Don't Raise Your Voice, Improve Your Argument: Learning to Retrieve Convincing Arguments

Sara Salamat[1](✉), Negar Arabzadeh[2], Amin Bigdeli[1], Shirin Seyedsalehi[1], Morteza Zihayat[1], and Ebrahim Bagheri[1]

[1] Toronto Metropolitan University, Toronto, ON, Canada
{sara.salamat,abigdeli,shirin.seyedsalehi, mzihayat,bagheri}@torontomu.ca
[2] University of Waterloo, Waterloo, ON, Canada
narabzad@uwaterloo.ca

Abstract. The Information Retrieval community has made strides in developing neural rankers, which have show strong retrieval effectiveness on large-scale gold standard datasets. The focus of existing neural rankers has primarily been on measuring the relevance of a document or passage to the user query. However, other considerations such as the convincingness of the content are not taken into account when retrieving content. We present a large gold standard dataset, referred to as CoRe, which focuses on enabling researchers to explore the integration of the concepts of convincingness and relevance to allow for the retrieval of relevant yet persuasive content. Through extensive experiments on this dataset, we report that there is a close association between convincingness and relevance that can have practical value in how convincing content are presented and retrieved in practice.

1 Introduction

There has been an increasing attention on mining and identifying argumentative structures from monologues (micro-level) and dialogues (macro-level) in the context of discussion forums and social networks [2,3,12,18–20], which are often referred to as *argument mining*. The works in the argument mining literature explore various tasks including argument detection [3], argument component classification [12], as well as inter and intra argument relation identification [2], to name a few. The major objective of these tasks is to identify arguments, understand their structure and model their relations with each other within a formal argumentation framework [16]. A specific strand of research in this area has focused on identifying, modeling, and predicting *persuasiveness* of arguments. These works are interested in determining what types of arguments and what forms of argumentative structures are capable of convincing the target audience [4,6,8,15,17,26]. There have been a variety of methods that focus on argument persuasion (convincingness) including those that leverage surface textual, social interaction, and argument-related features for ranking arguments [22], as well as others that adopt an end-to-end approach for modeling convincingness [6].

J. Kamps et al. (Eds.): ECIR 2023, LNCS 13981, pp. 589–598, 2023.
https://doi.org/10.1007/978-3-031-28238-6_50

Table 1. Broad areas of related work.

Reference	Task		
	Argument mining	Convincingness	Relevance
[2,6,12,16,19,20]	✓	✗	✗
[4,6,8,15,17,22,26]	✗	✓	✗
[5,21]	✗	✗	✓
Our work	✗	✓	✓

Other researchers have ventured into modeling argument quality [5,21]. While researchers have explored various aspects of argumentative structures, to the best of our knowledge, the *notion of convincingness* of content has not been explored within the context of Information Retrieval (IR). We believe that it is important to understand the process behind the effective retrieval of convincing content because as discussed by Vecchi et al. [20], a careful treatment of such content could be used for social good in areas such as retrieving factual and convincing information for purposes including countering misinformation.

The work in the literature, shown in Table 1, can broadly be classified as those that (1) perform argument mining, (2) measure content convincingness, and (3) determine content relevance. We note that there are no earlier works that have considered the retrieval of convincing information. In other words, retrieval tasks are often focused on optimizing relevance without necessarily taking convincingness of content into account. As such, our work in this paper is among the first to explore how IR ranking models can capture and incorporate the notion of convincingness and integrate it into the retrieval process. Our objective is to rank documents to be *both relevant and convincing*. We systematically curate and publicly release a gold standard of queries and relevant documents, each of which comes with an explicit degree of convincingness. We benefited from the Change My View (CMV) subreddit (`r/changemyview`) in order to capture content convincingness. The CMV subreddit allows users to exchange information with each other on specific topics with the hope of changing each others' opinion, and to explicitly specify how much and to what extent their opinions have changed. We consider content that have changed the opinion of a larger number of users to be more convincing.

