# Fact Checking Misinformation Using Recommendations from Emotional Pedagogical Agents

Ricky J. Sethi[0000−0001−5254−3750], Raghuram Rangaraju, and Bryce Shurts

Fitchburg State University, Fitchburg MA 01420, USA rickys@sethi.org

**Abstract.** Dealing with complex and controversial topics like the spread of misinformation is a salient aspect of our lives. In this paper, we present initial work towards developing a recommendation system that uses crowd-sourced social argumentation with pedagogical agents to help combat misinformation. We model users' emotional associations on such topics and inform the pedagogical agents using a recommendation system based on both the users' emotional profiles and the semantic content from the argumentation graph. This approach can be utilized in either formal or informal learning settings, using threaded discussions or social networking virtual communities.

## 1 Introduction

Dealing with *complex and controversial* questions has always been a salient aspect of our lives; questions like "Do we use only 10% of our brain?" or "Were the crowds truly larger at Donald Trump's inauguration than at Barack Obama's?" Ironically, such *emotionally* charged topics often elicit the **backfire effect**, where people's opinions harden in the face of facts to the contrary [8].

In this paper, we present initial work towards developing a recommendation system that uses crowd-sourced social argumentation with pedagogical agents

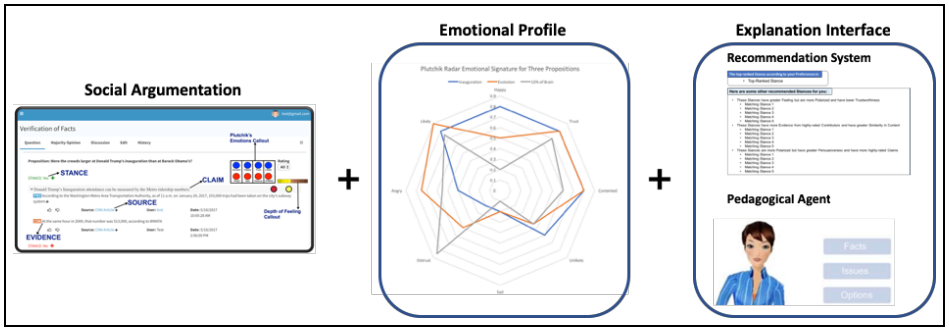


**Fig. 1.** System Overview: Social Collaborative Argumentation + Emotional Profiles + [Multi-Attribute Utility Theory Recommendation System + Pedagogical Agents]

(PAs) to help engender an analytical, evidence-based approach for combating misinformation. Our goal is to help learners *think critically* about such topics by using social collaborative argumentation supported by PAs that are informed by a recommendation system based on user' emotional and semantic profiles. An overview of our approach is shown in Figure 1.

### 1.1 Background

Our goal is to help people more effectively explore and understand their possibly subconscious biases in an effort to overcome the backfire effect and formulate more varied insights into complex topics. We utilize a *structured learning environment*, consisting of a **virtual community** that supports **social collaborative argumentation**, to build an argumentation graph which captures the semantic content of the propositions associated with such topics [9, 10].

We then address the role of emotion in reasoning by modeling the **emotional profiles** of users and contributors. We use both self-reported emotions and natural language processing with sentiment analysis from an **explanation interface** to create emotional profiles for users [8].

We now extend this approach to incorporate **Pedagogical Agents** (PA) to aid users' learning and gauge the impact of different PAs exhibiting varying degrees of emotion and knowledge. The PAs are informed by a **recommendation system** that fuses both the emotional profiles of users and the semantic content from the argumentation graph. We can use this emotional assessment to also gauge the extent of the backfire effect and the change in critical thinking as well as the ability of users to monitor and regulate their self-regulated learning processes by considering different types of information, evidence, and social influence delivered by the PAs.

