Frame-oriented Summarization of Argumentative Discussions

Shahbaz Syed [†] **Timon Ziegenbein** [‡] **Philipp Heinisch** [♠]
Henning Wachsmuth [‡] **Martin Potthast** [†◇]
[†]Leipzig University [‡]Leibniz University Hannover [♠]Bielefeld University [◇]ScaDS AI
<shahbaz.syed@uni-leipzig.de>

Abstract

Online discussions on controversial topics with many participants frequently include hundreds of arguments that cover different framings of the topic. But these arguments and frames are often spread across the various branches of the discussion tree structure. This makes it difficult for interested participants to follow the discussion in its entirety as well as to introduce new arguments. In this paper, we present a new rank-based approach to extractive summarization of online discussions focusing on argumentation frames that capture the different aspects of a discussion. Our approach includes three retrieval tasks to find arguments in a discussion that are (1) relevant to a frame of interest, (2) relevant to the topic under discussion, and (3) informative to the reader. Based on a joint ranking by these three criteria for a set of user-selected frames, our approach allows readers to quickly access an ongoing discussion. We evaluate our approach using a test set of 100 controversial Reddit ChangeMyView discussions, for which the relevance of a total of 1871 arguments was manually annotated.

1 Introduction

Web-based forums like Reddit facilitate discussions on all kinds of topics. Given the size and scope of some communities (known as "Subreddits"), multiple individuals regularly participate in the discussions of timely controversial topics, such as on ChangeMyView.[1] Notably, the volume of arguments tends to grow substantially in a tree-like response structure wherein each branch forms a concurrent discussion thread. These threads develop in parallel as different perspectives are introduced by the participants. After a discussion subsides, the resulting collection of threads and their arguments often represents a comprehensive overview of the most pertinent perspectives (henceforth, referred to as *frames*) put forth by the participants.

[1](CMV) https://www.reddit.com/r/changemyview/

Frames help shape one's understanding of the topic and deliberating one's own stance (Entman, 1993; Chong and Druckman, 2007). However, in large discussions, prominent arguments as well as the various frames covered may be distributed in arbitrary (and often implicit) ways across the various threads. This makes it challenging for participants to easily identify and contribute arguments to the discussion. Large online forums like Reddit typically provide features that enable the reorganization of posts, for example, based on their popularity, time of creation, or in a question–answer format. A popularity-based ranking may seem beneficial, but Kano et al. (2018) discovered that an argument's popularity is not well correlated with its informativeness. Furthermore, a popularity-based ranking does not cover the breadth of frames of a discussion, as we will show in this paper (Section 4.1).

In this paper, we cast discussion summarization as a ranking task with an emphasis on frame diversity, thereby introducing a new paradigm to discussion summarization in the form of *multiple* summaries per discussion (one per frame). Previous research has focused on creating a single summary per discussion instead (Section 2). As illustrated in Figure 1, we first assign arguments to one or more frames. Next, we re-rank arguments in a frame according to their topic relevance. Additionally, we also rank them based on their informativeness via post-processing. Finally, we fuse these rankings to create the final ranking from which the top-k candidates can be used as an *extractive* summary of the discussion centered around a specific frame.

In our experiments, we explore various state-of-the-art methods to realize the three steps of our approach. Our results suggest that: (1) Utilizing retrieval models together with query variants is an effective method for frame assignment, reducing the reliance on large labeled datasets. Here, our approach outperforms a state-of-the-art supervised baseline. (2) Re-ranking arguments of a frame

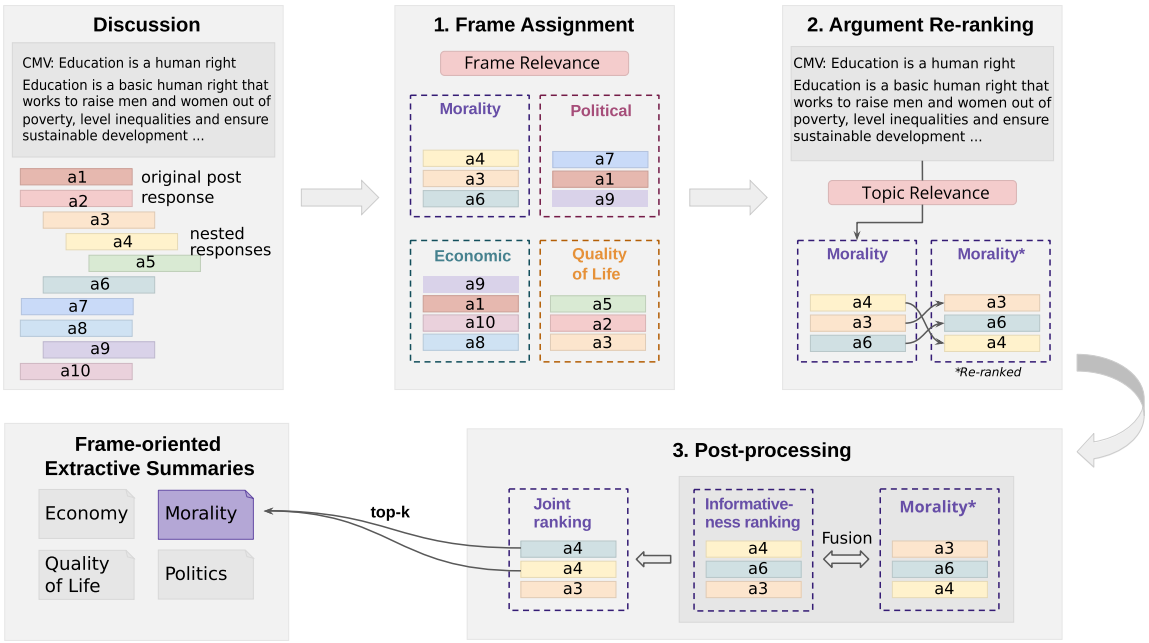

Figure 1: The proposed modular approach to frame-oriented discussion summarization: *1. Frame assignment* assigns arguments to frames ensuring frame relevance. *2. Argument re-ranking* ensures topic relevance of a frame's arguments (here, the *morality* frame is exemplified). *3. Post-processing* fuses the re-ranked arguments with an informativeness ranking. The top-k arguments are then taken as an extractive summary of the discussion.

based on content overlap with the discussion topic is more effective than retrieval-based approaches for ensuring the relevance of the frame's arguments to the topic. (3) Post-processing the argument rankings based solely on content features is insufficient to signal informativeness.

In summary, our contributions include: (1) A fully unsupervised frame assignment approach that assigns one or more frame labels to every argument within a discussion (Section 3.1). (2) An argument retrieval approach that ranks frame-specific arguments based on their topic relevance and informativeness (Section 3.2). (3) A dataset consisting of 1871 arguments sourced from 100 ChangeMyView discussions, where each argument has been judged in terms of frame relevance, topic relevance, and informativeness (Section 4) which forms the basis for an extensive comparative evaluation (Section 5).[2]

2 Related Work

Previous approaches to summarizing discussions can be broadly classified into two categories: *discussion unit extraction* and *discussion unit grouping*. We survey the literature on discussion summarization according to these two categories, followed by the literature on *argument framing*.

