Design Challenges for a Multi-Perspective Search Engine

Sihao Chen* **Siyi Liu*** **Xander Uyttendaele**
Yi Zhang **William Bruno** **Dan Roth**
University of Pennsylvania
{sihaoc, siyiliu, xanderu, yizhang5, wwbruno, danroth}@cis.upenn.edu

Abstract

Many users turn to document retrieval systems (e.g. search engines) to seek answers to controversial questions. Answering such user queries usually require identifying responses within web documents, and aggregating the responses based on their different *perspectives*.

Classical document retrieval systems fall short at delivering a set of *direct* and *diverse* responses to the users. Naturally, identifying such responses within a document is a natural language understanding task. In this paper, we examine the challenges of synthesizing such language understanding objectives with document retrieval, and study a new *perspective-oriented* document retrieval paradigm. We discuss and assess the inherent natural language understanding challenges in order to achieve the goal. Following the design challenges and principles, we demonstrate and evaluate a practical prototype pipeline system[1]. We use the prototype system to conduct a user survey in order to assess the utility of our paradigm, as well as understanding the user information needs for controversial queries.

1 Introduction

In the past two decades, web search has developed as a ubiquitous way for users to retrieve web information. Typically, A web search query is driven by a specific intent, or information need, i.e. the users' desire to seek answers to an open-ended question (Shneiderman et al., 1997; Rose and Levinson, 2004). Among the various types of questions, one of the most challenging types, from the perspective of designing an information retrieval (IR) system, is *debate-worthy*, *controversial questions*, for which an objectively-true answer or consensus seems impractical to reach. Rather, explicit or implied responses with *different perspectives*, can be found within topically-related web documents.

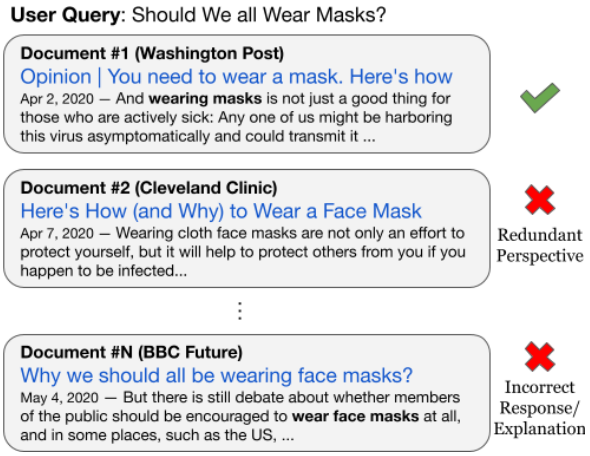

Figure 1: Example retrieved articles with snippets from a search engine with respect to the query "Should we all wear masks" dated from April to May 2020, during which mask wearing was a controversial topic. The retrieved articles contain redundant and often inconsistent information in headlines and the extracted snippets.

Classical IR systems fall short at delivering responses with a diverse set of perspectives to the user query. Instead, they typically provide a ranked list of references to relevant but not necessarily trustworthy web documents, with limited justifications on what part(s) of the document could answer the user query (Metzler et al., 2021). We argue that a document retrieval system should ideally be equipped with the ability to extract or generate such responses from documents conditioned on the user query. Furthermore, a retrieval system should recognize the semantic difference of responses in cross-document settings, and in turn organize and deliver a set of documents from diverse perspectives. These objectives can inherently be formulated into a suite of natural language understanding tasks (Chen et al., 2019b; Lamm et al., 2020).

[1]Demo at http://dickens.seas.upenn.edu:4011

* Equal Contributions

The goal of this paper is to discuss and validate the users' need of *direct and diverse* responses when issuing controversial, debate-worthy queries to a search engine, and examine the ideas of formulating such objectives as natural language understanding tasks, and incorporate them in a novel document retrieval paradigm. Through the lens of key design principles of a task-specific search engine (Rieger, 2009), we first discuss the challenges in designing such a multi-perspective search engine. We introduce a paradigm that revolves around the abstraction of *responses*, i.e. a text segment that serves as direct answer to users, and *perspectives*, the semantic implfications of the opinions expressed in a response in the context of a given user query (Chen et al., 2019b).

To assess the utility of the proposed paradigm, we develop and demonstrate a prototype open-domain news search engine. Given a user query of controversial topics, The search engine retrieves news articles; identify/generate responses within the articles along with evidence paragraphs; categorize and display the responses according to their perspectives differences. With the prototype, we conduct a user study in comparison to Google Search. With the goal of understanding and uncovering the user needs related to controversial queries, we discuss and analyze the advantages/disadvantages of our prototype through the user study responses. We then summarize and offer insights on the technical challenges and trade-off to cater to the specific user information needs with respect to controversial queries. We hope our findings will facilitate future research on this topic.

In summary, our contributions are as follows:

1. We study a document retrieval paradigm with the objective of delivering *direct* and diverse summary responses to user query. We outline a list of natural language understanding challenges to achieve the goal.

2. We demonstrate and evaluate a prototype perspective-oriented search engine that retrieve, organize and summarize the perspective behind news articles.

3. We conduct a user study with our prototype against Google Search to assess the utility of our paradigm, as well as to uncover the information needs by users with respect to controversial, debate-worthy queries.

2 Related Work

The importance of retrieval result diversity has long been recognized by the information retrieval community (Goffman, 1964; Clarke et al., 2008). The problem has mostly been studied in the context of topic or entity ambiguity in the user query (Agrawal et al., 2009), and most solutions focus on resolving the ambiguity and uncovering the user intent, or hierarchy of the target topics (Sakai and Song, 2011; Dang and Croft, 2013; Hu et al., 2015; Wang et al., 2017). The most notable hypothesis for this line of work, which our study also investigates, is the importance of modeling the inter-dependencies among documents during retrieval. The TREC 2009 Web track's diversity task (Clarke et al., 2009) introduces an evaluation protocol where the relevance of next retrieval item is dependent on the previous result.

