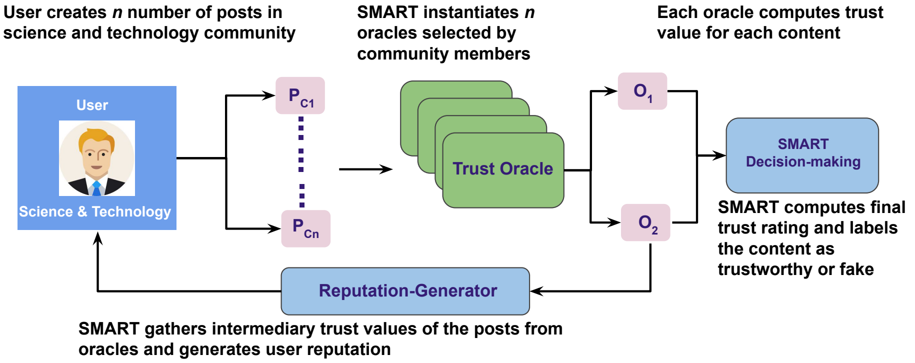

User creates *n* number of posts in science and technology community
SMART instantiates *n* oracles selected by community members
Each oracle computes trust value for each content
User
Science & Technology
P C1
P Cn
Trust Oracle
O 1
O 2
SMART Decision-making
SMART computes final trust rating and labels the content as trustworthy or fake
Reputation-Generator
SMART gathers intermediary trust values of the posts from oracles and generates user reputation

Fig. 1: SMART architectural workflow for trust and reputation management.

2.1 Trust Oracle

SMART computes the trust ratings and their content using a set of oracles with their own unique inference logic. SMART currently supports two types of trust oracles by design and plans to integrate several others in the future (e.g., online fact-checker tools).

Community voting based oracle O_1 utilizes the percentage of upvotes gathered by a post P_{Ci} in a community C to compute its trust rating rescaled between [-1,1] as follows:

$$O_1(P_{Ci}) = 2 \cdot \frac{Upvotes\,(P_{Ci})}{Votes\,(P_{Ci})} - 1, \tag{1}$$

where $Upvotes\,(P_{Ci})$ and $Votes\,(P_{Ci})$ are the number of endorsements and total votes of the post P_{Ci}.

ML classification based oracles O_2 represent binary machine learning models that classify a post P_{Ci} as trustworthy $(O_2\,(P_{Ci}) = 1)$ or fake $(O_2\,(P_{Ci}) = 0)$:

Trust $T(P_{Ci})$ computed by SMART decision-maker represents the aggregated normalized trust ratings of each oracle for a post P_{Ci} in a community C:

$$T\,(P_{Ci}) = \frac{O_1\,(P_{Ci}) + O_2\,(P_{Ci})}{2}. \tag{2}$$

The average trust computation is easily extensible to more oracles. A positive trust indicates trustworthy content, while a negative value suggests the opposite.

2.2 Reputation Generator

The SMART reputation generator computes the reputation rating of a user and classifies it as trustworthy and not. We define two types of reputation ratings.

Local reputation rating represents the trustworthiness of a user in a community C.

Global reputation rating reflects the accumulated trust of a user across all the communities of a social application.

The reputation generator follows three stages to compute the local and global reputation of a user.

Intermediary reputation RI_C is the first stage that initially gathers the final trust ratings $T(P_{Ci})$ (computed using Equation 2) of all posts P_{Ci} of a user in a community C. Essentially, each post created by the user varies in quality and trust and contributes to the intermediary reputation differently. Hence, we utilize content volume $V(P_{Ci})$ (measured in the number of characters) to distinguish the quality of different posts, assuming that a larger and more detailed content has a higher contribution to the user reputation:

$$RI_C = \frac{\sum_i T\left(P_{Ci}\right) \cdot V\left(P_{Ci}\right) \cdot \delta\left(P_{Ci}\right)}{\alpha_C}, \tag{3}$$

where α_C is the maximum content volume threshold V of a post in a community C, and $\delta\left(P_{Ci}\right)$ represents a weighted bias that rewards trustworthy and penalises fake posts using two weights p and r, respectively:

$$\delta\left(P_{Ci}\right) = \begin{cases} p, & T\left(P_{Ci}\right) < 0; \\ r, & T\left(P_{Ci}\right) > 0; \\ 0, & T\left(P_{Ci}\right) = 0. \end{cases} \tag{4}$$

We use $p = -2$ to penalize the fake posts and $r = 1$ to reward trusted ones in the current implementation. However, our design allows the community members to freely decide the reward and penalty weights based on consensus.