[2]Code and data: https://github.com/webis-de/SIGDIAL-23

2.1 Discussion Unit Extraction

Extraction-based approaches use either heuristics or supervised learning to identify important units, such as key phrases, sentences, or arguments within a discussion, then presented as the summary.

Tigelaar et al. (2010) identified several features for identifying key sentences from the discussion, such as the use of explicit author names to detect the response-tree structure, quoted sentences from the preceding arguments, and author-specific features such as participation and talkativity. They found that, while these features can be helpful, summarizing discussions primarily involves balancing coherence and coverage in the summaries. Ren et al. (2011) developed a hierarchical Bayesian model trained on labeled data to track the various topics within a discussion and a random walk algorithm to greedily select the most representative sentences for the summary. Ranade et al. (2013) extracted relevant and sentiment-rich sentences from debates, using lexical features to create indicative summaries. Bhatia et al. (2014) leveraged manually annotated dialogue acts to extract key posts as a concise summary of discussions on question-answering forums (Ubuntu, TripAdvisor). This dataset was further extended with more annotations by Tarnpradab et al. (2017) who proposed a

hierarchical attention network for extractive summarization of forum discussions. Egan et al. (2016) extracted key content from discussions via "point" extraction derived from a dependency parse graph structure, where a point is a verb together with its syntactic arguments.

Closely related to the domain we consider, Kano et al. (2018, 2020) studied the summarization of non-argumentative discussions on Reddit. They found that using the karma scores of posts was not correlated with their informativeness and that combining both local and global context features for comments was the most effective way to identify informative ones. Therefore, we do not rely on karma scores in our post-processing module (Section 4.2) and instead extract several content features for computing informativeness.

The outlined approaches all create a single summary for the entire discussion via end-to-end models. In contrast, we model the extraction of informative arguments organized by frames, thus enabling diverse summaries for a discussion. Furthermore, our experiments with unsupervised retrieval models for frame assignment (Section 4.2) enable us to assess the need to create labeled datasets beforehand to develop strong frame-oriented summarization models tailored to discussions.

2.2 Discussion Unit Grouping

Grouping-based approaches first categorize a discussion's units into explicit (or implicit) classes, such as queries, aspects, topics, dialogue acts, argument facets, or expert-labeled keypoints, and then generate individual summaries for each class. They rely on specific reference points to organize a discussion's units, providing flexibility to the readers by allowing them to choose from diverse summaries that best fit their information needs.

Qiu and Jiang (2013) modeled the discovery of latent viewpoints to group arguments based on two user characteristics: *user identity*, as arguments from the same user are likely to contain the same viewpoint; and *user interaction*, as users with different viewpoints may express disagreement or attack each other, while those with similar viewpoints may support each other. Misra et al. (2015) used summarization to discover repeating arguments and grouped them into facets. Reimers et al. (2019) proposed agglomerative clustering via contextual embeddings to identify similar arguments on a sentence level based on their aspects.

Nguyen et al. (2021) proposed an unsupervised approach to class-specific abstractive summarization of customer reviews with the goal of reducing generic and uninformative content in summaries. They model reviews in the context of topical classes of interest, which are treated as latent variables. These classes represent their reference points as latent variables to be discovered through supervised or reinforcement learning. In contrast, our frame inventory provides a more controlled—and thus more interpretable—set of reference points for discussion summarization. More recently Shapira et al. (2022) proposed a query-assisted, sentence-level interactive summarization approach for news reports using reinforcement learning. Their approach consists of two subtasks of query-based sentence selection and generating query suggestions to enable an interactive setting. In our scenario, we enable this interaction via the predefined set of frames.

Summarizing public debates, Bar-Haim et al. (2020a,b) investigated mapping similar arguments to expert-written key points. Bražinskas et al. (2021) summarized product reviews by selecting subsets of informative reviews, treating the choice of review subset as a latent variable that is learned by a model trained on a dataset compiled from professional product review forums. Amplayo et al. (2021) proposed aspect-controlled opinion summarization via employing multi-instance learning on a labeled dataset to identify aspects in reviews for grouping followed by summarization. The reference points of these approaches are defined either through manual annotations or distant supervision. Some of these reference points are highly topic-specific, requiring them to be created manually for each topic, for instance, the key points from Bar-Haim et al. (2020a). In contrast, we use a fixed and topic-independent set of reference points, namely media frames (Boydstun et al., 2014), grounded in framing theory (Chong and Druckman, 2007).

2.3 Argument Framing

Framing theory was initially utilized to categorize (political) newspaper articles in order to manifest the specifically reported perspective (Neuman et al., 1992; Semetko and Valkenburg, 2006; Boydstun et al., 2014). It was first introduced to the field of argumentation by Naderi and Hirst (2017). Later, Ajjour et al. (2019) modeled framing in argumentation more systematically, introducing automatically extracted, fine-grained, issue-specific frame labels.

Heinisch and Cimiano (2021) successfully combined computational argumentation with framing theory by showing a latent connection between the different frame granularities for the media frames defined by Boydstun et al. (2014). Hartmann et al. (2019) also used frame-labeled data from newswire corpus to successfully train frame classifiers for political discussions via multi-task and adversarial learning. Following the literature, we use the media frames due to their wide adoption in categorizing arguments (Card et al., 2015; Chen et al., 2021).

3 Ranking-based Summarization

This section describes our ranking-based approach to the extractive summarization of online discussions, centered around argumentation frames (Figure 1). First we describe our novel unsupervised approach for frame assignment, followed by methods for re-ranking arguments of a frame based on their relevance to the discussion topic and informativeness. The top-k arguments from the joint ranking are taken as the frame's summary.

3.1 Frame Assignment

Our approach to frame assignment IRFRAME is completely unsupervised in that it employs information retrieval models to rank arguments in a discussion by their *frame relevance*. Here, we consider arguments as documents and frames as queries. This offers a basic and interpretable alternative to frame assignment that does not require labeled data to train supervised models. We investigated both lexical and dense retrieval models.

We used an existing inventory of media frames to organize the arguments in a discussion. This originates from Boydstun et al. (2014) and consists of the 15 frames listed in Table 1. This inventory aims to support an issue-generic frame categorization of political communication. In the context of discussions on Reddit CMV, these issue-generic frames ideally cover a wide variety of controversial topics. The *other* frame is a catch-all category for frames that do not fit into any of the others. We excluded it from our experiments as it is not well-defined, and thus difficult to evaluate. For full frame descriptions see Table 4 in the appendix.

Employing query variants—semantically related queries derived from the primary query—has been shown to improve the retrieval performance (Benham et al., 2019). Thus, we manually created ten query variants for each frame to retrieve and rank all arguments in the discussion based on their frame relevance. Each variant is a high-quality sentence describing the various *aspects* of a frame. We manually curated these sentences from the Wikipedia pages of the frame labels as well as those of the various aspects mentioned in their descriptions (in Table 4). For example, a query variant for the frame *cultural identity* is: "Cultural identity is defined as the identity of a group or culture or of an individual as far as one is influenced by one's belonging to a group or culture and is similar to, and overlaps, with identity politics". The complete list of query variants for all frames is provided in the supplementary material. The output of this module is a ranked list of arguments for each frame, which is then used for extractive summarization (Section 3.2).

