Extinguishing the Backfire Effect: Using Emotions in Online Social Collaborative Argumentation for Fact Checking

Ricky J. Sethi
Fitchburg State University

Raghuram Rangaraju
Fitchburg State University

***Abstract*—Controversial or complex topics often exhibit the backfire effect, where users' opinions harden in the face of facts to the contrary. We present initial work towards developing an online social collaborative argumentation system to verify alternative facts and misinformation by also including users' emotional associations with those stances. Our goal is to help users more effectively explore and understand their possibly subconscious biases in an effort to overcome the backfire effect and formulate more varied insights into complex and controversial topics. In order to aid this process, we model their emotional profile on such topics and combine it with a proposition profile, based on the semantic and collaborative content of propositions. We develop an algorithm to generate sentiment-based models of claims and propositions which we can filter based on users' inferred beliefs and the strength of those beliefs.**

***Keywords*-fact checking; fake news; backfire effect; misinformation; social collaborative argumentation;**

I. Introduction

The effect of alternative facts in recent elections starkly illustrates the critical need to have an electorate that is able to use evidence-based reasoning. Providing new facts alone, however, can lead to the **backfire effect**, where just giving people additional facts can, counter-intuitively, lead to users' opinions hardening in the face of evidence to the contrary [1].

Recent research in educational psychology suggests that epistemic emotions can mediate this backfire effect [2]. Epistemic emotions result from cognitive processes and can include emotions like surprise, frustration, etc. Psychoevolutionary psychologists examined [3] the connection between cognitive functions and emotions. They identified eight primary emotions that are essential for cognition and found that feelings can affect cognition and cognition, in turn, can affect feelings. This is best captured in Plutchik's Wheel of emotion [3] which consists of four paired positive-negative emotions:

1) Fear $\Rightarrow$ Anger
2) Joy $\Rightarrow$ Sadness
3) Acceptance $\Rightarrow$ Disgust
4) Expectation $\Rightarrow$ Surprise

There are many synonyms for these emotions and the number of basic emotions can range from 3 - 11 but psychoevolutionary theory sets 8 basic emotion dimensions [3], each with a number of synonyms or related terms.

As shown further in [3], fear is actually a measure of safety and, since Plutchik explains that emotions are feedback processes, rather than linear events, we favour the equivalent term of "contentment with the current safety level" instead of fear; as a consequence, in this paper, we use the synonym Contented instead of Fear. The full list of equivalent synonymous terms for these 8 basic emotions are shown in Table I.

In this paper, we thus propose an online system to help overcome the backfire effect and engender an informed populace that is capable of thinking critically by adding emotional associations while reasoning with evidence. Our prototype combines emotional profiles of users for each proposition along with various dimensions of user ratings, trust, and authority. We model users' emotional associations on complex, controversial topics and combine it with a proposition profile, based on the semantic and collaborative content of propositions. We develop an algorithm to generate sentiment-based models of claims and propositions which we can filter based on users' inferred beliefs and the strength of those beliefs.

II. Cognitive Fact Checking via Collaborative Argumentation

Modern attempts to combat the spread of falsehoods have focused mainly on these kinds of computerized, automated tools. Such tools flag previously identified hoaxes, or automatically detect fake news articles using natural language processing techniques with pre-existing ground truth, or track the viral-like transmission of hoaxes [5], [6], [7],

Plutchik	**Bader**	**Our Synonym**
Fear	Fear	Contented
Anger	Anger	Angry
Joy	Joy	Happy
Sadness	Sadness	Sad
Acceptance	Trust	Trust
Disgust	Disgust	Distrust
Expectation	Anticipation	Likely
Surprise	Surprise	Unlikely

Table I
Comparison of Synonymous Terms for Plutchik's 8 Basic Emotions from Plutchik [3], Bader [4], and our system.

Figure 1. Formal Argumentation Components: Stance ⇒ Claims ⇒ Evidence and Sources

[8], [9]. None of them, however, focus on verifying the statements contained in news and opinion articles.

Critical thinking and evidence-based reasoning are essential for countering propaganda and misinformation intended to manipulate public opinion [10], [11]. In particular, formal argumentation has been shown to be a natural, substantiated approach for analyzing the veracity and reliability of assertions and claims [12], [13]. In fact, in considering how to assess critical thinking, [13] asserts the need to identify conclusions, reasons, and assumptions as well as judging the quality of arguments and developing positions on an issue. Using this sort of evidence based reasoning not only has the potential to identify fake news to a greater extent but also to imbibe users with the critical thinking ability to navigate future fake news articles.