Our study takes a different angle from query ambiguity, and hypothesize that in the case of debate-worthy query, users wish to see *direct* responses from diversified perspectives in the retrieved documents (Metzler et al., 2021). Such objectives requires recognizing the argumentation structure of a document (Stab and Gurevych, 2014; Ein-Dor et al., 2020), and more importantly comparing the semantics and implications of the arguments made cross documents (Bar-Haim et al., 2020). Chen et al. (2019b,a) formulate a argument retrieval problem that conceptually resembles our formulation. In their task, a system is expected to return single-sentence arguments from different perspectives given a user query. Our study instead focuses on the more practical and challenging setting of document-level retrieval.

The task of question answering (QA) aims to identify *direct* answers to users questions expressed in natural languages from either close- or open-domain information resources (Kolomiyets and Moens, 2011). Naturally, document retrieval has been an essential step during answer candidate generation (Chen et al., 2017). Among various types of QA tasks, Machine reading comprehension (Rajpurkar et al., 2016; Khashabi et al., 2018) and open-domain QA (Joshi et al., 2017; Dunn et al., 2017) resemble our problem closely, as both involves identifying a concise response from large- to web-scale corpora. Under such settings, researchers have found benefit in jointly learning QA and retrieval tasks (Das et al., 2018). The key difference between most QA tasks and our problem

is that objective correctness does not exist for most controversial queries, and so they cannot be responded with a single correct answer. Instead, there exist dimensions such as stance (Bar-Haim et al., 2017) and persuasiveness (Carlile et al., 2018) to measure the semantic difference between relevant and arguably equally valid arguments. To this end, recent researchers have attempted to extract or generate explanations alongside search results (Lamm et al., 2020).

3 Design Challenges

3.1 User Information Need

We start by speculating and discussing the needs of users when issuing controversial and debate-worthy queries to a search engine. Rieger (2009) introduces three key factors in building a task-specific search engine that caters to user's information needs. We introduce each factor below, and discuss the connections and implications to the case of controversial, debate-worthy user queries.

Satisfaction refers to overall success of the retrieved results in providing help to the user's specific search intent. Here there are two levels of user satisfaction to consider. The first dimension is whether the retrieved results *contain* the answer that user needs. The second more challenging yet crucial dimension is whether search engine is able to highlight and display the answer (Metzler et al., 2021). In the context controversial queries, identifying direct answers from retrieved documents seems essential, as it will greatly reduce the amount of readings user need to do before deriving the answer. However, automating the process comes with technical challenges, and such automation often ties closely with concerns of model biases, as well will further discuss in x 6.

Search Intent refers to the type and format of answers user wishes to see when issuing a query. In the case of controversial queries, as users typically do not know a priori the potential aspects or perspectives that could lead to the answer in their context, organizing and highlighting a *comprehensive* set of results with *diverse* perspectives in a parallel fashion becomes important in this case (Chen et al., 2019a).

Trust is originally defined by Pan et al. (2007) as users' trust in a search engine's capability of retrieving relevant results to the query. However, as concerns over the veracity of web information grow (Thorne et al., 2018), user trust revolves more closely around the trustworthiness of the sources and content (Pasternack and Roth, 2013). For controversial queries, as multiple perspectives exist for every issue, whether each perspective is corroborated by evidence from trustworthy sources becomes the key to ensure user trust (Chen et al., 2019b).

In the following sections, we take a closer look into the specific problems we discussed for the three factors, and the challenges in formulating NLP tasks to address the problems.

3.2 Identifying Direct Responses to Query

From the classical view of information retrieval, identifying responses, or *explanations* (Lamm et al., 2020) is inherently a challenging problem for term-based retrieval models. To this end, neural re-ranking has become popular technique to re-score the retrieved candidate documents (Xiong et al., 2017).

Our problem formulation for this part first takes inspiration from neural re-ranking, where we rely on a classical yet more scalable retrieval function to identify document candidates that potentially contain the responses. The next part of the problem, response identification, closely resembles the task of question answering. Given a document d and a user query q, we want to identify one of more text segments $fs_i g\ 2\ d$ such that each $s\ 2\ fs_i g$ can be used as a direct response to q. The key difference, as we discussed in the previous section, lies within the difficulty to define a single, valid response. Previous research in argumentation community has developed a few dimensions, such as argument stance, relevance, and persuasiveness to measure the validity of a text segment being a valid argument towards a user query. Using the corpus resources developed for such tasks (Bar-Haim et al., 2017; Ein-Dor et al., 2019), we can build models to score the text segments within a document along dimensions; aggregate the dimensions and retrieve a ranked list of the valid responses.

One particular challenge for some genres of web documents, such as opinion pieces (Liu et al., 2021), is that the responses are usually implied instead of explicitly written. To address this, an alternative strategy is to use a text generation model to generate the responses, though it introduces additional challenge, such as evaluating the integrity of the text to the original article (Maynez et al., 2020).

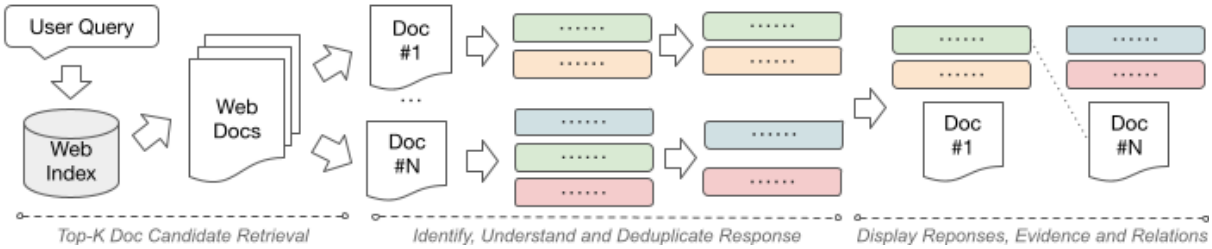

Figure 2: An conceptual illustration of our multi-perspective document retrieval paradigm, as discussed in § 3.5. Instead of document, our retrieval pipeline revolves around *direct responses* within articles. The responses are extracted or generated from individual candidate documents, and then they are re-ranked, de-duplicated and categorized based on the implied perspectives across documents. Finally, the responses from *diverse perspectives*, supporting evidence in context and relation between results are visualized in a parallel fashion.