Frame Inventory	
Capacity & Resources	Health & Safety
Constitutionality & Jurisprudence	Morality
Crime & Punishment	Policy Prescription & Evaluation
Cultural Identity	Political
Economic	Public Opinion
External Regulation & Reputation	Quality of Life
Fairness & Equality	Security & Defense
	Other

Table 1: Inventory of frames proposed by Boydstun et al. (2014) to track the media framing on policy issues.

We first obtained ten rankings of the arguments (one for each query variant) and then combined these via reciprocal rank fusion (Cormack et al., 2009) to obtain the final list of ranked arguments for a frame. We also compare our approach with a supervised baseline, SUPERFRAME, a classifier finetuned on a set of labeled arguments (details in Section 4.2).

3.2 Extractive Summarization

Building upon the frame assignment component described above that ensures frame relevance, we now perform an *extractive* summarization of the discussion by re-ranking the frame-relevant arguments based on their relevance to the discussion topic and informativeness. This modular approach to summarizing discussions does not require expensive ground-truth summaries, and is thus more scalable than supervised approaches. We first describe the argument re-ranking module followed by the post-processing module.

Argument Re-ranking Besides being relevant to a frame, arguments in the summary must also be relevant to the discussion topic. Thus, we re-rank the frame's arguments according to their *topic relevance*. In our scenario, a "topic" is the combination of the title and the reasoning of the original post on CMV. We propose two approaches for computing topic relevance. The first approach computes content overlap (lexical and semantic) between each argument and the topic. We used Jaccard similarity for lexical overlap, and for semantic overlap, we used the cosine similarity between the contextual sentence embeddings of an argument and the topic. Arguments within a frame are then re-ranked by their overlap scores. The second approach employs retrieval models and (re-)ranks the frame's arguments using the entire topic as the query (details in Section 4).

Post-processing Parallel to the aforementioned re-ranking by topic relevance, we derive a separate re-ranking of the frame's arguments based on their *informativeness*. Our goal is to prioritize content-rich and argumentative texts in the top-k arguments of our approach. We operationalize this through *content scoring* and *argumentativeness scoring*. For content scoring we employed a set of content-specific features such as named entities, noun phrases, the number of discourse markers, and the number of children an argument has in the discussion. Next, for argumentativeness scoring, we trained a topic-based argumentativeness scoring model (details in Section 4). The informativeness score of an argument is the sum of its content score and the argumentativeness score. We then re-rank the frame's arguments by this score.

Frame-oriented Extractive Summaries Given the list of arguments first ranked by frame relevance, then re-ranked by topic relevance, we fuse this ranking with the standalone informativeness ranking from the post-processing module (via reciprocal rank fusion) to derive the final ranking. The top-k arguments from this ranking are taken as the *extractive* summary of the discussion. A key benefit of our ranking-based extractive summarization approach is the flexibility to determine the summary length (i.e., k) by the user according to the discussion's length and their information need. Thus we refrain from setting a specific length budget for the summary.

4 Data and Experiments

This section describes the dataset on which our approach was evaluated, the various retrieval models with their respective parameters, and the content features that we used in our experiments. Also described is the supervised baseline for frame classification SUPERFRAME that we implemented to assign multiple frames to each argument.

4.1 Data

We constructed a dataset of 100 long discussions from CMV, dated January 2020, using the Pushshift Reddit dataset (Baumgartner et al., 2020). For the purpose of this study, we defined a long discussion as a post with at least 100 comments. As preprocessing, we filtered out comments that were deleted by their authors, removed by moderators due to violating community rules, or posted by bots (e.g., DeltaBot, RemindMeBot). The average length of the posts in our dataset is 304 words, with a minimum of 83 words and a maximum of 1611 words. These posts have a total of 25,385 comments, with an average of 253 comments per discussion. The shortest discussion has 105 comments, while the longest has 1066 comments. The average length of a comment is 90 words, with a minimum of 2 words and a maximum of 1589 words excluding the quoted text from either the post or the parent comments they responded to.[3]

Popularity Ranking We investigated to what extent does ranking the arguments only by their popularity (via karma scores on Reddit) cover all the top-k arguments of the frames in the discussion (as assigned by our approach). To quantify this, we computed the mean coverage of the top 10 arguments across all frames and models by their popularity ranking. We considered discussions with at least 500 arguments and ranked them by their popularity scores provided by the Reddit API. Then, at each rank, we computed the percentage of top 10 arguments from all frames that have been covered by the popular arguments. Figure 2 shows that in order to completely cover the top 10 arguments from all frames, a user must read through hundreds of arguments. This encourages us to investigate novel approaches to group arguments in a discussion via

[3]The strict community guidelines of CMV (https://www.reddit.com/r/changemyview/wiki/rules) ensure that comments are primarily argumentative. Therefore, in this paper, we consider each comment to be an argument and do not perform any argument mining.

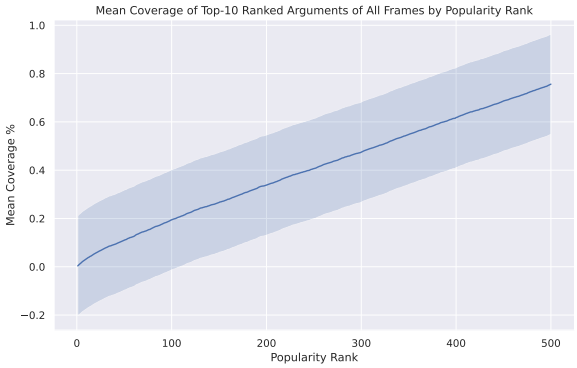

Figure 2: Mean coverage percentage by popularity rank of the top 10 (unique) frame arguments as assigned by our approaches.

frames instead of solely relying on their popularity. A similar conclusion was drawn by Kano et al. (2018) who investigated the effectiveness of popularity scores as a feature for summarizing Reddit discussions.

4.2 Experiments

We first describe the models and parameters for our approaches to frame assignment and extractive summarization. We then describe the supervised baseline for frame assignment.

Frame Assignment We experimented with three retrieval models for IRFRAME to retrieve frame-relevant arguments: BM25 (Robertson et al., 1994), SBERT (Reimers and Gurevych, 2019), and ColBERT (Khattab and Zaharia, 2020). The latter two are dense retrieval models based on contextual embeddings to match arguments to frames, addressing the limitation of BM25 not finding arguments with exact lexical matches to our query variants. We used the Okapi BM25 model with default settings (k=1.5, b=0.75),[4] initialized SBERT with the `all-mpnet-base-v2` model, and used ColBERT-v2 (Santhanam et al., 2022).[5]

Argument Re-ranking We experimented with two approaches to re-rank the arguments retrieved by IRFRAME: content overlap and retrieval-based re-ranking. Content overlap considers both lexical and semantic overlap between the topic and the argument. For lexical overlap, we used Jaccard similarity and for semantic overlap, we used SBERT (`all-mpnet-base-v2` model). For the retrieval-based re-ranking, we experimented with BM25 and ColBERT, with the topic as the query, to (re-)rank the frame's arguments. We excluded SBERT as an additional retrieval model since it is already integrated in the content overlap approach.