A. Our Argumentation Model

Following the approach of [14], we use formal argumentation to help analyze alternative facts. Thus, we also define an argument as being composed of Stances, Claims, and Evidence as shown in Figure 1. Both Claims and Evidence are supported by Sources, typically web documents. A Claim is either an inference or a conclusion while Evidence (sometimes called a Premise) provides the support for that Claim.

We thus create an argumentation Graph, $G_A = (V, E, f)$, composed of a set of vertices, V, edges, E, and a function, f, which maps each element of E to an unordered pair of vertices in V. Each fundamental Claim, Evidence, or Source in an argument thus constitutes an atomic argumentation component, v_a, and is embedded as a vertex in the graph such that $v_a \in V$. The vertices contain not just the component's semantic content, but also the ratings, authority, trust, and other attribute dimensions of each atomic argumentation component, including Sources. The edges $e \in E$ contain weights along the various dimensions of user ratings, trust, and authority as well as a pro/con designation for the connection. The argumentation framework also incorporates semantic web and linked open data principles.

III. Emotional Fact Checking

People are interested in alternative opinions [15] and social opinions on controversial topics are of significant interest to the public. This can become a problem as people sometimes form echo chambers with selective exposure to similar views. Even approaches like collaborative filtering for fake news detection can often lead to these echo chambers of like-minded people and ideas, thus further entrenching ideas in people's minds [16].

In fact, helping mitigate or at least assess the polarization of opinions can reduce this selective exposure to information [17], [18], [19]. Without the availability of such mechanisms, people only give heed to those views that are consistent with their own extant views.

On the other hand, tools that help people see and examine opposing views, or even just point out different perspectives on the topic, can help reduce the selective exposure. Such social opinions on controversial topics often exhibit emotions of opposite polarity when people have opposite stances on a topic. Quantifying the emotions associated with topics helps users gauge the general sentiment of the public [15], as well as allowing people to pay attention to those topics that have the most significance for them.

IV. The Emotional Signature

As discovered by [4], opinions express emotion towards the subject as well as emotions elicited *by* the subject. Although [4] explored this in the the context of opinions about movies, we use the general case of opinions on any subject in this paper. Following that line of thought, we quantify the emotional signature of each of the claim nodes, v_{claim}, in an argumentation graph, G_A.

We therefore modify the emotional signature [4] to be:

$$e_n = \frac{r_n^C}{P^C} \tag{1}$$

where the **emotional signature**, e_n, represents the weight of each of the eight emotions, $n \in 1 \ldots 8$, from Plutchik's Wheel, normalized to the range $[0, 1]$ for a particular proposition, P. We calculate r_n^C as the total number of ratings of emotion, n, in a Claim, C, and P^C as the total number of Claims, C, in a Proposition, P.

In Figure 2, we show Plutchik's Radar for Emotional Signatures of three Propositions; the first proposition is shown in Figure 3. The emotions on Plutchik's Radar as shown in Figure 2 are mapped to the emotions on the original Plutchik's Wheel of emotions [3].

Individual propositions can be clustered into topics, either semantically or via topic-modeling approaches. We can then use such Plutchik Radar charts to compare emotional signatures of different topics. This allows us to classify different topics by their overall emotional signatures distributions, as well.

V. Interface Design Principles

We derived three design principles, based on previous research [15], [16], [18], [20], that would support our goal to allow people to get multiple viewpoints on a complex topic, use evidence-based reasoning, and incorporate emotional

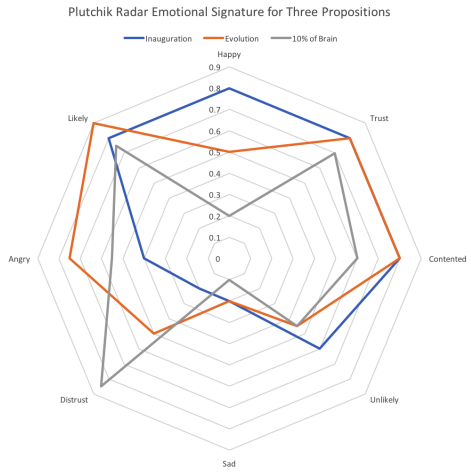

Figure 2. Plutchik's Radar for Emotional Signatures of three Propositions. The emotions on Plutchik's Radar are mapped to the emotions on Plutchik's Wheel.

associations to allow people to filter different perspectives. These design principles are:

1) Summarize support for different stances on a topic as determined by people's reactions and ratings, degree of emotion associated with a proposition, and the variety of social opinions.
2) Show degree and kind of emotional investment in each topic at different granularity. Users should be able to view emotional association at Stance and Claim levels and generalization.
3) Provide filters based on semantic support and emotional reactions inferred via NLP from semantic content as well as from sentiments self-expressed by users via Plutchik's emotions and depth of feeling.

