ABSTRACT

Online deliberation platforms such as Consider.it and the Deliberatorium explore a variety of approaches to address polarization resulting from internet-based self-selection that can amplify inherent confirmation bias and result in an "echo chamber" where initial biases are reinforced rather than explored and resolved. This paper presents initial results from DebateCAFE v1.0, a prototype platform that introduces an incentive mechanism to encourage participants to articulate persuasive arguments on both sides of an issue. It uses collaborative filtering to highlight and rank the most persuasive arguments. The score for each participant is the minimum of the Wilson score for each argument, ie, based on the “weaker” of the two arguments entered. To evaluate performance, we used the topic of "personal privacy vs. national security" which was widely discussed in the U.S. in Spring 2016 when Apple refused FBI's request to unlock a terrorist's cellphone. Initial results with 94 participants suggest that DebateCAFE v1.0 can effectively encourage participants to articulate arguments for both sides of an issue, and identify persuasive arguments. Participant’s score adequately reflects participant’s capability of articulating adversarial arguments. However, most participants are more capable of providing persuasive arguments for their own position and they tend to give higher ratings to arguments that align with their position. The report summarizes data and system performance from a preliminary study with 94 participants who entered 170 arguments on both sides and 1754 peer-to-peer ratings of the arguments.

1. INTRODUCTION

Online deliberation systems are platforms where groups of participants form structured discussion and debate to reach informed decisions that usually involve complex issues and difficult trade-offs [8, 15, 31, 32, 34]. One potential drawback is selective exposure, when participants seek out confirming information and resist changing their opinions. Numerous studies have noted that this is an important problem online [1, 6, 9, 10, 13, 24, 26, 28]. This problem has been considered before by platforms such as Consider.it [16] and the Deliberatorium [12] but as the number of participants grows, these frameworks can place a significant burden on the moderators. Novel online deliberation platform is needed to scalably automate or assist the moderation task of filtering out valuable items efficiently.
This platform builds on Opinion Space [17, 18] and is a variation of the CAFE framework [22, 38].

DebateCAFE uses Collaborative Filtering (CF), an approach widely adopted by companies such as Amazon and Netflix, to recommend products on the basis of tested algorithms. CF allows DebateCAFE to scale to an expanding population of participants and process unstructured textual data on subjective access (i.e., how persuasive is this idea). While collaborative filtering is effective in bringing some structure to lists of items by assigning reputation scores, this has to be coupled with incentive mechanisms and interfaces to avoid the biases of selective exposure [4].

DebateCAFE v1.0 primarily addresses the issue of collecting and managing large-scale textual data while mitigating selective exposure. In future versions, we aim to observe how people change positions to reach consensus of views or find effective approaches in the middle ground of various arguments. Proposed approaches include argument synthesis, where the platform requests participants to articulate a new argument based on a pair of existing arguments (whether or not they are similar or distinct); argument rebuttal, where participants are requested to provide a counterargument in response to an existing one; and an opinion tracker which enables participants to adjust their opinion after each activity.

This paper explores a case study on the contentious issue of “Apple vs. FBI”, inspired by the ongoing debate of whether private companies (i.e., Apple Inc.) should cooperate with government (i.e., FBI) requests to have backdoor access to encrypted information technologies. We report data and system performance from a preliminary study with 94 UC Berkeley students who entered 170 arguments on both sides and 1754 peer-to-peer ratings of the arguments. Results offer valuable insights into platform mechanism design but the statistics should not be interpreted to reflect the general population because of the biased sample.

2. RELATED WORK
2.1 Online Deliberation Platforms
In the past ten years, with growing numbers of researchers and government officials recognizing the effectiveness of public deliberation, many online deliberation platforms have been created. Klein’s MIT Deliberatorium structures knowledge into issues/problems, ideas/solutions, and arguments for or against [12]. However, the platform is unable to harvest the best ideas when crowds start to scale, especially when dealing with complex problems. Chilton et al. introduced a collaborative application, Frenzy, to gather the entire program committee for conference session-creation [7]. This crowd-sourced approach significantly reduced the time needed to make a conference program. Krilean et al. developed a plug-and-play platform called Consider.it, which enables participants to view and articulate arguments on a linear scale of a particular issue. Participants are exposed to pro/con arguments and are encouraged to adopt arguments they find persuasive and articulate pro/con points [16]. Consider.it then requests participants to position themselves on a linear scale to reflect their stand on the issue. Debategraph is another deliberation tool that visually organizes complex debates into graph, where each node represents an argument and each edge reflects arguments relationships [3]. Participants are also encouraged to provide evidence for each sub-argument. While these platforms are valuable in organizing arguments and collecting feedback, arguments on both sides can be greatly lopsided and the large volume of textual arguments make it difficult to identify valuable insights. DebateCAFE, by design, is able to balance the number of arguments on contradicting viewpoints and uses collaborative filtering to tackle the scale issue. DebateCAFE further introduces a scoring mechanism to incentivize persuasive arguments on both sides.