Based on the curated dataset, we explore whether it would be possible to learn the notion of convincingness through training different neural ranking models. The idea is that given recent state-of-the-art neural rankers are becoming increasingly better at learning the concept of relevance when shown pairs of queries and their relevant documents, we hypothesize that it might be possible to learn the concept of convincingness by using a similar strategy. Through extensive experiments, we make an important observation that the concepts of relevance and convincingness are (at least on the CMV subreddit) highly correlated phenomena. We find that highly relevant documents to a query are those that are considered to be the most convincing for the users. Our findings align very closely with those of researchers in *cognitive psychology* [13] who have shown

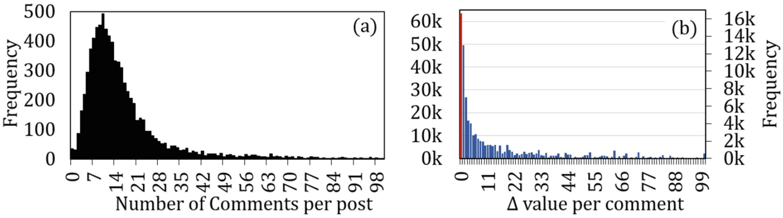

Fig. 1. (a) Distribution of the number of comments per post in CoRe; (b) Distribution of Δ values for the comments. Number of comments with Δ=0 is scaled with the left y-axis and the number of comments with non-zero Δ are scaled with the right y-axis.

that people tend to be convinced more easily when presented with highly relevant information. We show that retrieving documents that are highly relevant would lead to the retrieval of highly convincing ones. This observation suggests that relevance could be a significant contributing factor to convincingness; and therefore, users who would like to persuade others would need to focus their arguments on highly relevant content.

The **contributions of our work** can be summarized as follows: (1) We collect and publicly release a dataset, referred to as *CoRe (Convincing Retrieval)*, which includes 7,937 topics along with subsequent arguments on each topic that have explicit labels for their convincingness at 5 levels; (2) We adopt state-of-the-art neural rankers to learn concepts of relevance and convincingness using our CoRe dataset in order to rank content based on both criteria; (3) We systematically show that the concepts of relevance and convincingness are highly correlated where a retrieval process that maximizes the likelihood of relevance will also be effective for retrieving convincing content.

Reproducibility : The CoRe dataset is publicly available: https://github.com/sara-salamat/CoRe.

2 The Convincing Retrieval (CoRe) Dataset

Most gold standard datasets for the ad hoc retrieval task capture the concept of relevance between a query and its related documents. The objective of our work in this paper is to additionally introduce the concept of convincingness in order to facilitate the process of retrieving relevant and convincing content. To curate such a dataset, we leverage the popular subreddit known as the Change My View subreddit. This subreddit is a community, with over 1.5 million members, on which users post their opinions on a particular topic and challenge others to convince them to change their viewpoints. The community works based on a scoring system, called deltas (Δ), which provides the means to assign credits to convincing arguments. Users are expected to reply to the comment that has changed at least one aspect of their opinion by rewarding it a delta (Δ) and explain how they were convinced to change their opinion [1]. The more convincing a comment is, the more deltas it will receive.

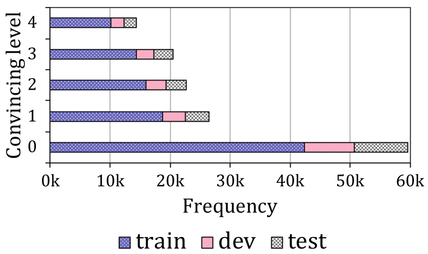

Fig. 2. Convincingness frequency in CoRe.

Table 2. CoRe dataset statistics.

# comments	153,755
# posts	7,937
# users	46,419
Avg length of posts' content	330.93 words
Avg length of posts' title	14.42 words
Avg length of comments	120.27 words

Table 3. CoRe dataset train/dev/test set statistics.