This approach can have broad applicability for improving *online classroom learning* that utilizes threaded discussions; for *facilitating decision-making* amongst domain experts; and for *creating an informed electorate* that can assess the trustworthiness of information and information sources and lessen the risks of untrustworthy information like "alternative facts" and "fake news." It can also be used to *alter existing social networking sites* like Facebook to leverage their current userbase and aid in mitigating destructive online behaviour like spreading misinformation.

The three parts of our overall approach, shown in Figure 1, consist of the:

1. Social Collaborative Argumentation and Virtual Community
   Our social collaborative argumentation framework allows arguments to be expressed in a graph structure with input to be provided by a crowd that is mediated by experts. The fundamental idea is to incorporate content, ratings, authority, trust, and other properties in a graph-based framework combined with a recommendation system which explains the tradeoff of various competing claims based on those attributes. [9, 10, 12, 11]
2. Emotional and Semantic Profiles
   We model users' emotional associations on complex, controversial topics as

well as create a proposition profile, based on the semantic and collaborative content of propositions. Our framework combines emotional profiles of users for each proposition along with the semantic proposition profile. [9, 8]
3. Explanation Interface Recommendation System and Pedagogical Agents Interaction, as developed next.

## 2 Recommender System Explaining Interface

Long, complex arguments have a tendency to overwhelm users and leave them unable to decide upon which Stance is best supported for complex or controversial topics [2]. Recommendation systems can help target content to aid users' decision making and come in many varieties [3, 14].

Although there are advantages to these models and other explanation interfaces [14], including case-based and knowledge-based, combating misinformation requires the ability to not just look deeply but to look laterally and be able to account for multiple attributes [15]. Multi-Attribute Utility Theory (MAUT) [7] based explanation systems have been shown to do exactly this by increasing trustworthiness and persuasiveness in users' decision-making [5, 4].

Since our goal is to also increase the trustworthiness and persuasiveness of examining information online, we use the MAUT-based approach to create a novel explanation interface that can help users reason about controversial or nuanced topics comprehensively, including analyzing the semantic and emotional content of a complex argument in order to overcome both cognitive and emotional biases that contribute to echo chambers and the backfire effect. As such, not only does the system need to address the trust and authority of the information and its sources but the Viewer needs to be able to assess these components independently, as well. [9, 10]

We therefore create a recommender system that can recommend different Stances for a Proposition depending on the emotional and cognitive content of the collaborative argument. We anticipate that most Propositions will represent complex, nuanced Topics and so will have a number of Stances in general. The system will thus suggest supported Stances based on weights from the Viewer.

We do so by forming both Semantic and Emotional Profiles. Analogous to the standard recommendation models, we map {Products} $\rightarrow$ {Stances} and {Attributes} $\rightarrow$ {Claims, Evidence, Sources, Contributors}. Wherein the traditional recommender system utilizes the content of {Reviews}, we consider the content of {Claim, Evidence} nodes in the argumentation graph, $G_A$.

We also conduct a sentiment analysis on these Claims and Evidence nodes to create a Proposition Sentiment Profile. Using Viewer ratings of the emotional and depth of feeling callouts, we construct a Viewer Emotional Profile as well as a Proposition Emotional Profile. [9, 8] We further construct a Proposition Semantic Profile by using a Text Analysis approach for the semantic content and combining it with the collaborative cognitive content as well as the weights along the various dimensions of Contributor ratings, trust, and authority encapsulated

in the edges $e \in E$ of the $G_A$. Finally, we create the following utility model as per MAUT [7]:

$$U_v(S) = \sum_{i=1}^{m} w_i \cdot [\alpha \cdot V_i(S) + (1-\alpha) \cdot O_i(S)] + \sum_{j=m+1}^{n} w_j \cdot O_j(S) \tag{1}$$

where $U_v(S)$ is the utility of a Stance, $S$, for a Viewer, $v$, in terms of the Viewer's preferences. This contains an attribute model value, $V_i$, which denotes the Viewer's preference for each attribute, $a_i$, as well as a sentiment model value based on opinion mining, or sentiment analysis, $O$, as detailed next. The Sentiment Stance Model is expressed as:

$$O_i(S) = \frac{1}{|R(a_i, S)|} \sum_{r \in R(a_i, S)} O_i(r) \tag{2}$$

where $R(a_i, S)$ is the set of Claims and Evidence, analogous to Reviews in traditional recommender systems, with respect to a Stance, $S$, that contain the opinions extracted on attributes, $a_i$, which consist of Claims, Evidence, Sources, and Contributors. The sentiment for each review component, $O_i(r)$, is defined as:

$$O_i(r) = \frac{\sum_{e \in E(a_i, r)} polarity(e)^2}{\sum_{e \in E(a_i, r)} polarity(e)} \tag{3}$$

where $E$ is the set of sentiment elements associated with all the features mapped to an attribute $a_i$ in a review component $r$ and $polarity(e)$ is the polarity value for the element $e$ as derived using standard statistical sentiment analysis [6].

The utility model, $U_v(S)$, establishes a Viewer's preferences using a standard weighted additive form of the value functions [4, 5, 7]. This model allows us to calculate the tradeoffs in order to explicitly resolve a Viewer's preferences. We calculate this tradeoff among the top $k$ recommended Stance candidates; these are limited to $min[5, |S|]$, where 5 is the optimum number of options in a category less than 6 as discovered by [4], and $|S|$ is the number of returned Stances. We then run the Apriori algorithm [1] over all candidates' tradeoff vectors in order to calculate the Categories for Stances, as motivated by [4, 5] .

All Stances with the same subset of tradeoff pairs are grouped together into one Category; these tradeoff vectors take the form of sentiment and feature combined into a phrase like, "more Evidence", "more Polarized", etc. The Categories, in turn, use the tradeoff vectors as their descriptors so that we end up with Category Titles like, "This Stance has more Evidence from highly-rated Contributors and have greater similarity in terms of Content." The category title is thus the explanation that shows the advantages and disadvantages of the Stances. A mockup of how this would appear is shown in Figure 2.

Finally, we want category titles that are different to maximize how informative they are . As such, we also map the Diversity, $D$, of each Category, $c$, in the

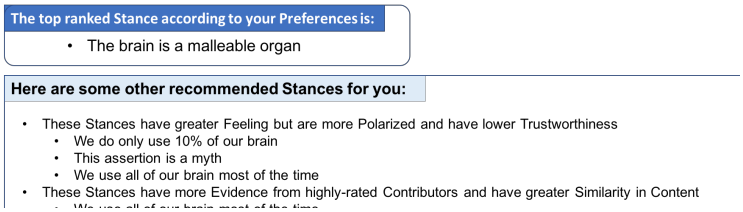


**Fig. 2.** A sample of the recommendations for Stances organized according to Category Titles made up of tradeoff vectors. Tradeoff vectors are of the type, "more Evidence", "lower Trustworthiness", etc., while Category Titles are of the form, "These Stances have greater Feeling but are more Polarized and have lower Trustworthiness".

set of Categories, $C$, in terms of both the Category Title, which is simply the set of tradeoff vectors, and the set of Stances in $c$, $S(c)$, as:

$$D(c, S(c)) = \min_{c_i \in C}[(1 - \frac{c \cap c_i}{|c|}) \times (1 - \frac{S(c) \cap S(c_i)}{|S(c)|})] \tag{4}$$

In the last step, the Viewer can update their preferences by using a button to offer better matching Stances. By also incorporating our structured discussion metrics [13], this framework can be applied to everything from analyzing misinformation to structuring discussions in online courses to ensure the information is trustworthy and the information sources assessable via the Category Titles.

## 3 Pedagogical Agents (PAs)

This explanatory recommendation infrastructure also utilizes PAs which will guide the user's experience. For example, suppose the topic a user wants to analyze is the crowd size at the 2017 inauguration. Viewers can examine the social argument with the aid of a PA. They can choose the kind of PA with whom to interact, like Friendly Republican or Objective Democrat or Angry Independent, etc.