Local reputation RL_C of a user in a community combines the intermediary reputation rating RI_C with a rescaled sigmoid [13, 19] function. We use the sigmoid function due to its capability to model natural growth and decay rate in the non-deterministic environment such as social media platforms and compute the local reputation of a user as follows:

$$RL_C = \frac{2}{1 + e^{-RI'_C}} - 1, \tag{5}$$

where $R'_I \in [-\gamma, \gamma]$ is the *reputation growth and decay constraint* that prevents infinite accumulation of trust:

$$RI'_C = \begin{cases} -\gamma, & RI_C < -\gamma; \\ RI_C, & RI_C \in [-\gamma, \gamma]; \\ \gamma, & RI_C > \gamma. \end{cases} \tag{6}$$

We utilize a *reputation threshold* β decided by community members with consensus to classify a user into three categories:

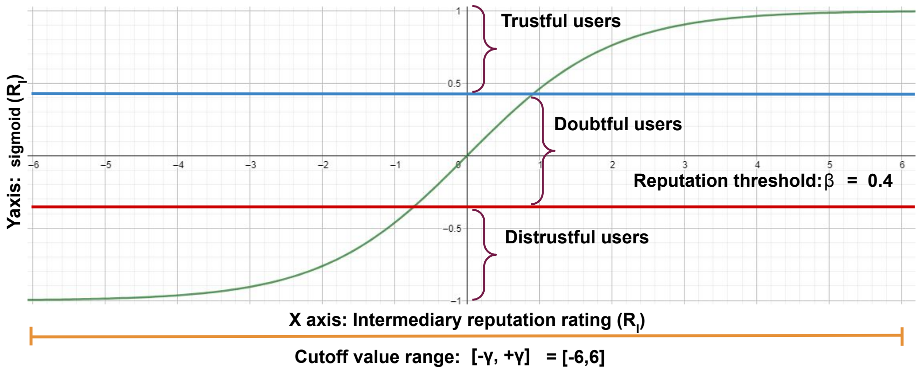


Fig. 2: Sigmoid representation of user reputation.

Trustful with a high positive local reputation: $RL_C > \beta$;
Distrustful with a low negative local reputation: $RL_C < -\beta$;
Doubtful with a local reputation in the range $RL_C \in [-\beta, \beta]$.

Figure 2 illustrates a sigmoid curve initialized with a reputation threshold $\beta = 0.4$ and a reputation growth and decay range $\gamma \in [-6, 6]$. We observe that a trustful user has a local reputation rating $RL_C > 0.4$, while $RL_C < -0.4$ classifies a user as distrustful. Additionally, we observe that the reputation growth and decay constraint γ prevents infinite accumulation of reputation by a user and instead limits a finite range of values.

Global reputation RG of a user averages the local reputations RL_C across all communities C weighted by the volume of the total posts in each community:

$$RG = \frac{\sum_{\forall C} V_C \cdot RL_C}{\sum_{\forall C} V_C}, \tag{7}$$

where $V_C = \sum_i V(P_{Ci})$ is the total content volume of all posts P_{Ci} published by a user in a community C.

3 Implementation

We developed a social media application similar to Reddit named `SocialApp` to pilot our research and development. Figure 3 shows a sample instance of the `SocialApp` application with two communities labeled `science and technology`, and `international politics`. A `SocialApp` user can join one or more communities based on topics of interest. For example, the users in the `international politics` community discuss ongoing affairs and the latest news across the world. In contrast, the `science and technology` community users create research and innovation-related content.