	Train	Dev	Test
Number of posts	5,555	1,189	1,193
Average number of comments	18.26	17.36	18.06
Median number of comments	13	12	13
Average number of Δ per comment	33.58	35.37	33.42
Median number of Δ per comment	1	2	2
Average number Δ per post	613.6	614.19	603.91
Median number Δ per post	515	509	516

In order to gather our CoRe gold standard, we collected all posts and comments published on CMV for a period of 15 months starting from January 2021. To avoid *recency bias*, we did not include any posts that were still active as the deltas on their comments may not have yet reached a steady state. Furthermore, in order to avoid *topical bias*, we did not prioritize the collection of any topics and all content were collected as available on CMV. Table 2 shows the statistics of the content included in our CoRe dataset. For each post, we obtained all of its first level responses and considered them to be the relevant documents for that post. We consider this to be a reasonable assumption since according to CMV rules, any irrelevant responses to the post will be removed by the CMV administrators. As shown in Fig. 1(a), the majority of the posts received between 7 to 14 comments, i.e., the majority of topics in our gold standard have between 7–14 relevant documents. Furthermore, for each of the comments, we collected their delta values whose distribution is depicted in Fig. 1(b). As seen in the Figure, from 153k comments in CoRe, 63k (41%) of these comments did not receive any deltas indicating that no user on CMV considered them to be convincing.

We map delta values into five different levels where comments with no Δ are placed in level zero and are considered not to be convincing at all. The other four levels consist of comments with increasing convincingness with 1–5, 6–20, 21–100 and 100 and more deltas, respectively. We have created splits of the CoRe dataset so it can be used for training neural models by randomly assigning 70% of the posts to the train set, 15% to the development set and 15% to the test set. Table 3 shows the statistics of the data in each split, which have a similar distribution in terms of number of posts, comments and average number of Δ per comment and per post. Additionally in Fig. 2, we depict the frequency of comments placed in the different levels as well as how comments with varying levels of convincingness are placed in different splits. We have ensured that convincingness levels retain a similar ratio in each split. CoRe is structured in TREC format where each CMV post is a query, its first-level comments are its relevant documents, and the convincingness level of each comment is related to its deltas.

3 Evaluation Tasks

We introduce two independent retrieval tasks for the CoRe dataset, namely (1) **relevance ranking**: to retrieve and rank all relevant comments to a post, and (2) **convincingness ranking**: to rank-order the comments of a post based on their degree of convincingness. **The Relevance Ranking Task.** The goal of this task is to perform ad hoc retrieval on the CoRe dataset. Given a query q, the goal of an ad hoc retriever is to use method M to retrieve a ranked list of documents D_q from a collection of items (i.e., C) such that $M(q, C) = D_q$. Given q, D_q is compared to a judged set of items R_q to evaluate the performance of M. In the context of CoRe, each post is considered to be a query, which needs to be satisfied through a retrieval method M based on the set of all comments in the corpus. Each post p is accompanied with a set of comments $C_p = \{C_{p^1}, C_{p^2}, ...C_{p^n}\}$. Given p, all the comments in C_p are considered as relevant, i.e., C_i is only relevant to p if $C_i \in C_p$, and comments not in C_p are considered to be irrelevant to p. The goal of the relevance ranking task is to identify a ranked list of comments D_p for a given post p from a collection of comments using retrieval method M, i.e., $D_p = M(p, C)$.

In order to operationalize M, we employ widely-used bi-encoder-based dense neural retrievers, which have shown promising performance on other tasks [7, 10, 11, 23–25]. Neural rankers need to be trained on a gold dataset. For this purpose, we adopt two strategies: **(1)** In the first strategy, we train the ranker on a completely different relevance judgment dataset, which is non-overlapping with CoRe. The reason for this is that we would like to ensure that the ranker only learns the concept of relevance and does not have a chance to observe the concept of convincingness (as present in CoRe). To this end, we adopt the MS MARCO dataset, which consists of over 500k queries and their relevant judgment documents. **(2)** In the second strategy, we train the ranker on the training split of the CoRe dataset; however, when using this split, we only consider comments that are related to each post as being relevant and ignore the convincingness levels of the comments when training the ranker. The reason for this is that the goal of relevance ranking is to rank comments based on their relevance to the post.