3.3 Recognizing Different Perspectives in Responses

One key prerequisite for recognizing the perspectives, or semantic implications of response is to break the independence assumption of retrieved documents. When learning about controversial topic, a critical reader would naturally compare responses from various perspectives. We follow this intuition and argue for the need of a module in the document retrieval pipeline to recognize whether two responses share similar or different perspectives in the context of the query.

Given a user query q and a retrieved list of document candidates $fd_i g$, each with zero or more response text segments $fs_{d_i} g$. We want to build a similarity function $Sim(s_{d_i}; s_{d_j} jq; d_i; d_j)$. Compared to traditional sentence-level semantic similarity tasks (Cer et al., 2017; Ganitkevitch et al., 2013), the key difference is the conditional nature of similarity here. For example, in the context of "*Should I ride bike or take taxi to work*", "*Riding bike costs less*" and "*Taxis cost more*" are semantically equivalent, despite the fact that two statements having different predicate arguments.

In practice, apart from the difficulty in developing annotations and resources for such tasks, the bottleneck includes the fact that text segments are not standalone without document as context. For example, text segments may use pronoun references instead of named or nominal mentions exclusively. To address such problems, Choi et al. (2021) proposes a text generation task to de-contextualize an input segment using its document context.

3.4 Trustworthiness

User trust in the results from a search engine is largely governed by the empirical *authoritativeness* of their information sources and content. From the perspective of information sources, its trustworthiness is typically determined by the process it generates and publishes the content. For example, a news source is generally considered credible if all the published content goes through fact-checking and statement biases are carefully removed (Entman, 2007).

However, such information is typically neither transparent nor available from the users perspective. To increase access to information about trustworthiness of information sources, we argue for the need of two alternative strategies. First, meta information about the sources helps user recognize the type of content they are seeing, and how to assess the credibility of the content accordingly. For example, it takes drastically different approach assessing the trust value of a official guidance on COVID-19 published by a government entity, versus a opinion piece from a major news outlet criticizing the guidance. Second, for each identified or generated responses in the context of the user query, we corroborate it with supporting evidence extracted from the document (Aharoni et al., 2014). These two strategies work together to provide user a potential way to visually assess the credibility of the content they see.

3.5 Putting it All Together: A Pipeline Approach

Figure 2 illustrates the resulting conceptual paradigm for a multi-perspective search engine. The paradigm centers around the abstraction of a response to a query, represented by a text segment from a retrieved document. First, a set of document candidates can be retrieved using a general document retrieval system. A response identification module (§ 3.2) extracts and validates responses from individual documents. A cross-document

Figure 3: An example screenshot of results returned by our multi-perspective search prototype.

module (§ 3.3) works to recognize the perspective differences between responses across documents. After corroborating evidences for each response (§ 3.4), the responses categorized by their perspectives; evidence from document context; along with meta information about the sources are displayed in parallel to the user.

4 A Prototype of Multi-Perspective Search Engine

Following the design principles outlined in the previous section, we implement a prototype search engine. Our motivation of the prototype is two-fold. First, we want to understand the utility of our conceptual paradigm from the user perspective. Second, this prototype serves as a platform to study the user information needs with respect to controversial and debate-worthy queries. We defer the discussion of the study to § 5.

4.1 Overview

Our prototype consists of four component. First, given a user query, we use Google Search API to retrieve the list of top-k relevant articles from the web. For each article, we use a query-focused summarization model (Liu et al., 2021) to generate the key response to the query implied by the document. We employ a stance classification model (Chen et al., 2019b) to categorize the responses based on their stance (e.g. pro/con) to the query. Next, we use a evidence extraction model (Ein-Dor et al., 2019) to extract sentences from each article that support the generated perspective of this article. Last, we cluster the articles according to their conditional semantic similarity on the query, and present them to the users with their lists of evidence. Details of each of the components are described in § 4.2

An example search result is shown in Figure 3. The user searches a query of *Should wearing masks be mandatory?*, and our system presents the searched articles to the users in two clusters of perspectives, one supporting the query, and one opposing the query. In each cluster, our system displays articles that share this perspective, along with their evidence retrieved from the articles.

4.2 Components

Candidate Document Retrieval We use Google Programmable Search Engine and Custom Search JSON API[2] to facilitate our retrieval process. We configure our search engine to include a manually curated list of trustworthy websites to search from (See Table 2 in Appendix B for the complete list of sources), and use the API to retrieve the webpages returned by our search engine. We then use Newspaper3k[3] to clean and extract the article body.

Perspective Extraction/Summarization Given a query and an abstract from an article, a perspective summarization system is expected to generate a single sentence thesis statement that for the arti-

[2] https://developers.google.com/custom-search/v1/overview

[3] https://newspaper.readthedocs.io/en/latest/

cle (Liu et al., 2021). This is intended to be used as a dynamic, query-dependent title for the article. In addition, the generated perspective ideally should be relevant in the context of the query, and reflect the same stance as the article towards the query. We use a query-focused multi-task BART summarization model proposed by Liu et al. (2021) to summarize the perspective of each article. We conduct our evaluation on the MultiOpEd dataset (Liu et al., 2021). The best performing perspective summarization model achieves 11.92 Rouge-2 score on the test split of the dataset. We report the selection and performance of the baseline models in Table 1.

Evidence Mining Evidence mining is the task of retrieving supporting evidence from an article to a perspective in the context of the user query. We choose Evidence Sentences dataset (Ein-Dor et al., 2019), which most closely resembles our task. We train a BERT regression model that predicts the probability of a sentence being evidence towards the perspective. For each document and its perspective, we extract a ranked list of sentences with top-k highest output probabilities from the model. As the training objective and data does not pose any constraint on whether the evidence supports or opposes the input perspective, We pipe it with a stance classifier (Chen et al., 2019b) to select only the sentences that support the perspective, and use a relevance classifier (Chen et al., 2019b) to rank them and present the three most relevant evidence paragraphs to the user. We conduct our evaluation on the test split of Evidence Sentences dataset, and our evidence mining system achieved a mean squared error of 3.97%.