VI. Combining Emotions and Social Collaborative Argumentation

The final online system combines these two components of cognition via critical thinking and of emotional associations into one intelligent interface. We combine the emotional signature, the depth of feeling, and the ratings captured in the edges of the argumentation graph.

In the user interface, we record both emotions and depth of feeling as seen in Figure 3. The callouts pop out on an on-hover event and the emotions range from Blue to Red (for positive and negative emotions, respectively) while the depth of feeling ranges from Dark Yellow to Dark Brown (for positive and negative feeling, respectively).

In addition, users can pick more than one emotion in the emotional call-out; i.e., they can make up to four selections, one for each of the positive-negative emotional pairs. However, users can only select a single depth of feeling rating on the depth of feeling slider callout. We further keep track of emotions and depth of feeling/sentiment on both a per-user basis as well as a per-claim basis, which is used in the per-Stance and per-Proposition bases, as well. These also feed into per-Topic basis evaluations.

The emotional signature, as defined in Section IV, is normalized to the range $[0, 1]$. The depth of feeling, or sentiment, is also normalized in the range $[0, 1]$. For a set of claims, we extract the edges based on the filters employed by the user, from the argumentation graph, G_A. These edges $e \in E$ contain an amalgam weight, also normalized to the range of $[0, 1]$, of the weights along the various dimensions of user ratings, trust, and authority. Pro-Con designations determine the polarity of the rating as positive-negative.

These are combined in a user-level value-utility model for each user for each proposition [16]. This is expressed as a set of pairs for each Claim, C, and amalgamated for the entire proposition as:

$$U(\langle C_1, C_2, \ldots, C_n \rangle) = \sum_{i=1}^{n} w_i V_i(C_i) \tag{2}$$

Here, V_i is the value function and w_i is the weight for each claim, as determined by normalized user ratings proportion. We can use this to calculate the average utility of each proposition in a topic as:

$$\overline{U} = \frac{1}{|S^P|} \sum_{c \in S^P}^{|S^P|} U(c) \tag{3}$$

Here, S^P is the set of propositions that satisfy the criteria for inclusion in a certain topic, T.

VII. Conclusion

In this initial work towards characterizing the backfire effect, we were able to compute Plutchick Radar Emotional Signatures for a variety of complex issues (e.g., whether Donald Trump's Inauguration numbers were larger, whether evolution is real, whether we use 10% of our brain) and found these signatures have the ability to distinguish between propositions. Future work will involve conducting in-depth user studies across multiple populations to gauge the efficacy of the user-level value-utility model and the potential mitigation of the backfire effect in combating fake news and misinformation.

References

[1] B. Nyhan and J. Reifler, "When corrections fail: The persistence of political misperceptions," *Political Behavior*, vol. 32, no. 2, pp. 303–330, 2010.

[2] G. J. Trevors, K. R. Muis, R. Pekrun, G. M. Sinatra, and P. H. Winne, "Identity and epistemic emotions during knowledge revision: A potential account for the backfire effect," *Discourse Processes*, vol. 53, no. 5-6, pp. 339–370, 2016.