When training the rankers, for each post p, pairs of (p, C_i) are positive samples if $C_i \in C_p$, otherwise, (p, C_i) is a negative sample. The ranker is trained to predict the label for each (p, C_i). We set the maximum sequence length to 300, the number of training epochs to 30 and learning rate to 2e-5. We use Faiss [9] for efficient indexing. Table 5 illustrates the performance of the rankers based on which we make several observations: **(a)** Consistent with earlier findings on base language models for neural rankers, the best performance on relevance ranking is seen when `BERT` is used [14]. **(b)** The first strategy that uses MS MARCO to train the ranker is more effective for relevance ranking, which shows that relevance learnt on a different corpus is transferable to CoRe; and, **(c)** In two of the language models with the largest number of parameters, i.e., `BERT` and `RoBERTa`, the model trained on CMV shows weaker performance compared to one trained on MS MARCO. On the other hand, on the smaller language model, i.e., `DistilBERT`, the model trained on CMV shows better performance. This can be due to the need for a large number of training samples to tune language models with a large number of parameters, i.e., `BERT` and `RoBERTa`.

The Convincingness Ranking Task. The objective of the second task is to learn the concept of convincingness and rank comments according to their degree of convincingness. Formally stated, given a pair of post p, and comment C_i where $C_i \in C_p$, i.e., C_i is a comment related to post p, our goal is to learn the level of convincingness of a pair (p, C_i) while minimizing the difference between the predicted convincingness level through function $S(p, C_i)$ with its actual level of

Table 4. Results on the convincing ranking task.

	Training		Evaluation metric	
Model	Dataset	Task	Recall@10	ndcg@10
`DistilBERT`	MS MARCO	relevance	0.739	0.688
	CoRe	relevance	0.732	0.674
	CoRe	convincing	0.738	0.689
`BERT`	MS MARCO	relevance	**0.741**	0.697
	CoRe	relevance	**0.741**	0.695
	CoRe	convincing	**0.741**	**0.699**
`RoBERTa`	MS MARCO	relevance	0.738	0.684
	CoRe	relevance	**0.741**	0.696
	CoRe	convincing	0.729	0.655

Table 5. Results on relevance retrieval task.

	Training dataset	Recall		MAP		nDCG	
		@10	@100	@10	@100	@10	@100
`DistilBERT`	MS MARCO	0.212	0.394	0.164	0.204	0.384	0.381
	CoRe	0.234	0.454	0.185	0.234	0.414	0.424
`BERT`	MS MARCO	**0.260**	**0.466**	**0.213**	**0.266**	**0.462**	**0.454**
	CoRe	0.236	0.462	0.183	0.233	0.409	0.424
`RoBERTa`	MS MARCO	0.227	0.414	0.183	0.228	0.415	0.404
	CoRe	0.192	0.378	0.144	0.179	0.351	0.357

convincingness. To learn the representation for function S, we adopt two strategies: **(1)** In the first strategy, we benefit from the convincingness levels in our CoRe dataset to learn which comments are convincing in the context of the post. We train a bi-encoder based dense-retriever architecture discussed in the first task to train a model based on comment convincingness levels available in CoRe. In contrast to the first task where there were only two relevance levels, here we are dealing with five levels of convincingness. **(2)** In the second strategy, we use the same neural rankers that were trained for the relevance ranking task to estimate the convincingness of a comment. We use rankers that have learnt the concept of relevance to rank comments based on their convincingness to investigate whether there are any meaningful relationships between the concepts of relevance and convincingness.