Similarity Between Perspectives We use RoBERTa plus a linear layer to train a stance classifier on the Perspectrum dataset (Chen et al., 2019b). We use the model to predict the stance label among {support, refute, or neutral} of each perspective on the query, and group the articles into three clusters in accordance with their stances: support, reject, or neutral. Our stance classifier achieves 91.6% accuracy on the test split of the Perspectrum dataset.

5 A Study on User Information Needs

With the prototype search engine, we revisit the discussions in § 3. Our goal for the user study is to validate our hypothesis on user information need, and discover new insights from users' experience with an actual prototype model.

Task + Dataset	System	Metric	Score
Perspective Summarization (Liu et al., 2021)	BART	ROUGE$_2$	11.34
	+REL		11.51
	+STANCE		11.53
	+REL & STANCE		11.92
Evidence Mining (Ein-Dor et al., 2019)	Linear Regression	MSE	6.39
	SVM		11.71
	GRADIENTBOOST		7.74
	BERT$_{base}$		3.97
Stance Classification (Chen et al., 2019b)	BERT$_{base}$	F1	70.8
	BERT$_{large}$		79.6
	ROBERTA$_{base}$		85.3
	ROBERTA$_{large}$		91.6

Table 1: Evaluation results of the baselines and models on the three components. All the numbers are in percentage. Top performing systems are bold. The setup of different baselines in Perspective Summarization is described more in detailed in (Liu et al., 2021).

5.1 Setup

We collect and construct 30 queries of 6 controversial topics from 2020 United States presidential debate[4]. The 6 topics include COVID-19, Race & Justice, Immigration, Healthcare, Climate change, and Policy. We conduct a survey that compares the search results of these queries returned by Google search with the results returned by our prototype. The complete list of queries can be found in Appendix C.

We use a Google Programmable Search Engine to retrieve the same list of ten articles for each query. Our prototype system would further process the articles and present users with extracted responses, evidence paragraph in different order, i.e. organized based on the predicted stance. For each controversial query, we show ten participants five different questions that compare the retrieved result from Google with our prototype's. The first three questions ask them to choose which system do they think has a better organization; shows a more comprehensive view of the topic; and is more informative in the context of the topic. The last two questions inquire their the general preferences of the system, and ask for their explanations of their choices.

To eliminate user interface (UI) as a confounding factor for the study, we construct a web interface for the survey that presents two sets of results

[4] https://2020election.procon.org/

Figure 4: The survey result of comparing Google's search results to our prototype's on 30 different controversial queries. Each query is answered by 10 participants and has four binary questions that indicate their preference of the system. Our prototype is significantly more preferred than Google Search in Organization, Comprehensiveness, and Preference with more than 99% con dence level. See Appendix A for more details.

with nearly identical UI. A screenshot of the survey interface is shown in Appendix D. We hire participants from Amazon Mechanical Turk to collect responses. We compensate the participants $1 for answering 25 questions in total for ve queries. The compensation rates are determined by estimating the average completion time for the survey.

5.2 Quantitative Analysis

The statistics from our survey responses are shown in Figure 4. 63% of the responses indicate that they generally prefer our prototype to Google Search on the query. 69% and 74% of the responses state that our prototype has a better organization of the results and offers a more comprehensive view of the query. The noticeable advantages of these two aspects align with our assumptions. We believe that users prefer seeing search results in different clusters of perspectives instead of in a list ranked by relevance, and are convinced that such organization presents users a broader and more diverse set of views of the topic. On the other hand, only 55% of the responses indicate that our prototype presents more information on the topic. This meets our expectation as well given that the two systems are presenting the same set of articles for each query.

We perform hypothesis testing to validate that users prefer our prototype over Google Search on all of the aspects. We resample from the distribution of the results of each question by boostrapping and compute the mean and standard deviation of each distribution. We repeat this process for 1000 times and use the average mean and standard deviation as the sample mean and sample standard deviation for each question. We then perform one-tailed z-test for each of the sample distribution. The p-values of all aspects except Informativeness are smaller 0.01, giving us 99% con dence to reject the null hypothesis that less than half of the participants prefer our prototype under on these aspects, and the p-value of Informativeness is 0.043, showing a 95% con dence level in improvements over Google Search.

5.3 Qualitative Analysis

To understand the information needs of participants, we look into the explanations they provided for their general preferences on the system. We nd that some participants prefer the search results provided by our prototype because they seek for direct answers to the query. The search is asking a question, and the results of System 1 (Google) offer only a collection of articles without any real answers. I like that System 2 (Our prototype) lters the results by organizing the "yes" articles together, the "no" articles together, and even the "maybe" articles together. This is a more convenient and helpful way for the user to nd what they're really looking for. This nding con rms our hypothesis that some users expect to see direct answers when searching for a controversial query.

Some other participants prefer the way we present the information by showing bullet points instead of snippets. : prefer system 2 because each result has bullets of information which makes it easier for me to decide if it is worth clicking. Also, it is separated by results that answer yes/no to the question. It shows that some users prefer to see an organized list of key information in an article and use it to decide whether or not the article is worthwhile reading.

Despite our intent of providing a diverse set of perspectives to alleviate selection bias, a small portion of participants express concern that the automation provided by our prototype would in turn create biases comparing to Google's. . prefer system 1 because it gives a more balanced set of articles and gives the impression of being more unbiased, whereas System 2 has a statement right at the beginning that lends one to think of bias. Even when

given the exactly same set of articles, a participant thinks that our prototype gives an impression of being more biased because our system explicitly states the perspective of an article, whereas Google Search results don't show that explicitly and require the users to look into the articles to find out. This finding shows that the information needs of users may vary, and explicitly displaying the stances of articles may have biased connotations.

6 Discussion

The results from the user study offer validation our main hypothesis for the paper, that users typically wish to see direct and diverse responses when issuing a controversial, debate-worthy query to the search engine. At the same time, we also see concerns over potential model mistakes and intrinsic biases, which lead to decreased trust compared to existing document retrieval services.