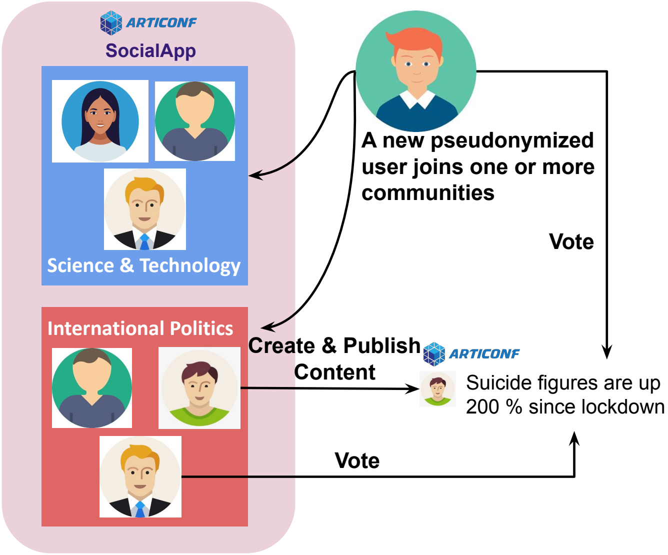


Fig. 3: `SocialApp` pilot use case application.

A `SocialApp` user can create and publish content in the form of text or multimedia. `SocialApp` also allows users to vote for content in their own community, reflecting their opinion about the content authenticity and quality. In its current form, `SocialApp` offers three basic functionalities to pseudonymized users as any other generic social media platform:

Post functionality allows a `SocialApp` user to create and publish content in its own community. `SocialApp` does not allow a user to post content in a community without joining it.

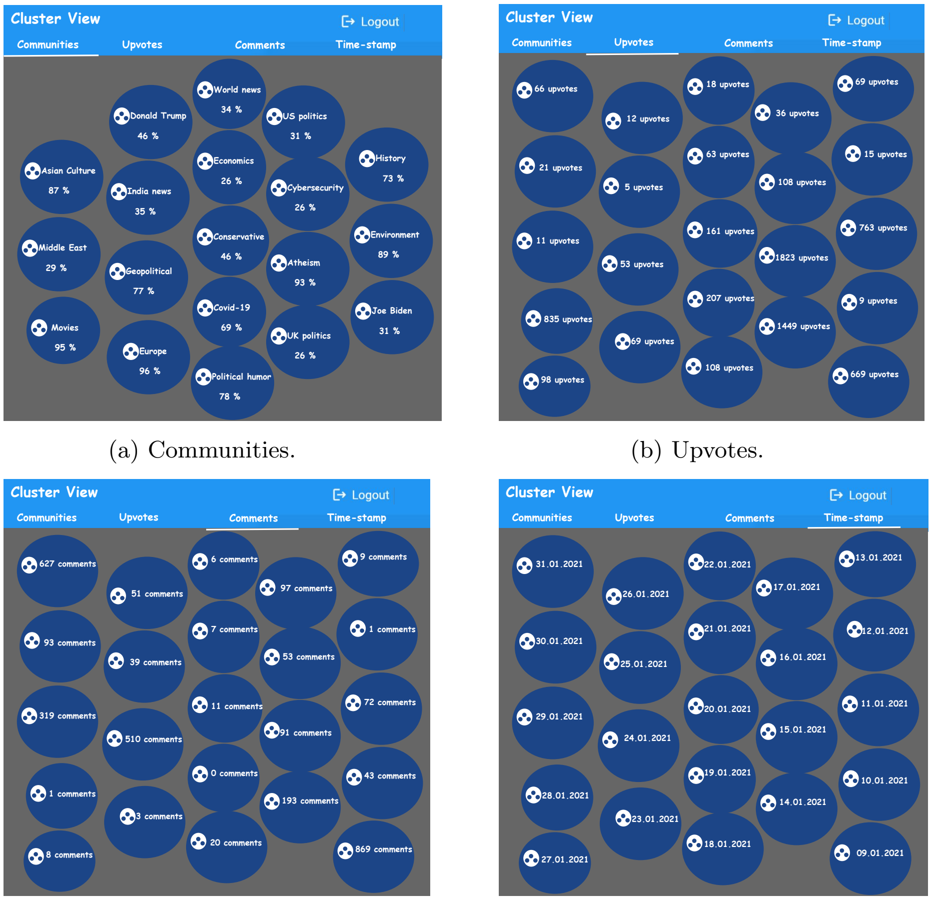
Vote functionality allows a `SocialApp` user to either upvote or to downvote a published post across the associated community. Similar to post functionality, a user cannot vote a content without joining it.