In our approach, an argument is composed of Stances, Claims, Evidence, and Sources. This crowd-sourced argument is built out by the contributors to our system. [9, 8] Once the argument is constructed, Viewers can interact with the argument and the PA. This PA can then guide the Viewer through the examination of the argument using the biases captured in the PA and help critically analyze the topic. Viewers can change the PAs in their preferences or they can change them depending on affective data, either implicit or explicit, as collected by our framework.

In particular, the PAs can help viewers navigate the claims and especially those that might contradict their initial stance on a topic by helping them evaluate both the evidence and feelings for alternative claims. Since the PA will be built upon the Multi Attribute Utility Theory, it will be able to provide the explanation for why the alternative claims are presented from both a cognitive and affective perspective.

**Acknowledgments** We would like to gratefully acknowledge support from the Amazon AWS Research Grant program. We would also like to thank Roger Azevedo for the valuable discussions and support.

## References

1. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: Acm sigmod record. vol. 22, pp. 207–216. ACM (1993)
2. Baram-Tsabari, A., Sethi, R.J., Bry, L., Yarden, A.: Asking scientists: A decade of questions analyzed by age, gender, and country. Science Education **93**(1), 131–160 (jan 2008). https://doi.org/10.1002/sce.20284, http://doi.wiley.com/10.1002/sce.20284
3. Chen, L., Chen, G., Wang, F.: Recommender systems based on user reviews: the state of the art. User Modeling and User-Adapted Interaction **25**(2), 99–154 (2015)
4. Chen, L., Pu, P.: Experiments on the preference-based organization interface in recommender systems. ACM Trans. Comput.-Hum. Interact. **17**(1), 1–33 (2010). https://doi.org/10.1145/1721831.1721836
5. Chen, L., Wang, F.: Sentiment-enhanced explanation of product recommendations. In: Proceedings of the 23rd international conference on World Wide Web. pp. 239–240. ACM (2014)
6. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 168–177. ACM (2004)
7. Keeney, R.L., Raiffa, H., Rajala, D.W.: Decisions with multiple objectives: Preferences and value trade-offs. IEEE Transactions on Systems, Man, and Cybernetics **9**(7), 403–403 (1979)
8. Sethi, R., Rangaraju, R.: Extinguishing the backfire effect: Using emotions in online social collaborative argumentation for fact checking. In: 2018 IEEE International Conference on Web Services, ICWS 2018, San Francisco, CA, USA, July 2-7, 2018. pp. 363–366 (2018). https://doi.org/10.1109/ICWS.2018.00062, https://doi.org/10.1109/ICWS.2018.00062
9. Sethi, R.J.: Crowdsourcing the Verification of Fake News and Alternative Facts. In: ACM Conference on Hypertext and Social Media (ACM HT) (2017). https://doi.org/10.1145/3078714.3078746
10. Sethi, R.J.: Spotting Fake News : A Social Argumentation Framework for Scrutinizing Alternative Facts. In: IEEE International Conference on Web Services (IEEE ICWS) (2017)
11. Sethi, R.J., Bry, L.: The Madsci Network: Direct Communication of Science from Scientist to Layperson. In: International Conference on Computers in Education (ICCE) (2013)
12. Sethi, R.J., Gil, Y.: A Social Collaboration Argumentation System for Generating Multi-Faceted Answers in Question & Answer Communities. In: CMNA at AAAI Conference on Artificial Intelligence (AAAI) (2011)
13. Sethi, R.J., Rossi, L.A., Gil, Y.: Measures of Threaded Discussion Properties. In: Intelligent Support for Learning in Groups at International Conference on Intelligent Tutoring Systems (ITS) (2012)
14. Tintarev, N., Masthoff, J.: Explaining recommendations: Design and evaluation. In: Recommender Systems Handbook, pp. 353–382. Springer (2015)
15. Wineburg, S., McGrew, S.: Lateral reading: Reading less and learning more when evaluating digital information (2017)