Surrounding the three design factors introduced in x 3, we discuss a few existing trade-offs in our conceptual paradigm through the lens of technical challenges involved in developing a multi-perspective search system.

Direct Response vs. Trust As the responses act as an extra layer of abstraction over the documents themselves, the question on the trustworthiness of the machine-identified response naturally rises as a concern.

Such concerns involve two conflicting factors of research and design considerations. First, explanation-driven design of natural language understanding tools is beneficial, not only for improvements on model performance, but allows for more efficient design iteration of models. For this reason, explanation-driven task and data design has drawn more attention for information retrieval applications (Lamm et al., 2020).

However, from the user experience point of view, the harm of displaying the wrong explanation might overwhelm the benefit of increased model performance by incorporating explanation. In the case of controversial queries, users are expected to be more sensitive to such trade-offs.

Diverse Response vs. Trust An unfortunate yet existing factor that influences user trust is confirmation bias. For example, Meppelink et al. (2019) discovers in the domain of health information retrieval, the user tend to only agree with results that conform with their prior belief. This raises a question about whether displaying a diverse set of perspectives can influence or correct the potential biases within users' prior belief.

Assessing such hypothesis is out of scope of this current project, and we will defer a study of such to the future. However, as empirical evidence show that such confirmation bias can be attributed in part to the users' access pattern and consumption of news information (Knobloch-Westerwick et al., 2015), we tend to believe that improving how search engines present controversial information will also have a positive impact on alleviating confirmation bias among consumers of news information.

7 Conclusion

This study aims to understand and uncover the user information needs for controversial and debate-worthy queries to search engine. We argue for and examine the need of delivering direct and diverse responses to such queries. To demonstrate the utility of the paradigm, we develop a prototype multi-perspective search engine by synthesizing a suite of relevant NLP and IR tasks. We use the tool to facilitate a user study that confirms the benefit of serving direct and diverse responses to user queries. Through the positive and negative feedback we receive on the prototype, we discuss a few trade-off between technical design and user experience. We hope the findings in the paper offers insights, guidance and opportunities for future development on web document retrieval systems.

Acknowledgments

The authors would like to thank Disha Jindal and Hegler Tissot for their valuable feedback throughout the development of the prototype search engine. In addition, the authors are grateful for the collaboration with 10clouds [5] to research, design and develop the front-end interface for the prototype. This work was supported in part by a Focused Award from Google, and a gift from Tencent.

[5] https://10clouds.com/

References

Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying search results. In Proceedings of the second ACM international conference on web search and data mining, pages 5–14.

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In Proceedings of the first workshop on argumentation mining, pages 64–68.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 251–261, Valencia, Spain. Association for Computational Linguistics.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020. From arguments to key points: Towards automatic argument summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4029–4039.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 621–631.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. arXiv preprint arXiv:1708.00055

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1870–1879.

Sihao Chen, Daniel Khashabi, Chris Callison-Burch, and Dan Roth. 2019a. Perspectroscope: A window to the world of diverse perspectives. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 129–134.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019b. Seeing things from a different angle:discovering diverse perspectives about claims. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 542–557, Minneapolis, Minnesota. Association for Computational Linguistics.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. Transactions of the Association for Computational Linguistics, 9:447–461.

Charles LA Clarke, Nick Craswell, and Ian Soboroff. 2009. Overview of the trec 2009 web track.

Charles LA Clarke, Maheedhar Kolla, Gordon V Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pages 659–666.

Van Dang and Bruce W Croft. 2013. Term level search result diversification. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, pages 603–612.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, and Andrew McCallum. 2018. Multi-step retriever-reader interaction for scalable open-domain question answering. In International Conference on Learning Representations.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. arXiv preprint arXiv:1704.05179

Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. 2019. Corpus wide argument mining – a working solution.

Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, et al. 2020. Corpus wide argument mining-a working solution. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 7683–7691.

Robert M Entman. 2007. Framing bias: Media in the distribution of power. Journal of communication, 57(1):163–173.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 758–764.

William Goffman. 1964. A searching procedure for information retrieval. Information Storage and Retrieval, 2(2):73–78.

Sha Hu, Zhicheng Dou, Xiaojie Wang, Tetsuya Sakai, and Ji-Rong Wen. 2015. Search result diversification based on hierarchical intents. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pages 63–72.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1601–1611.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 252–262.

Silvia Knobloch-Westerwick, Benjamin K Johnson, and Axel Westerwick. 2015. Confirmation bias in online searches: Impacts of selective exposure before an election on political attitude strength and shifts. Journal of Computer-Mediated Communication, 20(2):171–187.

Oleksandr Kolomiyets and Marie-Francine Moens. 2011. A survey on question answering technology from an information retrieval perspective. Information Sciences, 181(24):5412–5434.

Matthew Lamm, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, and Michael Collins. 2020. Qed: A framework and dataset for explanations in question answering.

Siyi Liu, Sihao Chen, Xander Uyttendaele, and Dan Roth. 2021. MultiOpEd: A corpus of multi-perspective news editorials. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4345–4361, Online. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1906–1919.

Corine S Meppelink, Edith G Smit, Marieke L Fransen, and Nicola Diviani. 2019. "i was right about vaccination": Confirmation bias and health literacy in online health information seeking. Journal of health communication, 24(2):129–140.

Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. 2021. Rethinking search: making domain experts out of dilettantes. In ACM SIGIR Forum, volume 55, pages 1–27. ACM New York, NY, USA.

Bing Pan, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. 2007. In google we trust: Users' decisions on rank, position, and relevance. Journal of computer-mediated communication, 12(3):801–823.

Jeff Pasternack and Dan Roth. 2013. Latent credibility analysis. In Proceedings of the 22nd international conference on World Wide Web, pages 1009–1020.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2383–2392.

Oya Y Rieger. 2009. Search engine use behavior of students and faculty: User perceptions and implications for future research. First Monday.