Comment functionality allows a `SocialApp` user to comment on a post either in the form of text or multimedia.

Each post in `SocialApp` has a data schema consisting of ten fields: the unique identifier, pseudonymized user identifier, community label, title, content, timestamp, comments, as well as the number of votes, endorsements, and dislikes. To demonstrate the SMART prototype, we integrated the Mockaroo[4] random data generator into the `SocialApp` interface. Mockaroo enables the creation of realistic test data in CSV, JSON, and SQL formats, which we used to generate 2000 users and 12 000 posts according to this schema.

Figure 4 shows the SMART cluster visualization with four interfaces:

[4] `https://www.mockaroo.com/`



(a) Communities.

(b) Upvotes.

(c) Comments.

(d) Timestamp.

Fig. 4: SMART cluster visualization snapshots.

Communities interface provides an aggregated view of the posts based on the community labels and identifiers. Figure 4a shows the clustered visualization of 12 000 `SocialApp` posts across 19 communities with unique labels. Each community label represents the context and type of the social media posts created by its members based on their interest topics. Social media users who are not members of a specific community cannot post content.

Upvotes interface shows the clustered posts based on the number of endorsements and dislikes by `SocialApp` users in their respective communities. Figure 4b shows 12 000 posts clustered across categories with upvotes ranging between 1893 and 5. The upvotes interface enables users to find the posts with the most positive reviews and contemplate if they match the content trust ratings. Such an interface aggregates the most endorsed `SocialApp` posts and promotes trustworthy content propagation.

Comments interface depicts the aggregated posts based on their number of comments, reflecting the general interest across community members. Figure 4c shows the clustered visualization of `SocialApp` posts based on the number of comments, ranging between zero and 800. This visualization allows users to obtain awareness of the trending posts and topics of discussion, generating higher interest.

Timestamp interface shows the clustered posts based on their creation date and time across different communities. Figure 4d shows the `SocialApp` posts clustered with different timestamp across 23 days. This clustered view allows `SocialApp` users to understand the timeliness of the content contained within each post. Additionally, this visualization indicates up-to-date or expired content and focuses on recent events.

A `SocialApp` user can click any cluster in these interfaces and explore different posts, content, and comments. Figure 5a shows the example of a 44% trustworthy post published in the `conservative` community by a pseudonymized user `itsanoobsgame`. This enables `SocialApp` users to check the trust ratings of a post and track the user who created the post and its corresponding community.