2.2 Selective Exposure
Sears and Freedman describe selective exposure as: “People prefer exposure to ideas that agree with their pre-existing opinions” [30]. An inherent characteristic of the Internet is the freedom to select from a range of content. However, this freedom enlarges the concern that individuals only view information that aligns with their personal attitude. Research has shown that people differ in their willingness to be exposed to adverse information, especially political content [25, 29, 33], and that exposure to online content is tailored by algorithms, creating what Pariser (2011) refers to as “filter bubbles” that limit exposure to divergent viewpoints [23]. This phenomenon of unbalanced exposure could largely hinder people from making rational and informed decisions and can impede political tolerance [14]. Our platform is designed to mitigate this bias by encouraging participants to articulate reasoning for a pair of contradicting positions.

2.3 Collaborative Filtering (CF)
Collaborative filtering combines subjective ratings provided by humans to assign a numerical reputation to each item [11]. In many CF systems, such as Amazon and Netflix, the reputation of each item is based on the aggregated ratings from a neighborhood [20]. CF can also be applied globally when the reputation of an item depends on the aggregated ratings from ALL participants, reflecting a common opinion of the crowd.

Most CF systems adopt a list-based presentation, resulting in unbalanced exposure of items, where highly rated items are shown on top [37]. Though the system does not set out to bias any item, the self-selection nature of humans could largely hinder new items due to limited exposure [21]. This can be particularly counterproductive for online deliberation platforms, where arguments on one side could enjoy greater exposure, discouraging participants to articulate and rate arguments with contradicting views. In DebateCAFE v1.0, we balance the exposure of arguments by simultaneously displaying a subset with balanced positions and mixed rankings, permitting the system to present arguments on both sides equally and collect feedback on all arguments.

3. PLATFORM DESCRIPTION
3.1 The Apple vs. FBI Issue
To evaluate platform performance, we chose the topic of “personal privacy vs. national security” for our first deployment.

The issue arose when the FBI requested that Apple help investigators gain access to an iPhone used by Syed Rizwan Farook in the December 2015 mass shooting in San Bernardino, CA [2]. Apple refused the request because it would require writing new software to bypass encryption features of the iPhone and would create “the potential to unlock any iPhone in someone’s physical possession.” [5]. This particular incident has attracted attention from the public and led to debates on social media platforms and related online forums. Starting with this incident, the discussion has expanded to a more general theme of “personal privacy vs. national security”. We perceive this issue as a highly complex and controversial topic that may be clarified...
through deliberation allowing the public to articulate ideas side-by-side.

3.2 INTERFACE
DebateCAFE guides participants through three stages: assessment, argument articulation, and peer-to-peer argument evaluation.

3.2.1 Assessment Phase
To begin, participants assess their current security/privacy beliefs by rating three Initial Bias Assessment questions (IBAs) on a 10-point scale from 0 (Strongly Disagree) to 9 (Strongly Agree).
1. I am willing to give up some privacy for increased security.
2. Personal privacy should be guaranteed by the US Constitution.
3. There are reasonable arguments on both sides of the security and privacy debate.

These statements were chosen as they succinctly participants’ initial stance (pro-Apple vs. pro-FBI) on digital privacy, as well as their degree of open-mindedness to opposing arguments. Participants have the option to skip any question they choose not to answer by either hitting the skip button or leaving the response blank.

Participants are then asked to provide their zip code. We find zip code to be an informative demographic statistic, while not being so intrusive as to hinder further participation in the system. After that, DebateCAFE displays the histograms of the responses to the three IBAs so that the participant learns his/her stance on these questions among all participants.

3.2.2 Argument Articulation Phase
In this phase, participants are prompted to formulate their own arguments in response to the central discussion question: “In the future, should Apple cooperate with FBI requests for personal data?” Compared with previous CAFE instances, DebateCAFE is novel in that participants are prompted to enter two arguments: one for each side. To help bootstrap participant’s writing, we pre-populate the textboxes with “Apple should cooperate because”, and “Apple should not cooperate because.” The interface strongly encourages, but does not require that a participant fill in both arguments. DebateCAFE also requests that participants supply their email address so we can update them with peer-rating scores on their arguments (explained further in section 3.3).