Daniel E Rose and Danny Levinson. 2004. Understanding user goals in web search. In Proceedings of the 13th international conference on World Wide Web, pages 13–19.

Tetsuya Sakai and Ruihua Song. 2011. Evaluating diversified search results using per-intent graded relevance. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pages 1043–1052.

Ben Shneiderman, Don Byrd, and W Bruce Croft. 1997. Clarifying search: A user-interface framework for text searches.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 46–56.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 809–819.

Xiaojie Wang, Ji-Rong Wen, Zhicheng Dou, Tetsuya Sakai, and Rui Zhang. 2017. Search result diversity evaluation based on intent hierarchies. IEEE Transactions on Knowledge and Data Engineering, 30(1):156–169.

Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 2017. End-to-end neural ad-hoc ranking with kernel pooling. In Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval, pages 55–64.

A Hypothesis Testing

Our null hypothesis is $P \quad P_0$ and alternate hypothesis is $P > P_0$, where $P_0 = 0.5$ and P is the percentage of responses of the survey that prefer our prototype.

We use the one-sided Z-Test as our hypothesis testing statistics. The z-score of a standard z-test can be calculated as

$$Z = \frac{X}{\overline{\sqrt{N}}}$$

We perform Boostrapping resampling to get the sample distribution and standard deviation. Specifically, we randomly sample data points with replacement from the distribution of each question's results to a sample size of 300, and repeat this process 1000 times to compute an average mean and average standard deviation. We then regard these mean and standard deviation as the sample mean and sample standard deviation. We then use these statistics to calculate the z-score for each type of question and get $z = 7.29, 9.57, 1.72,$ and 4.79. We show that users prefer our prototype over Google Search with significance levels 0.01, 0.01, 0.05, and 0.01 for questions on *organization*, *comprehensiveness*, *informativeness*, and *general preference*, respectively.

B List of Sources

See Table 2.

C List of Controversial Queries

See next pages.

D Screenshots of Demo Website and Human Study Interface

nytimes.com
theguardian.com
npr.org
reuters.com
bbc.com
cnn.com
washingtonpost.com
nbcnews.com
cbs.com
abc.com
foxnews.com
who.int
clevelandclinic.com
medicalnewstoday.com
noah-health.org
familydoctor.org
medlineplus.gov
mayoclinic.org
webmd.com
cebm.net
sciencenews.org
sciencemag.org
yalemedicine.org
nejm.org

Table 2: List of urls of the sources we include as information source for our prototype multi-perspective search engine.

Topic	Query
COVID-19	Should wearing masks be mandatory? Should we all get vaccinated? Is herd immunity for COVID-19 achievable? Should COVID-19 vaccines be mandatory? Will the Covid-19 pandemic have a lasting impact on society?
Race & Justice	Sshould the US ban assault weapons? Should police departments be defunded, if not abolished? Should the death penalty be allowed? Should the use of private prisons continue? Should the US continue to build a border wall at the US-Mexico border?
Immigration	Will immigrants take american jobs? Should the US end the Deferred Action for Childhood Arrivals (DACA) policy? Should the US decriminalize illegal border crossings? Should immigrants who entered the US illegally be denied a path to citizenship? Should America use a merit-based immigration system?
Health Care	Should we support medicare for all? Does racial inequality persists in health care coverage? Should abortion be legal? Do the prescription drug costs need to be lowered? Should imported prescription drugs be allowed in the U.S.?
Climate Change	Should the US rejoin the Paris climate agreement? Should the US adopt a climate change plan? Should the US expand fossil fuel extraction on public land? Should fighting climate change be a priority? Is climate crisis inevitable?
Policy	Should the US have withdrawn from the Open Skies Treaty? Should the tariffs imposed on China by president Trump be maintained? Should confederate statues be taken down? Should the federal government adopt Net Neutrality Rules? Should the US re-enter a nuclear deal with Iran?

Table 3: List of controversial topic and queries we include in our user study. The topics and queries are selected from the 2020 United States presidential debates. The list of topics are selected from https://2020election.procon.org/.

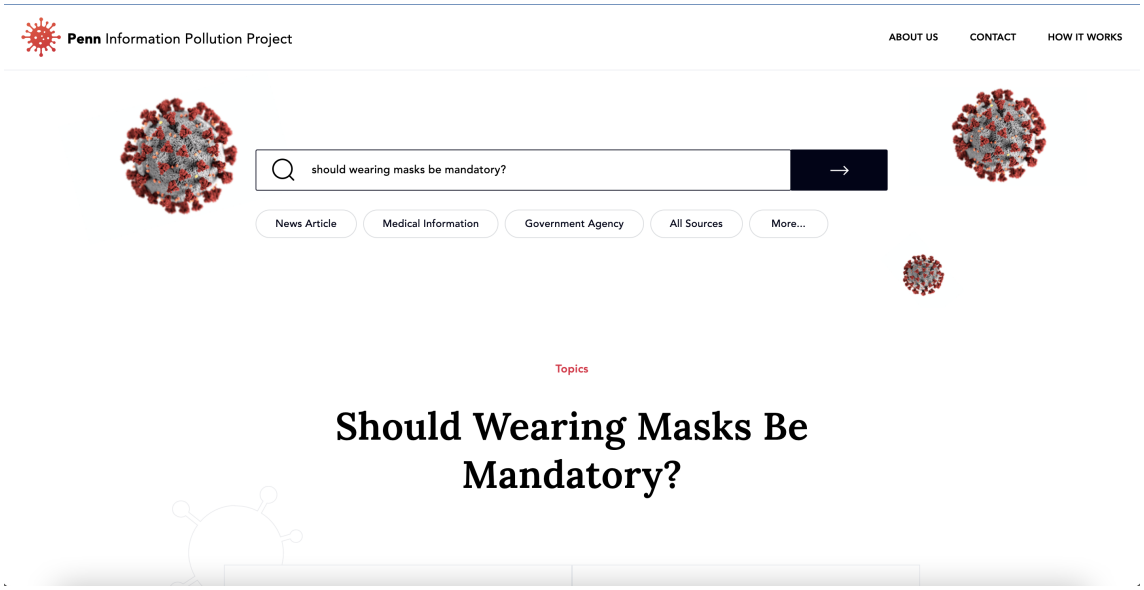

Figure 5: Screenshot 1 of our prototype's demo website.