Post-processing Informativeness is computed based on the content richness and the argumentativeness of the arguments. Content is scored as the sum of the ratios of named entities, discourse markers, and noun phrases found in the argument and the number of children for an argument in the discussion. We used spaCy (Honnibal et al., 2020) for text tokenization and extraction of the named entities and noun phrases.[6] For discourse markers, we used a lexicon of claim-related words constructed by Levy et al. (2017) for identifying claim-containing sentences. The ratios of named entities and noun phrases were on the token level, while the ratio of discourse markers was on the word level, all normalized by the arguments' lengths. For argumentativeness, we developed *ArgDetector*,[7] a RoBERTa model (Liu et al., 2019) fine-tuned on the dataset by Schiller et al. (2022), containing 150 controversial topics with 144 sentences labeled for their argumentativeness, given the topic. Implementation details are described in Appendix A.

SUPERFRAME This is the supervised baseline for frame assignment. Extending the state-of-the-art frame classification model of Heinisch and Cimiano (2021), we developed a new classifier trained on an external frame-labeled dataset. The existing classifier of Heinisch and Cimiano (2021), utilizes a recurrent neural network to assign a *single* frame to an argument, and combines it with a model that predicts a cluster of frame labels from the inventory of Ajjour et al. (2019) in a multi-task setting. Particularly longer arguments, however, often contain multiple frames. Thus, assigning a single frame to an argument may not be sufficient (Reimers et al., 2019). We therefore extend the model to predict *multiple* frames for an argument. Given the probability distribution of the classification model $P = (p_{f_1}, \cdots, p_{f_k})$ over a set of frames $\mathcal{F} = \{f_1, \ldots, f_k\}$, $k \geq 2$, we apply nucleus sampling (Holtzman et al., 2020) to predict multiple frames for an argument. Specifically, given a cumulative probability mass threshold τ, we assign

[4]We used the Rank BM25 toolkit (Brown, 2020)

[5]We used PyTerrier (Macdonald and Tonellotto, 2020) for the ColBERT pipeline

[6]We used the `en_core_web_md` model.

[7]https://huggingface.co/pheinisch/roberta-base-150T-argumentative-sentence-detector

the minimal subset of frames $F \subseteq \mathcal{F}$ such that:

$$\sum_{f \in F} p_f \geq \tau$$

When the model is very confident in predicting one frame, it is hence likely that an argument is classified to that frame. In cases where the model has lower confidence in its prediction, the argument may consist of multiple frames. This overcomes the limitation of clustering-based approaches and classifiers which strictly assign a single frame to arguments that may contain multiple ones (Reimers et al., 2019; Heinisch and Cimiano, 2021).

To train SUPERFRAME, we used the Media Frames Corpus by Card et al. (2015) consisting of 14,515 news articles with text spans manually annotated for the frame classes in Table 1. Following Heinisch and Cimiano (2021), we trained two variants of the classifier, a *single-task* and a *multi-task* classifier which additionally used the framing dataset by Ajjour et al. (2019) with 12,326 labeled arguments. Both models were based on BiLSTMs, used GloVe embeddings,[8] and trained up to 12 epochs using early stopping. We truncated the input to 75 words with a batch size of 64. To choose between the *single-task* and *multi-task* variants, three of the authors first manually assigned frame(s) for 150 arguments. We then predicted the frames for these arguments using both variants.[9] We opted for higher precision as our goal is to minimize mislabeling arguments with an unrelated frame that can negatively impact the resulting frame-oriented summaries. Since frame assignment is a subjective task (Card et al., 2015) and the boundaries of the frame classes are fuzzy (Reimers et al., 2019; Budzynska et al., 2022), we observed some diversity in our manual annotations. Specifically, we observed that 92% of all the annotated arguments have at least one frame, which was assigned by only a single annotator (minority), indicating different perceptions of observing specific frames in texts. On average, an argument was assigned 3.8 frames (or 1.3 and 0.4 considering the majority and full agreements, respectively).

Table 2 presents the precision scores of both variants with cumulative probability threshold $\tau = 0.9$. Assigning only the most probable frame as predicted by the *single-task* model results in a precision of 59.6% (micro-average) and 49.6% (macro-average), respectively. The *multi-task* model is slightly better at predicting rare frame classes (+0.5% macro-average) but worse at predicting the frequent ones (-7.2% micro-average). Assigning multiple frames per argument increases the effectiveness of the *single-task* model by +0.9% (micro-average), and especially the prediction of rare frame classes, increasing the macro-average prevision by +5.8% (at $\tau = 0.9$).

Model	Minority	Majority	Full
single-task	59.6 / 49.6	**41.7 / 34.1**	**38.8 / 28.8**
single-$\tau = .8$	55.0 / 45.5	32.6 / 27.6	34.8 / 27.6
single-$\tau = .9$	**60.5 / 55.4**	27.8 / 24.5	30.4 / 23.7
multi-task	52.4 / 50.1	27.9 / 22.7	38.4 / 29.5
multi-$\tau = .8$	56.4 / 55.0	33.0 / 26.6	27.4 / 20.1
multi-$\tau = .9$	51.0 / 46.9	26.7 / 21.7	25.4 / 17.9

Table 2: Precision scores (micro / macro %) of the SUPERFRAME model variants at different annotator agreements and thresholds τ for multi-frame prediction.

Considering only the majority-labeled frame classes as ground truth restricts the set of manually assigned frame classes, and hence, reduces the precision scores. On this restricted subset of frame labels, the *single-task* model performs best in nearly all cases, by predicting only the most probable frame class due to the sparsity of the manually assigned frame classes. This variant of the *single-task* model which predicts only a single frame for an argument has a micro-averaged precision of 41.7% and 38.8% in the majority and full agreement scenarios, respectively. Despite this, we extended the *single-task* variant to predict multiple frames per argument, resulting in a high overlap with ground truth frame labels from at least one annotator as well as benefiting from a higher recall. This also avoids having sparse sets of arguments assigned under rare frames.

In conclusion, our internal evaluation supports using the *single-task* model, as opposed to the findings of Heinisch and Cimiano (2021) due to our emphasis on precision while the *multi-task* variant primarily encourages the model in its recall-generalization ability. On average, SUPERFRAME (*single-task* variant) assigned 2.6 frames per argument, with a minimum of 1 and a maximum of 8.2. The frequency counts of all frames in both posts and arguments are shown in Appendix Table 5.

[8] https://nlp.stanford.edu/data/glove.840B.300d.zip

[9] We also experimented with multiple preprocessing methods (e.g. generating a conclusion or ranking the sentences) before automatically predicting the frames. However, these methods negatively impacted the frame prediction.

5 Evaluation

Given that our entire approach is based on retrieval models, we evaluated it manually via relevance judgments. We followed the evaluation style of TREC (Harman, 1993) as best practice. Our evaluation was comprised of judging the *frame relevance*, the *topic relevance*, and the *importance* (in the discussion's context) of arguments retrieved by our models. Following the TREC protocol, we first created 50 evaluation topics, each comprising a post's title, the post itself, and a frame of interest (see supplementary material). To obtain a sufficiently large set of arguments to pool from, we then selected only those discussions for which all models assigned at least 20 arguments to each of the five most frequent frames identified in the comments: *cultural identity*, *economic*, *quality of life*, *public opinion*, and *political* (see Table 5 in the Appendix for the full list). We retrieved arguments for each evaluation topic and performed pooling at depth 5 using TrecTools (Palotti et al., 2019), resulting in 1871 unique arguments to be judged.