3.2.3 Peer-to-Peer Argument Evaluation Phase
Next, participants enter the “discussion space,” a 2D visualization where other participants’ arguments are represented by spheres arranged across the space. This discussion space displays 8 argument spheres at a time, 4 “pro” and 4 “con” arguments. Seeking to ensure that all arguments are seen and rated, DebateCAFE prioritizes the display of arguments that have high uncertainty in their evaluation grades. We quantify uncertainty using the standard error = std/sqrt(N), a bound on how far away the sample mean is from the population mean.

The spheres are placed in the 2D space according to the first two dimensions of a Principal Component Analysis (PCA) applied to participants’ responses during the assessment phase [36]. (Skipped questions are assigned the mean response rating for that question.) Participants then click on the spheres in the 2D space to read other participants’ arguments. Finally, using a rating interface similar in design to that used during the assessment phase, participants evaluate their peers arguments on the question “How persuasive is this argument?” again using a scale from 0 (Not at all Persuasive) to 9 (Extremely Persuasive).

3.3 Scoring Mechanism
A key challenge of crowd sourcing platforms is the high dropout rate, where participants lack the incentive to click through the platform and spend time articulating persuasive arguments. To create such incentive, DebateCAFE not only provides instant feedback in the assessment phase, but also computes a “persuasiveness score” for each participant. Participants receive two scores \( s_1, s_2 \) for their two arguments “pro-Apple” and “pro-FBI” computed by the statistic described below. We record the overall quality of a participant's contributions as the minimum of these two scores \( \min(s_1, s_2) \). Intuitively, the minimum score is intended to disincentivize a participant from providing a low-quality response for the position they do not support, encouraging them to articulate persuasive arguments on both sides.

In crowdsourced rating systems, the average rating can give rankings that are not robust due to small sample size. In DebateCAFE, we calculate argument score with the lower bound of the binomial proportion confidence interval (also called the Wilson Score) [35]. Intuitively, this approach is more robust because it incorporates information about the uncertainty of the score estimate. For example, a suggestion that receives ratings \{10, 0\} is ranked lower than one that is rated \{5, 5\}.

4. PRELIMINARY RESULTS
In this section, we present results from a preliminary analysis of undergraduate students at UC Berkeley who were assigned to participate in DebateCAFE. The resulting dataset contains responses from 94 students with 284 IBA responses, 170 new arguments on both sides and 1754 peer-to-peer ratings. Note that the population in this study is not a representative sample but this informal study provides insights into the performance of DebateCAFE and suggestions for future design choices. For analysis purpose, we scale the raw ratings to 0-1.

4.1 The Initial Bias Assessment Questions
The first two questions provide a rough indication of the participant’s initial attitude toward the issue and the third question captures how open the participant is to deliberation. Figure 2 panels (a), (b) and (c) show the histogram of each of the three IBAs. The response to IBA1 has a wider spread than IBA2 and IBA3 with a mean at 0.54, indicating people have varied preference in giving up privacy for security. IBA2 is skewed left suggesting that most participants value personal privacy and regard it as a constitutional right. For IBA3, responses are clustered at 0.7-1.0, reflecting an appreciation of arguments on both sides of an issue.

The relation between IBA1 and IBA2 is interesting. Figure 2 panel (d) shows the scatter plot of IBA2 vs. IBA1. Observe that most points lie above the y=1-x line, demonstrating that participants who are not willing to give up privacy for security have a strong pursuit of personal privacy. The blue line is the regression line between the two IBAs. The slightly negative slope indicates negative correlation between the two issues. Participants on the top left of the plot are likely to favor Apple’s position, whereas participants on the bottom right are likely to favor the FBI. The participants’ initial estimated positions are determined by the following metric:

If IBA2 - IBA1 ≥ 0.3, this participant is “pro-Apple”;  
If IBA1 - IBA2 ≥ 0.1, this participant is “pro-FBI”. 

These criteria are not symmetric because the ratings received for IBA1 and IBA2 are asymmetric and we want two groups of
similar sizes. This metric results in 27 “pro-Apple” participants, 23 “pro-FBI” participants and 44 “neutral” participants as shown in Figure 2 panel (d).