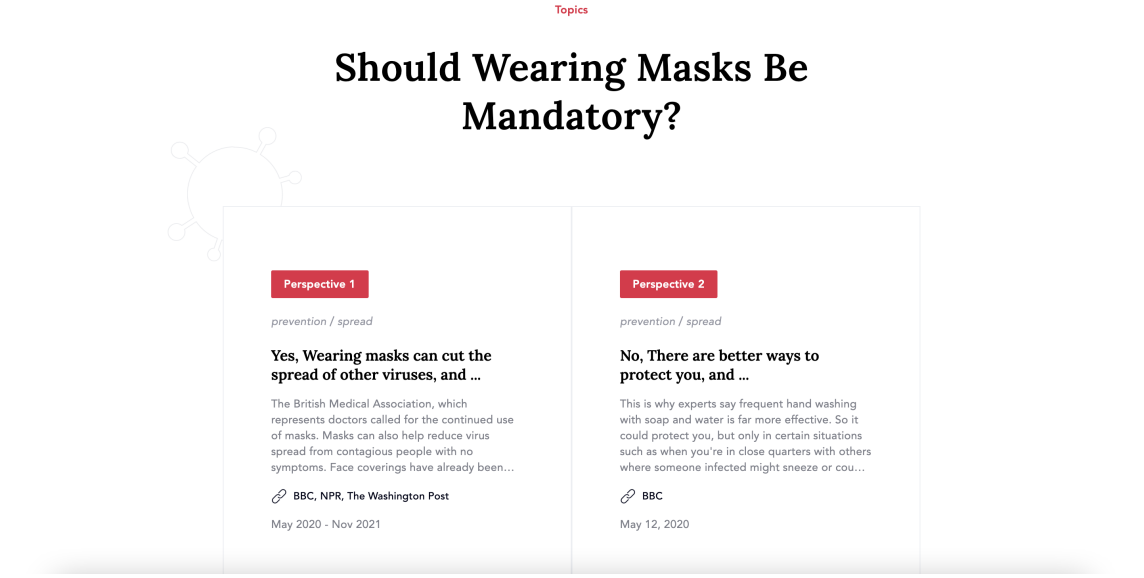

Figure 6: Screenshot 2 of our prototype's demo website.

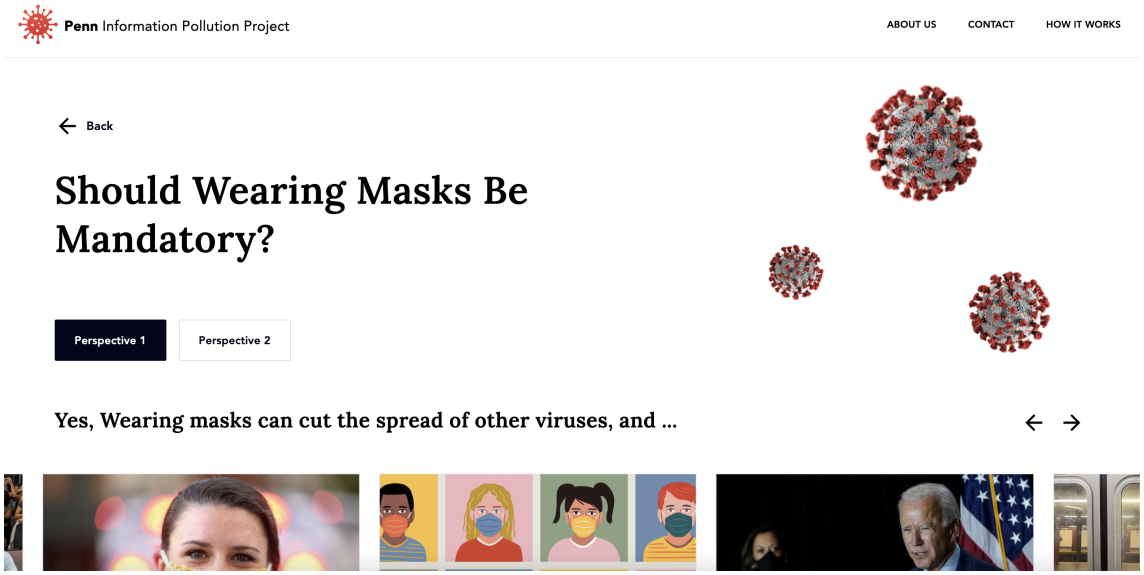

Figure 7: Screenshot 3 of our prototype's demo website.

Figure 8: Screenshot 4 of our prototype's demo website.

No, There are better ways to protect you, and ...

prevention / spread / News Article

It could protect you, but only in certain situations.

12 May 2020 BBC

- This is why experts say frequent hand washing with soap and water is far more effective.
- So it could protect you, but only in certain situations such as when you're in close quarters with others where someone infected might sneeze or cough near your face.
- One is that a mask is not seen as reliable protection as other methods, given that current research shows the virus is spread by droplets and contact with contaminated surfaces.

Figure 9: Screenshot 5 of our prototype's demo website.