5.1 Pilot Study

Multi-annotator relevance judgments can often result in low agreement due to the subjective nature of defining *relevance* and the varying perspectives of annotators (Voorhees, 1998; Bailey et al., 2008; McDonnell et al., 2016; Thomas et al., 2022). Additionally, judges may experience inconsistencies in their decisions as the task progresses (Scholer et al., 2011). To mitigate these issues, we conducted a pilot study with 100 arguments (not included in the main evaluation) to train three annotators and gather feedback for improving the main evaluation interface. The annotators were Computer Science graduates with backgrounds in NLP and IR.

Task Design Following McDonnell et al. (2016), we used a four-point scale for assessing the frame and topic relevance, and the importance of an argument with these options: *definitely not*, *probably not*, *probably*, and *definitely* relevant/important.[10] In assessing importance, we asked annotators to indicate the relevance of an argument to a discussion by answering this question: "How important is the argument to be included in a *summary* of the discussion?". We also experimented with an automatic summary (Nathan, 2016) for long arguments to reduce the cognitive load of the annotators. They were instructed to use the summary if they found it helpful, otherwise to read the entire argument (for details, see Appendix B, Figure 3).

Pilot Agreement and Feedback We measured the inter-annotator agreement (IAA) for the three evaluated criteria using Krippendorff's α, similar to Card et al. (2015). The resulting α values were 0.22 for frame relevance, 0.33 for topic relevance, and 0.22 for importance, respectively. While the agreement is thus limited, the values are consistent with the findings of Card et al. (2015) in their annotation of frame-relevant text spans for the Media Frames Corpus, particularly the frame relevance α value. From feedback, we improved the task design for the main evaluation. Firstly, we removed the automatic summary for each argument since it did not provide significant help. Secondly, we rephrased the importance question to "How important is the argument to be included in the *discussion* of the given topic?" to make it more straightforward, since we did not have ground-truth summaries of the discussions at hand. Annotators also reported that assessing the relevance of an argument for a *single* frame was too restrictive, since an argument may belong to multiple frames, which aligns with the observations of Card et al. (2015). Therefore, we allowed them to assign multiple frames to an argument if the currently-assigned one was not relevant. Accordingly, we proceeded with the main evaluation by assigning each annotator an independent set of arguments to judge. This allowed us to collect more relevance judgments while ensuring a certain level of *shared* understanding of the task.

5.2 Main Evaluation Results

The evaluated models are shown in Table 3.[11] We obtained relevance judgments for a total of 1871 arguments and calculated nDCG@5 (Järvelin and Kekäläinen, 2002) as the effectiveness measure (mean over all topics). Described below are the key findings for each module of our ranking-based extractive summarization framework.

Frame Relevance Our frame assignment approach (*IRFr* with BM25) outperforms other models for identifying frame-relevant arguments in a

[10] We mapped these labels to numerical values ranging from 0 (*definitely not* relevant/important) to 3 (*definitely* relevant/important) for computing nDCG scores.

[11] Model names in Table 3 shortened for brevity. SUPERFRAME → *SupFr* denotes the baseline, IRFRAME → *IRFr* denotes our frame assignment approach, Argument Reranking → *_rr* (via overlap and retrieval models), and Postprocessing → *_post*

discussion with an nDCG@5 of 0.573. Among the retrieval models, BM25 performs better than SBERT and ColBERT, also for re-ranking by topic relevance. Upon further inspection, we found that BM25 often retrieves longer arguments compared to the embedding-based SBERT and ColBERT models. This may provide annotators with more context for informed judgments compared to the shorter arguments. Given the computational costs of running dense retrieval models in real-time, it is promising that a relatively simple and explainable model performs well on our query variants. For the baseline (*SupFr*), combinations with argument re-ranking (via BM25 and topic overlap) also perform reasonably well. However, as various query variants can be easily designed, our *IRFr* approach is more flexible and can be adapted to other domains and topics without the need for labeled data.

Topic Relevance Argument re-ranking by overlap (**_rr_overlap*) outperforms retrieval models for ensuring topic relevance of a frame's arguments. This benefits both *IRFr* and *SupFr* frame assignment approaches with an nDCG@5 scores of 0.847 and 0.785 for the top two models, respectively. Among the retrieval models, BM25 slightly outperforms ColBERT. Given the intuitive nature of content overlap, we conclude that it is favorable to use for re-ranking arguments in a frame.

Importance None of the post-processed models (using informativeness) appear in the top-5 for ranking arguments by importance in the context of the discussion. Instead, argument re-ranking by topic relevance performs best, with nDCG@5 of 0.381 combined with *SupFr* for frame assignment. This contradicts our intuition of post-processing to promote important arguments in the final ranking. As future work, we plan to investigate using context features of the arguments (Kano et al., 2018), as well as pairwise judgments for importance (Zopf, 2018; Luo et al., 2022).

Model	**nDCG@5**		
	Frame	Topic	Imp.
Our Approach			
IRFr_BM25	**0.573**[1]	0.708	0.375[2]
IRFr_SBERT	0.480	0.525	0.303
IRFr_ColBERT	0.522	0.659	0.361[3]
IRFr_BM25_rr_BM25	0.516	0.781[3]	0.349
IRFr_BM25_rr_overlap	0.560[2]	**0.847**[1]	0.350[5]
IRFr_BM25_rr_ColBERT	0.540[4]	0.761	0.358[4]
IRFr_BM25_rr_BM25_post	0.489	0.735	0.297
IRFr_BM25_rr_overlap_post	0.522	0.755	0.339
IRFr_BM25_rr_ColBERT_post	0.526	0.719	0.325
Supervised Baseline			
SupFr_rr_BM25	0.545[3]	0.765[4]	**0.381**[1]
SupFr_rr_overlap	0.536[5]	0.785[2]	0.334
SupFr_rr_ColBERT	0.529	0.764[5]	0.348
SupFr_rr_BM25_post	0.493	0.714	0.322
SupFr_rr_overlap_post	0.493	0.734	0.348
SupFr_rr_ColBERT_post	0.487	0.709	0.329

Table 3: nDCG@5 for the manual relevance judgments for frame relevance, topic relevance, and importance. The best results for each evaluated criterion are highlighted in bold, alongside the rankings for the five best models. We evaluated our frame assignment approach (*IRFr*) against the supervised baseline (*SupFr*), combined with our argument re-ranking (*_rr*) and post-processing components (*_post*). We see that our approach to frame assignment results in the best models for frame and topic relevance and is also competitive for argument importance.