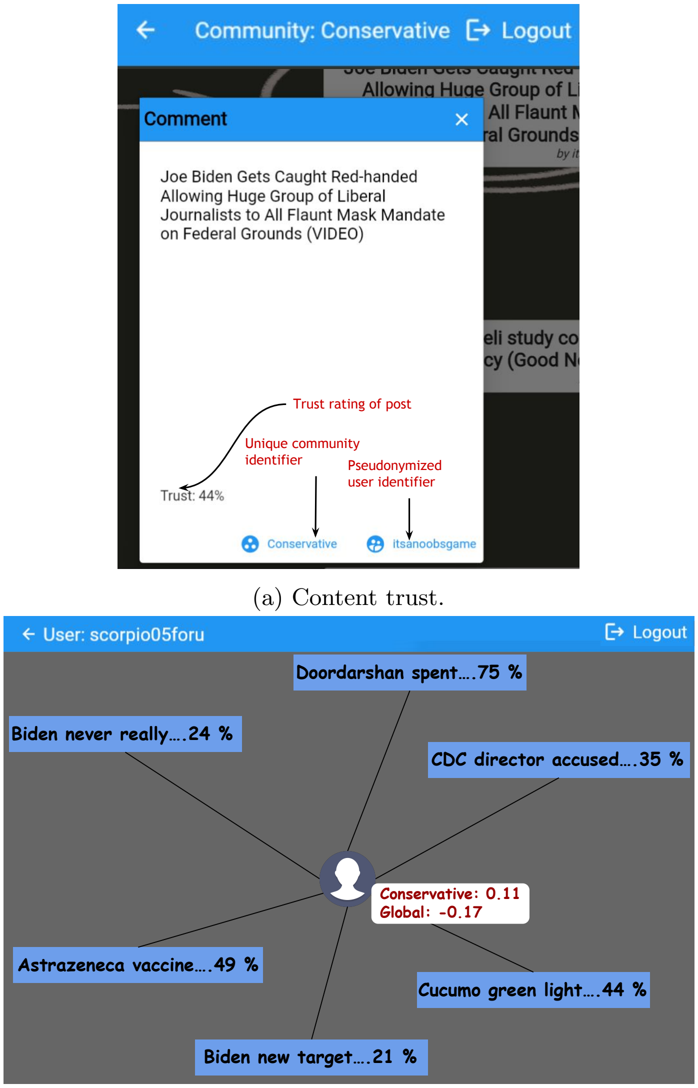
SMART also links each content to the user who published the post and their local and global reputation ratings. Figure 5b shows the snapshot of a pseudonymized `SocialApp` user `scorpio05foru` with a local reputation rating 0.11 in community `conservative` and a global reputation -0.17 across all joined communities. Additionally, SMART links other posts created by the same user along with their trust ratings. This allows a `SocialApp` user to get a historical overview of the content quality created and published by the user.

4 Related Work

Trust and reputation management is an extensively studied problem across many disciplines, including sociology [11], psychology [5], economics [7], and computer science [9,16]. Each discipline defined trust from different perspectives that may not fit into the diversified and digital social networks. In this section, we briefly describe some of the trust models across academia and industry.

Marsh et al. [10] proposed one of the earliest theoretical models for computational trust classified in three categories: basic, general, and situational. They characterize collaboration in digital networks, where a user who tends to trust others yields a higher reputation. Similarly, Sebater et al. [15] classify trust across four dimensions: the information source and the granularity that reflect the type and context of content for trust computation, the behavioral assumption that identifies manipulative activities by a social media user for trust enhancement, and the reliability that refers to the accuracy of the trust model.

In the industrial sphere, the eBay trust model is quite popular across online marketplaces. Online marketplaces such as Amazon use the eBay trust model to rate users and publicly reflect the historical users' activity and behavior in an online digital network public [14, 15]. The eBay computational trust model



Joe Biden Gets Caught Red-handed Allowing Huge Group of Liberal Journalists to All Flaunt Mask Mandate on Federal Grounds (VIDEO)

(a) Content trust.

User: scorpio05foru
Logout
Doordarshan spent....75 %
Biden never really....24 %
CDC director accused....35 %
Conservative: 0.11
Global: -0.17
Astrazeneca vaccine....49 %
Cucumo green light....44 %
Biden new target....21 %

(b) User local and global reputation.

Fig. 5: SMART trust and reputation visualization snapshots.

accumulates the positive, negative or neutral rating of other users over a period of six months. There are several potential problems with the eBay trust model. Firstly, the user reputation ratings are unbound and allow infinite accumulation of trust. As a consequence, new users find it difficult to compete with existing highly reputed ones. This also allows malicious users to accumulate high reputations by first performing trustworthy activities and scamming afterward. Hence, a trust and reputation management system requires time sensitivity and prioritizes recent activities. Sporas [20] trust model is similar to eBay but only considers the last user activity instead of accumulating the trust and reputation ratings over six months. However, Sporas also allows infinite accumulation of reputation by a user, similar to the eBay model.

5 Conclusion

Mitigating misinformation concerns and provisioning trustworthy content creation and propagation is essential for realizing next-generation social media. We propose in this paper a data-driven tool called SMART developed in the ARTICONF project that implements a trust and reputation management system based on community engagement and rescaled sigmoid model. We presented the SMART decision-making methodology that engages community experts in computing trust and reputation ratings of social media content and users and classifies them as trustworthy or not. We demonstrated the SMART trust and reputation management prototype using a generic social media application called `SocialApp` similar to Reddit, with user-friendly interfaces and trustworthy content exploration. In the future, we plan to integrate online fact-checkers to the SMART tool to improve fairness across computation of trust and reputation ratings. We also aim to validate its trust and reputation management for content co-creation, news marketplace, and other real industrial applications.