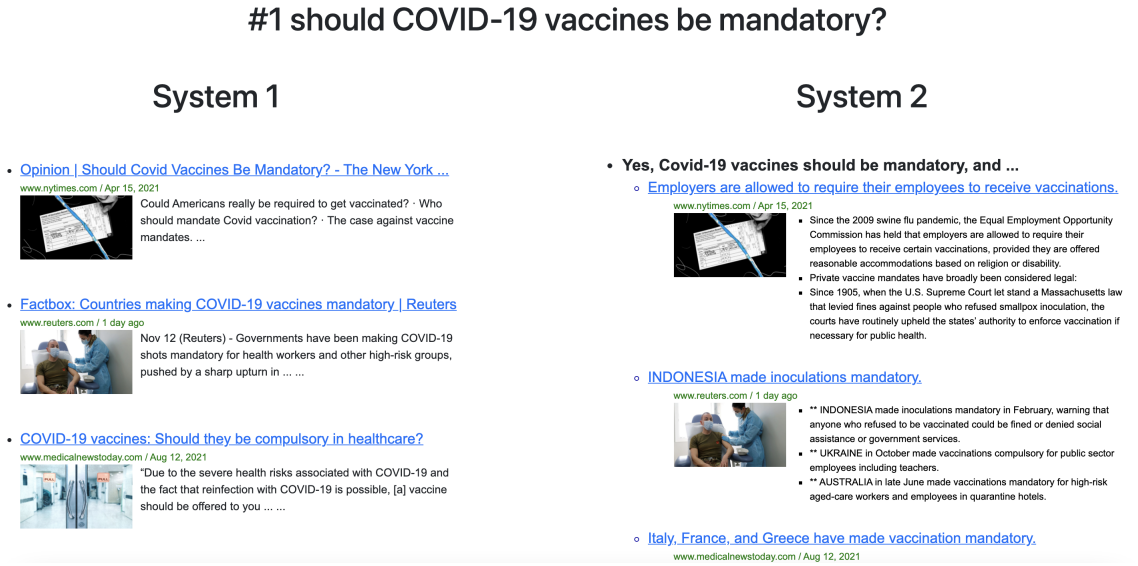

Figure 10: Screenshot 1 of the survey interface.

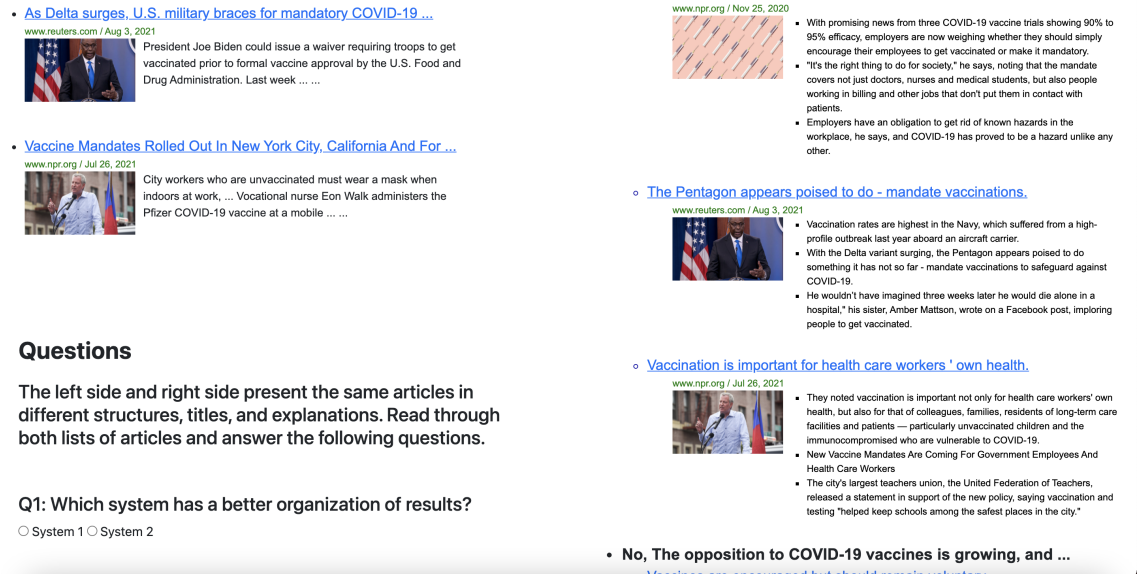

Figure 11: Screenshot 2 of the survey interface.

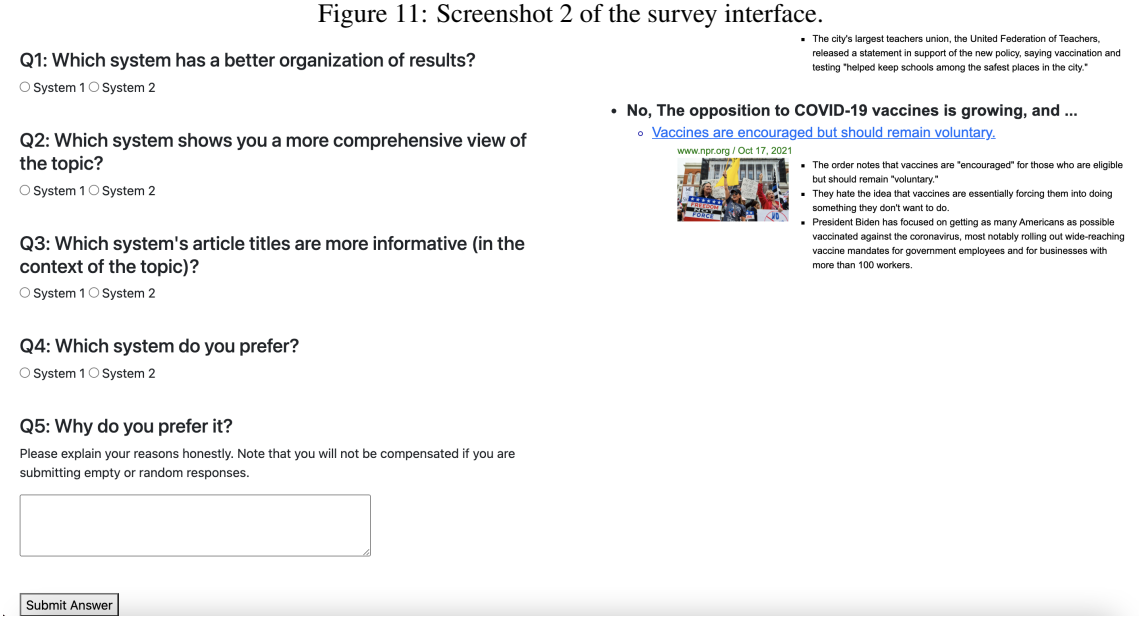

Figure 12: Screenshot 3 of the survey interface.