4.2 Arguments on Both Sides

Out of a total of 170 arguments, after ranking with the Wilson metric, 17 of the top 20 were “pro-Apple” arguments and 3 were “pro-FBI” arguments. Here we present the top 3 arguments on both sides of the issue.

Pro-Apple (personal security) arguments:
1. Apple should not cooperate because it sets a dangerous precedent with regards to privacy and the FBI and other investigative groups. It also has the potential for issues with the same opening being exploited by hackers.
2. Apple should not cooperate because allowing the FBI access to these systems opens a door that cannot be closed again at will. Creating a cyber-security loophole allowing access for the FBI means that anyone with the technical knowledge can also exploit this access; a weakness in the fundamental security creates vulnerability. Our personal information would be vulnerable to sophisticated attacks by hackers and terrorists themselves.
3. Apple should not cooperate because it violates customer's privacy. If customers do not trust Apple with their data anymore then sales will drop and Apple will no longer be relevant in tech. One thing that the government is asking is to create a backdoor for a particular OS, which could lead to hackers exploiting a bug in subsequent OS'. This is a huge security red light because Apple does not and should not knowingly create a backdoor which could lead to major security concerns in the future.

Pro-FBI (national security) arguments:
1. Apple should cooperate because it is a small price to pay for the increased understanding of this terrorist act.
2. Apple should cooperate because if there is a little compromise involved in guaranteeing the security of the nation, then as a resident and/or citizen of this nation, people should be willing to make that compromise.
3. Apple should cooperate because there is an increased security risk when they do not cooperate. If potential terrorist information is on the phones of those like the San Bernardino shooters, it will jeopardize the entire safety and security of the United States.

These arguments are rather well articulated with a clear position and concrete reasoning. However, argument duplication is an issue: we observe many arguments conveying the same idea with slightly different wording. For example, the top 2 “pro-Apple” arguments both mention that complying with the FBI’s request could lead to potential cyber-attack from hackers and the top 2 “pro-FBI” arguments point out the tradeoff between personal privacy and increasing national security from terrorist acts.

4.3 Clustering

4.3.1 Measuring Selective Exposure

As mentioned earlier, selective exposure is a major concern for online deliberation platforms, and DebateCAFE is able to quantify the extent of this behavior. Recall that DebateCAFE’s peer-to-peer argument evaluation interface presents 8 arguments covering both sides of an issue. Each argument is represented by a sphere in the space with a tag of either “Apple” or “FBI” indicating the position of the argument. Participants are free to click on any sphere in the space when they first land on this page. This design enables us to observe which side of the argument a participant chooses to view first given his/her initial bias estimated by the IBAs. 20/27 pro-Apple participants first selected an argument for “Apple” and 11/23 pro-FBI participants first selected an argument for “FBI”. The difference is not significant; the balanced number of spheres in the space may help mitigate selective exposure.

4.3.2 Articulation Bias

Here we would like to compare the quality of participants’ arguments for and against their initial position (as estimated from the results of their IBAs). Peer-ratings of argument persuasiveness reveal that among the pro-Apple participants who provided rated arguments, only 3/20 have a higher rated argument for “FBI”. However, among the pro-FBI participants who provided rated arguments, 8/21 have a higher rated argument for “FBI”. Some participants did not provide arguments and some arguments were not rated. Results indicate that the pro-Apple participants were less likely to articulate persuasive arguments for the opposing view, while pro-FBI participants had little trouble crafting persuasive arguments for the opposing view, but the difference is not significant.

4.3.3 Peer Rating Bias

We received 1754 valid peer ratings, among which 851 rated “pro-FBI” arguments and 903 rated “pro-Apple” arguments. By adopting the Welch two-sample t-test, with a t of 4.289 and p-value of 0, we conclude that the “pro-Apple” arguments are, on average, receiving significantly higher ratings than “pro-FBI” arguments. We conjecture this difference comes from the
inherent bias of participants, who tend to rate more highly those arguments that align with their position.

We classify the arguments as either “Consistent” or “Adversarial” with respect to initial bias. From participants in the pro-Apple and the pro-FBI groups, “Consistent” arguments received an average rating of 0.523 with a Standard Deviation (SD) of 0.25 and a Standard Error (SE) of 0.0118, while “Adversarial” arguments received an average rating of 0.485 with a SD of 0.25 and a SE of 0.0119. Although the difference is under 5% on the 1.0 scale, with a p-value of 0.022, the Welch two-sample t-test suggests that the true difference in means is not equal, suggesting a small consistency between initial bias and ratings bias. However, a controlled experiment is warranted to explore this.