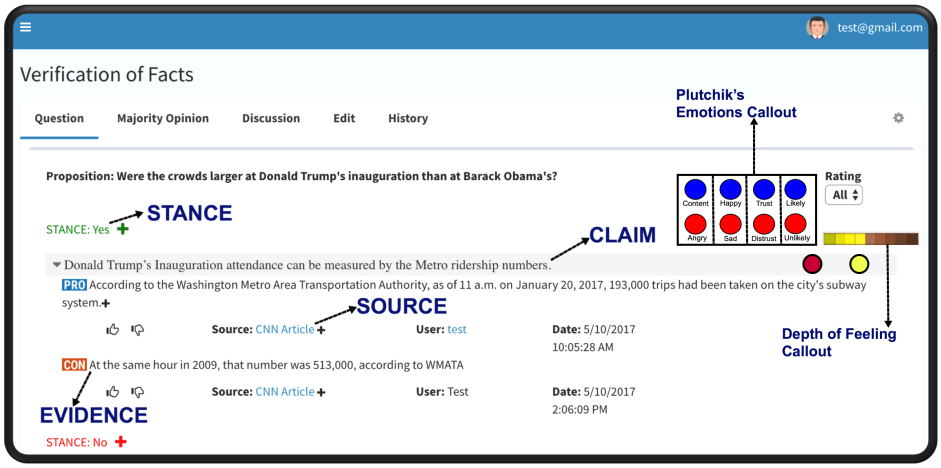

Figure 3. The final system annotated with Argumentation Components and Emotional Callouts. The primary interface shows the main Argumentation Components of Stances, Claims, Evidence, and Sources. This also shows the two emotional callouts: one for Plutchik's 8 Emotions, paired in groups of two, and the Depth of Feeling slider, which ranges from Positive (Dark Yellow) to Negative (Dark Brown). Finally, it shows the Proposition, which is the main question. In addition, this shows tabs for both the Majority Opinion, as well as the Discussion which allows interaction between members of the virtual community.

[3] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American Scientist*, vol. 89, no. 4, pp. 344–350, 2001. [Online]. Available: http://www.jstor.org/stable/27857503

[4] N. Bader, O. Mokryn, and J. Lanir, in *Intelligent User Interfaces (IUI)*, 2017. [Online]. Available: http://dx.doi.org/10.1145/3030024.3040982

[5] V. Rubin, Y. Chen, and N. J. Conroy, "Deception Detection for News: Three Types of Fake News," *ASIS&T*, 2015.

[6] N. J. Conroy, , V. L. Rubin, and Y. Chen, "Automatic Deception Detection: Methods for Finding Fake News," in *ASIS&T*, 2015. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/pra2.2015.145052010082/full

[7] Z. Jin, J. Cao, Y.-G. Jiang, and Y. Zhang, "News Credibility Evaluation on Microblog with a Hierarchical Propagation Model," *2014 IEEE International Conference on Data Mining*, pp. 230–239, 2014. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7023340

[8] A. Kucharski, "Post-truth: Study epidemiology of fake news," *Nature*, vol. 540, no. 7634, pp. 525–525, 2016. [Online]. Available: http://www.nature.com/doifinder/10.1038/540525a

[9] M. Tambuscio, G. Ruffo, A. Flammini, and F. Menczer, "Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks," *WWW*, pp. 977–982, 2015.

[10] M. J. Sproule, *Propaganda and democracy : the American experience of media and mass persuasion*. Cambridge, U.K. New York, NY: Cambridge University Press, 1997.

[11] M. B. Sethi, "Information, education, and indoctrination: The federation of american scientists and public communication strategies in the atomic age," *Historical Studies in the Natural Sciences*, vol. 42, no. 1, pp. 1–29, 2012. [Online]. Available: http://hsns.ucpress.edu/content/42/1/1

[12] R. Johnson, *The Rise of Informal Logic*. Windsor Studies in Argumentation, 1996. [Online]. Available: https://windsor.scholarsportal.info/omp/index.php/wsia/catalog/book/9

[13] R. H. Ennis, "Critical thinking assessment," pp. 179–186, 1993.

[14] Anonymized, "Anonymized," in *IEEE Anonymized Conference*, 2018.

[15] M. Gao, H. J. Do, and W.-T. Fu, "An Intelligent Interface for Organizing Online Opinions on Controversial Topics," in *Intelligent User Interfaces (IUI)*, 2017.

[16] L. Chen and P. Pu, "Experiments on the preference-based organization interface in recommender systems," *ACM Trans. Comput.-Hum. Interact.*, vol. 17, no. 1, pp. 1–33, 2010.

[17] S. Faridani, E. Bitton, K. Ryokai, and K. Goldberg, "Opinion space," *Proceedings of the 28th international conference on Human factors in computing systems - CHI 10*, 2010.

[18] Q. V. Liao and W.-T. Fu, "Can you hear me now?" in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14*. ACM Press, 2014. [Online]. Available: https://doi.org/10.1145%2F2531602.2531711

[19] Q. V. Liao, W.-T. Fu, and S. S. Mamidi, "It is all about perspective," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*. ACM Press, 2015. [Online]. Available: https://doi.org/10.1145%2F2702123.2702570

[20] S. A. Munson, S. Y. Lee, and P. Resnick, "Encouraging reading of diverse political viewpoints with a browser widget." in *ICWSM*, 2013.