Acknowledgement

This work received funding from the European Union's Horizon 2020 research and innovation programme, grant agreement 825134, "Smart Social Media Ecosytstem in a Blockchain Federated Environment (ARTICONF)".

References

1. Al Qundus, J., Paschke, A.: Investigating the effect of attributes on user trust in social media. In: International conference on database and expert systems applications. pp. 278–288. Springer. (2018)
2. Buntain, C., Golbeck, J.: Identifying social roles in reddit using network structure. In: 23rd International Conference on World Wide Web. pp. 615–620 (2014)
3. Chen, B.C., Guo, J., Tseng, B., Yang, J.: User reputation in a comment rating environment. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 159–167 (2011)
4. Cohen, R., Moffatt, K., Ghenai, A., Yang, A., Corwin, M., Lin, G., Zhao, R., Ji, Y., Parmentier, A., P'ng, J., et al.: Addressing misinformation in online social networks: Diverse platforms and the potential of multiagent trust modeling. Information **11**(11), 539 (2020)
5. Cook, K.S., Yamagishi, T., Cheshire, C., Cooper, R., Matsuda, M., Mashima, R.: Trust building via risk taking: A cross-societal experiment. Social psychology quarterly **68**(2), 121–142 (2005)

6. Habibi, M.R., Laroche, M., Richard, M.O.: The roles of brand community and community engagement in building brand trust on social media. Computers in human behavior **37**, 152–161 (2014)
7. Huang, F.: Building social trust: A human-capital approach. Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft pp. 552–573 (2007)
8. Kim, Y.A., Ahmad, M.A.: Trust, distrust and lack of confidence of users in online social media-sharing communities. Knowledge-Based Systems **37**, 438–450 (2013)
9. Maheswaran, M., Tang, H.C., Ghunaim, A.: Towards a gravity-based trust model for social networking systems. In: 27th International Conference on Distributed Computing Systems Workshops (ICDCSW'07). pp. 24–24. IEEE. (2007)
10. Marsh, S.P.: Formalising trust as a computational concept (1994)
11. Möllering, G.: The nature of trust: From georg simmel to a theory of expectation, interpretation and suspension. Sociology **35**(2), 403–420 (2001)
12. Prodan, R., Saurabh, N., Zhao, Z., Orton-Johnson, K., Chakravorty, A., Karadimce, A., Ulisses, A.: ARTICONF: Towards a smart social media ecosystem in a blockchain federated environment. In: Euro-Par 2019: Euro-Par 2019: Parallel Processing Workshops. LNCS, vol. 11997., pp. 417–428. Springer (May 2020)
13. Ren, J., McIsaac, K.A., Patel, R.V., Peters, T.M.: A potential field model using generalized sigmoid functions. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **37**(2), 477–484 (2007)
14. Resnick, P., Zeckhauser, R.: Trust among strangers in internet transactions: Empirical analysis of ebay's reputation system. In: The Economics of the Internet and E-commerce. Emerald Group Publishing Limited. (2002)
15. Sabater, J., Sierra, C.: Review on computational trust and reputation models. Artificial intelligence review **24**(1), 33–60 (2005)
16. Sikder, O., Smith, R.E., Vivo, P., Livan, G.: A minimalistic model of bias, polarization and misinformation in social networks. Scientific reports **10**(1), 1–11 (2020)
17. Tang, J., Liu, H.: Trust in social media. Synthesis Lectures on Information Security, Privacy, & Trust **10**(1), 1–129 (2015)
18. Wahlstrom, K., Ul-haq, A., Burmeister, O., et al.: Privacy by design. Australasian Journal of Information Systems **24** (2020)
19. Weisstein, E.W.: Sigmoid function. https://mathworld. wolfram. com/ (2002)
20. Yu, W.: Analysis on trust influencing factors and trust model from multiple perspectives of online auction. Open Physics **15**(1), 613–619 (2017)