4.4 Participant Score

DebateCAFE defines participant score as the minimum of the two argument scores: min(s1, s2), i.e., based on the weaker of the two arguments entered and thus reflecting a participant’s ability to articulate adversarial arguments. We scale all the scores to 0 to 1 for more intuitive comparison. From Figure 3, we observe that participant scores follow a normal distribution, where few participants are extremely capable or incapable of articulating adversarial arguments and most participants are mediocre. Among the top-scoring participants, we confirm that they are able to provide strong adversarial arguments. For example, the two arguments from a top-scoring participant are:

(Pro-FBI argument) Apple should cooperate because it is a small price to pay for the increased understanding of this terrorist act.
(Pro-Apple argument) Apple should not cooperate because if this gets scaled to all cases, it could seriously diminish the customers’ privacy.

Note the first argument is among the top 3 pro-FBI arguments and the second argument focuses on the consequence of scaling such action to all cases.

Among the low-scoring participants, the Wilson scores of their two arguments are highly lopsided. The participant with the lowest score only provided a pro-Apple argument, leaving the pro-FBI argument as “Apple should cooperate because”. Another participant with a low score gave a “pro-FBI” argument as “Apple should cooperate because public security?” while providing a well-articulated “pro-Apple” argument: “Apple should not cooperate because it has paid such amount of advertisement and technology to increasing security, so it should not yield all to FBI”.

5. Discussion

We present a novel deliberation platform, DebateCAFE, designed to encourage participants to review and articulate adversarial arguments on contentious issues. DebateCAFE implements collaborative filtering to identify persuasive arguments and balance the exposure of polarized viewpoints using importance sampling. By introducing a scoring system to participants, DebateCAFE incentivizes them to articulate persuasive arguments on both sides of the issue.

We describe an application of DebateCAFE to the contentious personal privacy (Apple) vs. national security (FBI) issue and demonstrate preliminary results from an informal study. By presenting an equal number of arguments on both positions, our platform seems to mitigate selective exposure, achieving a more balanced discussion around the issue. Even though a small sample size did not permit us to show that DebateCAFE helped participants change their minds on this particular issue, this first attempt did demonstrate the capacity of the system to measure and effectively discourage selective exposure bias without relying on high-cost human moderation.

6. FUTURE WORK

6.1 De-duplication

During the deployment, we found that argument duplication was a significant problem. Many participants provided similar arguments with slightly different wording. For example, among the “pro-Apple” arguments, many identified the potential that hackers could gain access to all Apple devices. Duplicate arguments should be consolidated to optimize the effectiveness of participants’ peer-ratings. One possible solution is to introduce a moderator, who reads and consolidates arguments. However, when the platform scales, it would be infeasible for the moderator to view every single argument. An alternative approach is to identify potentially similar arguments using Natural Language Processing techniques, and then enlist participants (in an interface similar to that used in the Argument Articulation Phase) to de-duplicate or synthesize those arguments.

6.2 Reconnecting with Participants

Another question that emerges from this initial case study is how to close the loop to connect back to the participants and measure opinion change over time. DebateCAFE offers participants the option to input their email addresses so that the actual scores of their arguments (delayed due to peer-rating) can be shared with them later, closing the loop of participation. However, very few participants entered an email address, indicating that this was insufficient incentive (or perhaps that this incentive only functions when participants believe they will be returning to the DebateCAFE community in the future). We are considering adding a leaderboard where top-rated arguments are displayed. Another potential approach would be to instantly offer an estimated score of the arguments’ persuasiveness based on the actual scores of similar arguments in the system.

6.3 Measuring Changes in Opinion

When participants rate persuasiveness of arguments, they might give high ratings to arguments that are logically coherent but are peripheral to the issue or of marginal importance. To better measure “persuasiveness” as opposed to “logical validity,” we will experiment with a slider indicating the participant’s position on the issue. It will be available throughout the discussion phase, allowing the participant to record her/his opinion change after
viewing each argument. The persuasiveness of each argument will then be captured by the total opinion changes of other participants.

6.4 Future Deployment and Design Choices

The current platform uses a keypad design for ratings. In the future, we would like to experiment with other designs such as sliders to explore how design choices affect response. In addition to improving the design of the platform, we would like to gather more participants with diverse demographics to participate in DebateCAFE.

7. REFERENCES