Based on results of the convincingness ranking task shown in Table 4, our most notable observation is that regardless of whether the training task was on relevance or convincingness ranking, the results of the convincingness ranking task is similar (0.741) regardless of whether the neural rankers were trained on the MS MARCO or the CoRe datasets. This is an important finding as it shows the neural rankers trained on MS MARCO for relevance ranking are competitive with those rankers trained on CoRe for convincingness. This might indicate that relevance and convincingness are correlated.

4 In-depth Analysis

In order to take an in-depth look into a possible correlation between relevance and convincingness, we first compare the rankings produced by models that were trained on MS MARCO with their counterparts trained on the convincingness levels in CoRe. Then, we compare the rankings produced by both approaches through a *stratified* strategy.

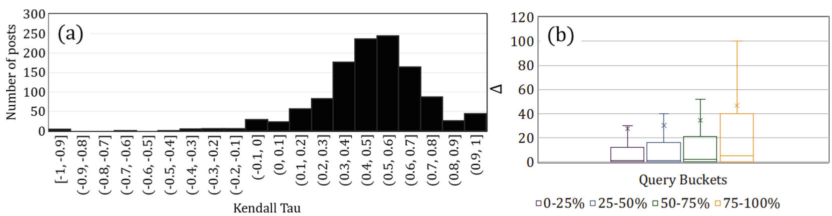

Fig. 3. (a) Distribution of Δ values; (b) distribution of Kendall Tau values.

Association Between Relevance and Convincingness. To assess the degree of association between the two concepts, we compare the retrieved list of comments for a given post when retrieved using the two different strategies, once using rankers trained on relevance and once through rankers trained on convincingness. We employ the Kendall Tau rank correlation to evaluate the correlation between the predicted scores for each of the comments in the retrieved lists for every post. Figure 3(a) presents the histogram of the Kendall Tau correlation values. The Figure shows how correlated the ranked list of comments from the BERT model trained on MS MARCO is to the BERT model trained on CoRe. For each post, the closer the value of Kendall Tau is to one, the higher the correlation between the two retrieved lists would be. From the Figure, we observe that the majority of comments experience a strong correlation (over 0.3), which indicates that the performance of the ranker trained on relevance is quite correlated with a ranker trained on CoRe on an individual post level. This shows that, at least on the CoRe dataset, the concepts of convincingness and relevance are correlated with each other.

Stratified Comparison of Relevance and Convincingness. We study the relationship between the performance of queries' relevance-based retrieval and their comments' convincingness. To do so, we categorize queries into 4 equally-size buckets based on the percentile of their performance (recall@100) where the worst performing queries are in the 0–25% bucket and the 75–100% bucket includes 25% of the highest performing queries. We plot the distribution of deltas associated with the comments on each post under each bucket, which are shown in Fig. 3(b). As shown, the best-performing query bucket, i.e., the yellow bucket, consists of comments with higher degrees of convincingness compared to lower performing query buckets. This finding shows that the easier the query is, the more convincing its comments are and vice versa, i.e., the comments for the hardest queries gain the lowest number of deltas compared to better performing buckets of queries. We find that in CoRe, relevance and convincingness are correlated, which means a ranker that has been effectively trained for relevance retrieval could be an effective out-of-the-box ranker for convincingness retrieval.

5 Concluding Remarks

In this paper, we have introduced the task of convincing IR and offered a systematically collected dataset, called CoRe. The dataset allows the community to explore the retrieval of persuasive content. Based on extensive experiments, we find that the concepts of relevance and convincingness may be correlated, which suggests that, at least in the context of the CMV subreddit, convincing content are those that are relevant to the topic of the query. This reinforces findings in cognitive psychology that indicate people are more likely to be convinced when they are presented with highly relevant content.