6 Conclusion and Future Work

We introduced a novel ranking-based approach to frame-oriented (extractive) discussion summarization in web-based forums, aiming to enhance the accessibility and comprehension of large-scale online discussions for participants. Our approach involves three key steps: frame assignment, argument re-ranking, and post-processing. Specifically, we developed unsupervised methods for both frame and topic assignment leveraging standard retrieval models. Extensive experiments on a dataset of 1871 arguments from 100 ChangeMyView discussions demonstrate the effectiveness of our approach in ensuring frame and topic relevance in the summary, outperforming a state-of-the-art supervised baseline for frame assignment. Nevertheless, further exploration is needed to enhance summary informativeness through post-processing.

In the future, we plan to develop practical applications that leverage our approach for scalable exploration of online discussions guided by argumentation frames. Moreover, we will explore the application of our approach to summarize discussions in various Subreddits beyond ChangeMyView and across different debate portals.

References

Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019. Modeling frames in argumentation. 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)*, pages 2915–2925, Hong Kong, China. Association for Computational Linguistics.

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Aspect-controllable opinion summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6578–6593, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P. de Vries, and Emine Yilmaz. 2008. Relevance assessment: are judges exchangeable and does it matter. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, July 20-24, 2008*, pages 667–674. ACM.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020a. From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4029–4039. Association for Computational Linguistics.

Roy Bar-Haim, Yoav Kantor, Lilach Eden, Roni Friedman, Dan Lahav, and Noam Slonim. 2020b. Quantitative argument summarization and beyond: Cross-domain key point analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 39–49. Association for Computational Linguistics.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, pages 830–839. AAAI Press.

Rodger Benham, Joel M. Mackenzie, Alistair Moffat, and J. Shane Culpepper. 2019. Boosting search performance using query variations. *ACM Trans. Inf. Syst.*, 37(4):41:1–41:25.

Sumit Bhatia, Prakhar Biyani, and Prasenjit Mitra. 2014. Summarizing online forum discussions – can dialog acts of individual messages help? In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2127–2131, Doha, Qatar. Association for Computational Linguistics.

Amber E. Boydstun, Dallas Card, Justin Gross, Paul Resnick, and Noah A. Smith. 2014. Tracking the development of media frames within and across policy issues.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2021. Learning opinion summarizers by selecting informative reviews. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9424–9442, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dorian Brown. 2020. Rank-BM25: A Collection of BM25 Algorithms in Python.

Katarzyna Budzynska, Chris Reed, Manfred Stede, Benno Stein, and Zhang He. 2022. Framing in communication: From theories to computation (dagstuhl seminar 22131). *Dagstuhl Reports*, 12(3):117–140.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Wei-Fan Chen, Khalid Al Khatib, Benno Stein, and Henning Wachsmuth. 2021. Controlled neural sentence-level reframing of news articles. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2683–2693, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dennis Chong and James N. Druckman. 2007. *Framing theory*, Annual Review of Political Science, pages 103–126.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, pages 758–759. ACM.

Charlie Egan, Advaith Siddharthan, and Adam Wyner. 2016. Summarising the points made in online political debates. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 134–143, Berlin, Germany. Association for Computational Linguistics.

Robert M Entman. 1993. Framing: Towards clarification of a fractured paradigm. *McQuail's reader in mass communication theory*, 390:397.

Donna Harman. 1993. Overview of the second text retrieval conference (TREC-2). In *Proceedings of The Second Text REtrieval Conference, TREC 1993,*

Gaithersburg, Maryland, USA, August 31 - September 2, 1993, volume 500-215 of *NIST Special Publication*, pages 1–20. National Institute of Standards and Technology (NIST).

Mareike Hartmann, Tallulah Jansen, Isabelle Augenstein, and Anders Søgaard. 2019. Issue framing in online discussion fora. 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 1401–1407, Minneapolis, Minnesota. Association for Computational Linguistics.

Philipp Heinisch and Philipp Cimiano. 2021. A multi-task approach to argument frame classification at variable granularity levels. *it - Information Technology*, 63(1):59–72.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Ryuji Kano, Yasuhide Miura, Motoki Taniguchi, Yan-Ying Chen, Francine Chen, and Tomoko Ohkuma. 2018. Harnessing popularity in social media for extractive summarization of online conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1139–1145, Brussels, Belgium. Association for Computational Linguistics.

Ryuji Kano, Yasuhide Miura, Tomoki Taniguchi, and Tomoko Ohkuma. 2020. Identifying implicit quotes for unsupervised extractive summarization of conversations. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 291–302, Suzhou, China. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 39–48. ACM.

Ran Levy, Shai Gretz, Benjamin Sznajder, Shay Hummel, Ranit Aharonov, and Noam Slonim. 2017. Unsupervised corpus–wide claim detection. In *Proceedings of the 4th Workshop on Argument Mining*, Copenhagen, Denmark. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ge Luo, Hebi Li, Youbiao He, and Forrest Sheng Bao. 2022. Prefscore: Pairwise preference learning for reference-free summarization quality assessment. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 5896–5903. International Committee on Computational Linguistics.

Craig Macdonald and Nicola Tonellotto. 2020. Declarative experimentation in information retrieval using pyterrier. In *ICTIR '20: The 2020 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Norway, September 14-17, 2020*, pages 161–168. ACM.

Tyler McDonnell, Matthew Lease, Mücahid Kutlu, and Tamer Elsayed. 2016. Why is that relevant? collecting annotator rationales for relevance judgments. In *Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2016, 30 October - 3 November, 2016, Austin, Texas, USA*, pages 139–148. AAAI Press.

Amita Misra, Pranav Anand, Jean E. Fox Tree, and Marilyn Walker. 2015. Using summarization to discover argument facets in online idealogical dialog. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 430–440, Denver, Colorado. Association for Computational Linguistics.

Nona Naderi and Graeme Hirst. 2017. Classifying frames at the sentence level in news articles. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 536–542, Varna, Bulgaria. INCOMA Ltd.

Paco Nathan. 2016. PyTextRank, a Python implementation of TextRank for phrase extraction and summarization of text documents.

W Russell Neuman, Russell W Neuman, Marion R Just, and Ann N Crigler. 1992. *Common knowledge: News and the construction of political meaning*. University of Chicago Press.

Thi Nhat Anh Nguyen, Mingwei Shen, and Karen Hovsepian. 2021. Unsupervised class-specific abstractive summarization of customer reviews. In *Proceedings of The 4th Workshop on e-Commerce and NLP*, pages 88–100, Online. Association for Computational Linguistics.

Joao Palotti, Harrisen Scells, and Guido Zuccon. 2019. Trectools: an open-source python library for information retrieval practitioners involved in trec-like campaigns. SIGIR'19. ACM.

Minghui Qiu and Jing Jiang. 2013. A latent variable model for viewpoint discovery from threaded forum posts. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1031–1040, Atlanta, Georgia. Association for Computational Linguistics.

Sarvesh Ranade, Jayant Gupta, Vasudeva Varma, and Radhika Mamidi. 2013. Online debate summarization using topic directed sentiment analysis. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, WISDOM 2013, Chicago, IL, USA, August 11, 2013*, pages 7:1–7:6. ACM.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578, Florence, Italy. Association for Computational Linguistics.