References

1. Change my view (cmv) (2018). https://www.reddit.com/r/changemyview/wiki/deltasystem/
2. Chakrabarty, T., Hidey, C., Muresan, S., McKeown, K., Hwang, A.: AMPERSAND: argument mining for PERSuAsive oNline discussions. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 2933–2943. Association for Computational Linguistics, Hong Kong, China, November 2019. https://doi.org/10.18653/v1/D19-1291, https://aclanthology.org/D19-1291
3. Cheng, L., Wu, T., Bing, L., Si, L.: Argument pair extraction via attention-guided multi-layer multi-cross encoding. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 6341–6353 (2021)
4. Dayter, D., Messerli, T.C.: Persuasive language and features of formality on the r/changemyview subreddit. Internet Pragmatics **5**(1), 165–195 (2022)
5. Dumani, L., Schenkel, R.: Quality-aware ranking of arguments. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 335–344 (2020)
6. Dutta, S., Das, D., Chakraborty, T.: Changing views: Persuasion modeling and argument extraction from online discussions. Inf. Process. Manage. **57**(2), 102085 (2020)
7. Gao, J., Xiong, C., Bennett, P., Craswell, N.: Neural approaches to conversational information retrieval. arXiv preprint arXiv:2201.05176 (2022)
8. Habernal, I., Gurevych, I.: Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1589–1599 (2016)
9. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. IEEE Trans. Big Data **7**(3), 535–547 (2019)
10. Karpukhin, V., et al.: Dense passage retrieval for open-domain question answering. arXiv preprint arXiv:2004.04906 (2020)
11. Lin, J., Nogueira, R., Yates, A.: Pretrained transformers for text ranking: bert and beyond. Synth. Lect. Hum. Lang. Technol. **14**(4), 1–325 (2021)
12. Lugini, L., Litman, D.: Contextual argument component classification for class discussions. arXiv e-prints pp. arXiv-2102 (2021)
13. Maio, G.R., Hahn, U., Frost, J.M., Kuppens, T., Rehman, N., Kamble, S.: Social values as arguments: similar is convincing. Front. Psychol. **5**, 829 (2014)
14. Reimers, N., Gurevych, I.: Sentence-bert: sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 (2019)
15. Simpson, E., Gurevych, I.: Finding convincing arguments using scalable bayesian preference learning. Trans. Assoc. Comput. Linguist. **6**, 357–371 (2018)
16. Stab, C., Gurevych, I.: Parsing argumentation structures in persuasive essays. Comput. Linguist. **43**(3), 619–659 (2017)
17. Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., Lee, L.: Winning arguments: interaction dynamics and persuasion strategies in good-faith online discussions. In: Proceedings of the 25th International Conference on World Wide Web, pp. 613–624 (2016)

18. Trabelsi, A., Zaiane, O.R.: Finding arguing expressions of divergent viewpoints in online debates. In: Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM), pp. 35–43 (2014)
19. Tran, N., Litman, D.: Multi-task learning in argument mining for persuasive online discussions. In: Proceedings of the 8th Workshop on Argument Mining, pp. 148–153 (2021)
20. Vecchi, E.M., Falk, N., Jundi, I., Lapesa, G.: Towards argument mining for social good: a survey. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 1338–1352 (2021)
21. Wachsmuth, H., Stein, B., Ajjour, Y.: "Pagerank" for argument relevance. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp. 1117–1127 (2017)
22. Wei, Z., Liu, Y., Li, Y.: Is this post persuasive? ranking argumentative comments in online forum. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 195–200 (2016)
23. Xiong, L., et al.: Approximate nearest neighbor negative contrastive learning for dense text retrieval. CoRR abs/2007.00808 (2020). https://arxiv.org/abs/2007.00808
24. Yang, W., Zhang, H., Lin, J.: Simple applications of BERT for ad hoc document retrieval. CoRR abs/1903.10972 (2019). http://arxiv.org/abs/1903.10972
25. Yu, S., Liu, Z., Xiong, C., Feng, T., Liu, Z.: Few-shot conversational dense retrieval. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 829–838 (2021)
26. Zeng, J., Li, J., He, Y., Gao, C., Lyu, M., King, I.: What Changed Your Mind: The Roles of Dynamic Topics and Discourse in Argumentation Process, p. 1502–1513. Association for Computing Machinery, New York, NY, USA (2020), https://doi.org/10.1145/3366423.3380223