Zhaochun Ren, Jun Ma, Shuaiqiang Wang, and Yang Liu. 2011. Summarizing web forum threads based on a latent topic propagation process. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, page 879–884, New York, NY, USA. Association for Computing Machinery.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume 500-225 of *NIST Special Publication*, pages 109–126. National Institute of Standards and Technology (NIST).

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. Colbertv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3715–3734. Association for Computational Linguistics.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2022. On the effect of sample and topic sizes for argument mining datasets.

Falk Scholer, Andrew Turpin, and Mark Sanderson. 2011. Quantifying test collection quality based on the consistency of relevance judgements. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, pages 1063–1072. ACM.

Holli A. Semetko and Patti M. Valkenburg Valkenburg. 2006. Framing European politics: A Content Analysis of Press and Television News. *Journal of Communication*, 50(2):93–109.

Ori Shapira, Ramakanth Pasunuru, Mohit Bansal, Ido Dagan, and Yael Amsterdamer. 2022. Interactive query-assisted summarization via deep reinforcement learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2551–2568, Seattle, United States. Association for Computational Linguistics.

Sansiri Tarnpradab, Fei Liu, and Kien A Hua. 2017. Toward extractive summarization of online forum discussions via hierarchical attention networks. In *The Thirtieth International Flairs Conference*.

Paul Thomas, Gabriella Kazai, Ryen White, and Nick Craswell. 2022. The crowd is made of people: Observations from large-scale crowd labelling. In *CHIIR '22: ACM SIGIR Conference on Human Information Interaction and Retrieval, Regensburg, Germany, March 14 - 18, 2022*, pages 25–35. ACM.

Almer S. Tigelaar, Rieks op den Akker, and Djoerd Hiemstra. 2010. Automatic summarisation of discussion fora. *Nat. Lang. Eng.*, 16(2):161–192.

Ellen M. Voorhees. 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, pages 315–323. ACM.

Markus Zopf. 2018. Estimating summary quality with pairwise preferences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1687–1696. Association for Computational Linguistics.

A Argumentativeness Scoring

The dataset from Schiller et al. (2022) consists of topics formulated as phrases as opposed to the topic titles in CMV, which are often formulated as claims. To unify this, we manually transformed their topics by appending them with stance-indicative phrases (e.g., "Abortion" → "Abortion should be banned"). We trained the RoBERTa model for the binary classification task with default training parameters: a learning rate of 5e-5, 5% of the training data for

Frame	Description
Capacity & Resources	The lack of or availability of physical, geographical, spatial, human, and financial resources, or the capacity of existing systems and resources to implement or carry out policy goals.
Constitutionality & Jurisprudence	The constraints imposed on or freedoms granted to individuals, government, and corporations via the Constitution, Bill of Rights and other amendments, or judicial interpretation. This deals specifically with the authority of government to regulate, and the authority of individuals/corporations to act independently of government.
Crime & Punishment	Specific policies in practice and their enforcement, incentives, and implications. Includes stories about enforcement and interpretation of laws by individuals and law enforcement, breaking laws, loopholes, fines, sentencing and punishment. Increases or reductions in crime.
Cultural Identity	The social norms, trends, values and customs constituting culture(s), as they relate to a specific policy issue.
Economic	The costs, benefits, or monetary/financial implications of the issue (to an individual, family, community or to the economy as a whole).
External Regulation & Reputation	A country's external relations with another nation; the external relations of one state with another; or relations between groups. This includes trade agreements and outcomes, comparisons of policy outcomes or desired policy outcomes.
Fairness & Equality	Equality or inequality with which laws, punishment, rewards, and resources are applied or distributed among individuals or groups. Also the balance between the rights or interests of one individual or group compared to another individual or group.
Health & Safety	Healthcare access and effectiveness, illness, disease, sanitation, obesity, mental health effects, prevention of or perpetuation of gun violence, infrastructure and building safety.
Morality	Any perspective—or policy objective or action (including proposed action)— that is compelled by religious doctrine or interpretation, duty, honor, righteousness or any other sense of ethics or social responsibility.
Policy Prescription & Evaluation	Particular policies proposed for addressing an identified problem, and figuring out if certain policies will work, or if existing policies are effective.
Political	Any political considerations surrounding an issue. Issue actions or efforts or stances that are political, such as partisan filibusters, lobbyist involvement, bipartisan efforts, deal-making and vote trading, appealing to one's base, mentions of political maneuvering. Explicit statements that a policy issue is good or bad for a particular political party.
Public Opinion	References to general social attitudes, polling and demographic information, as well as implied or actual consequences of diverging from or getting ahead of public opinion or polls.
Quality of Life	The effects of a policy, an individual's actions or decisions, on individuals' wealth, mobility, access to resources, happiness, social structures, ease of day-to-day routines, quality of community life, etc.
Security & Defense	Security, threats to security, and protection of one's person, family, in-group, nation, etc. Generally an action or a call to action that can be taken to protect the welfare of a person, group, nation sometimes from a not yet manifested threat.
Other	Any frames that do not fit into the above categories.

Table 4: Descriptions of frames as per Boydstun et al. (2014). We substituted the term "policy" with the phrase "actions/decisions" to align the frame definitions with the individualistic style of arguments in CMV. Similarly, in *External Regulation & Reputation*, we substituted "United States" with "country" to generalize it.

warmup, early stopping, and a batch size of 32. On the test split provided by Schiller et al. (2022), our fine-tuned model performs with a macro-F1 of 67%, which is comparable with the results from the best model reported in Schiller et al. (2022). A text is labeled as argumentative if the output probability from the finetuned classifier is higher than 50%. Given an input text and the discussion topic we take the mean scores of its constituent sentences as the text's argumentativeness score.

Posts		Comments	
Frame	**Count**	**Frame**	**Count**
Cultural Identity	53	Cultural Identity	13,540
Quality of Life	37	Economic	8931
Economic	33	Quality of Life	8559
Public Opinion	26	Public Opinion	7257
Health & Safety	22	Political	5177
Political	19	Health & Safety	4927
Morality	12	Morality	4237
Policy Prescription & Evaluation	10	Policy Prescription & Evaluation	4108
Fairness And Equality	10	Constitutionality & Jurisprudence	3226
Constitutionality & Jurisprudence	9	Fairness & Equality	2457
Security & Defense	1	Crime & Punishment	898
Crime & Punishment	1	Security & Defense	515
		External Regulation & Reputation	216
		Capacity & Resources	169

Table 5: Counts of frames in posts and comments in our dataset of 100 discussions as predicted by `SuperFrame`. Since each text can be assigned multiple frames, the counts include duplicates. Here, we observe that there are two additional frames found in the comments: *External Reputation & Regulation*, *Capacity & Resources* that are not found in the posts.

B Annotation Interface

Annotation interfaces for the pilot study and the main evaluation are shown in Figures 3 and 4, respectively. We improved the interface for our main evaluation based on annotator feedback from the pilot study with the following changes: (1) We substituted "probably" with "rather" in our scales to indicate a clearer relevance judgment. (2) For non-argumentative texts or meta-arguments (e.g. "I agree.", "I don't understand what you mean." etc.), we allowed annotators to mark the text as *noisy* and skip it. (3) We asked annotators to select at least one relevant frame if the current frame was (definitely/rather) not relevant, with the possibility of selecting multiple frames if required.

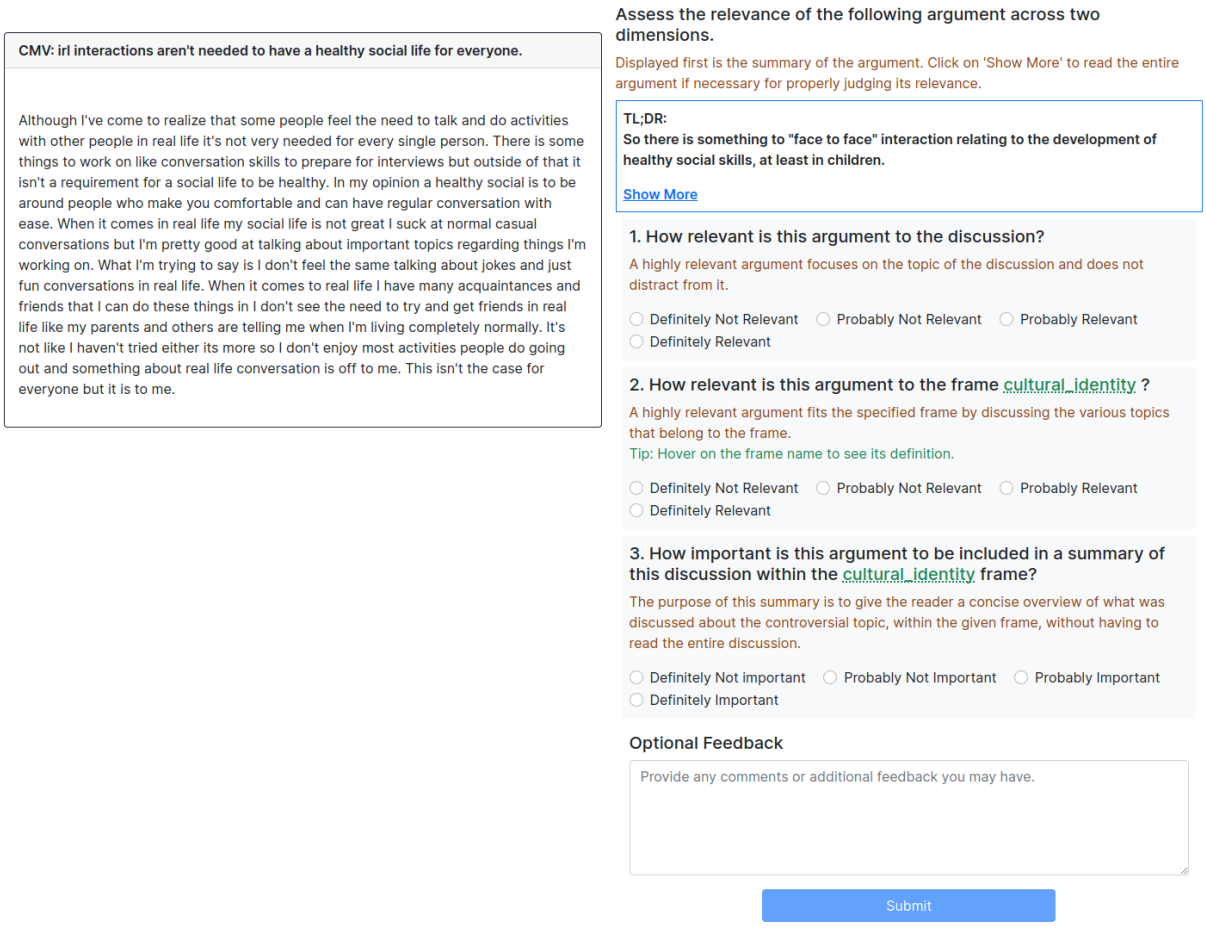

Figure 3: Annotation interface for the **pilot study**. Annotators were provided a summary of the argument alongside the entire argument. There was no option to mark a text as noisy/non-argumentative. Furthermore, the importance of an argument was assessed based on how likely it was to be included in a frame-oriented *summary* of the discussion.

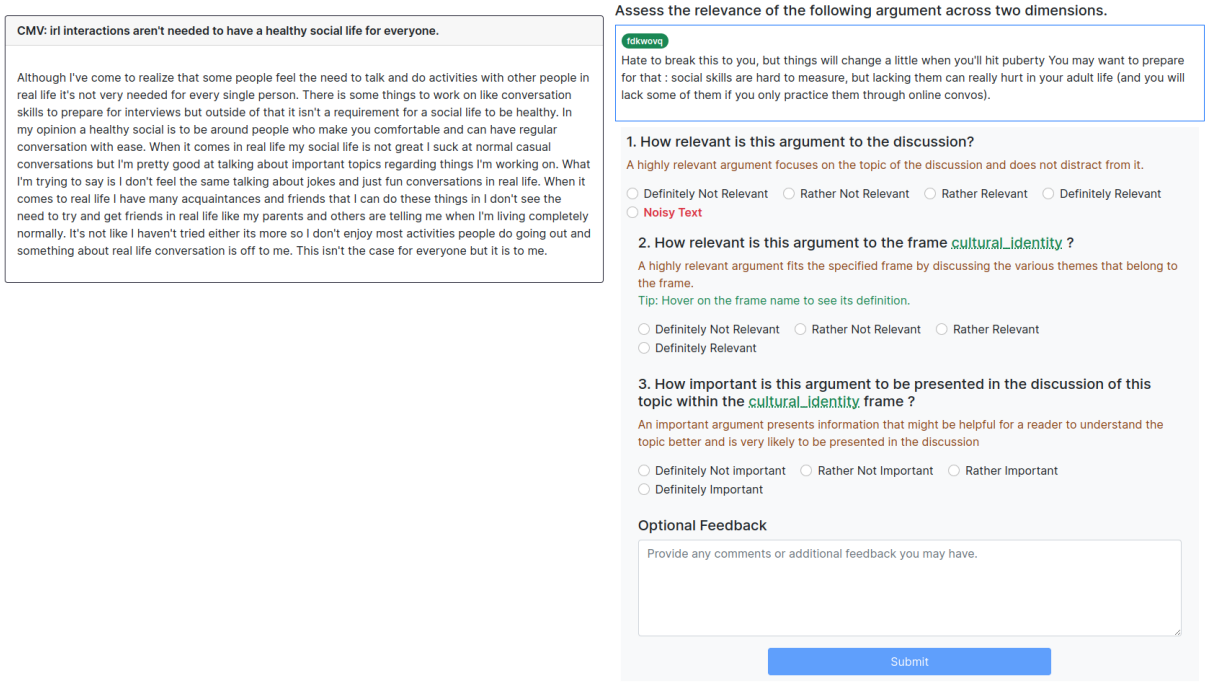

Figure 4: Annotation interface for the **main evaluation**. First, we removed the summary of the argument and always showed the complete argument. Next, we allowed marking a text as "noisy" and skip answering the remaining questions. Finally, as it was difficult to decide if an argument was important enough to be included in a summary of the discussion before reading the entire discussion, we rephrased the important question as the likelihood of including an argument in the *discussion* of the topic.