Corpus-Oriented Stance Target Extraction

Benjamin D. Steel
McGill University
benjamin.steel@mail.mcgill.ca

Derek Ruths
McGill University
derek.ruths@mcgill.ca

Abstract

Understanding public discourse through the frame of stance detection requires effective extraction of issues of discussion, or stance targets. Yet current approaches to stance target extraction are limited, only focusing on a single document to single stance target mapping. We propose a broader view of stance target extraction, which we call corpus-oriented stance target extraction. This approach considers that documents have multiple stance targets, those stance targets are hierarchical in nature, and document stance targets should not be considered in isolation of other documents in a corpus. We develop a formalization and metrics for this task, propose a new method to address this task, and show its improvement over previous methods using supervised and unsupervised metrics, and human evaluation tasks. Finally, we demonstrate its utility in a case study, showcasing its ability to aid in reliably surfacing key issues of discussion in large-scale corpuses.

1 Introduction

Disagreement is a critical part of discourse. As such, understanding discourse requires inferring the constituent disagreements. This task becomes increasingly complex as discussions scale to online environments (Gottfried, 2024), where the pressing need to ensure healthy dialogue is compounded by threats from inauthentic influence attempts and harmful platform mechanisms (Saurwein and Spencer-Smith, 2021; Goldstein et al., 2023; Commission, 2024). To address these challenges, we need both easy-to-use analytical methods and clear representations of discussion data. However, developing such tools presents challenges, given that online media documents typically mix many different related issues, topics, and contexts.

Stance detection (i.e. the task of identifying the attitude of the author of a text on a stance target (a claim, entity etc.) (Mohammad et al., 2016)) is a well-developed method for understanding disagreement. But the current state of stance detection is such that, unless one knows a priori the stance targets one wants to know the documents' stance on, one must undertake the difficult task of defining those stance targets oneself via the arduous task of understanding the entire corpus. While there are initial methods available for finding targets in documents, we propose that they are insufficient at the corpus-level, and that such a method needs four key features in order to faithfully and clearly capture stance in a discussion corpus:

1. The stance targets need not be known a priori to the researcher - avoiding human bias in issue selection, and improving scalability.
2. A single document can articulate a position on multiple, or hierarchical (i.e. more abstract, or more general), targets - which frequently occurs in the real-world - and as such, the method should map the document to these targets.
3. Targets should be determined in the context of the corpus - meaning both that the discussion as a whole aids the inference of the targets of a document, and that documents should be clustered to targets to allow aggregation for downstream application.
4. Documents should be mapped to clear representations of these stance targets, to aid understanding, and allow use in downstream tasks

Existing approaches do not address all of these features. Most stance target extraction methods produce a single stance target for a single given document, without attending to the broader context of a discussion, or allowing for multiple issues to be addressed in a document (Irani et al., 2024; Akash et al., 2024; Li et al., 2023; Zhang et al., 2021). Disagreement discovery methods from outside the stance-detection literature that *do* consider a corpus

Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025), pages 209–227

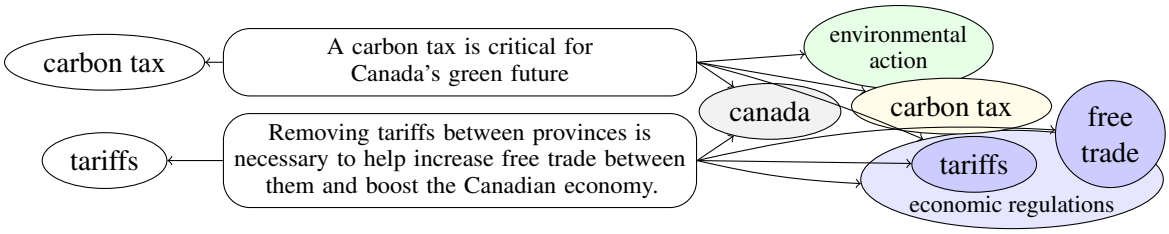


Figure 1: Comparison of assigning a single stance target to each document (left), versus assigning multiple hierarchical stance targets that overlap with other texts as proposed here (right).

as a whole (Paschalides et al., 2021; He et al., 2021) do not produce a clear mapping of documents to stance targets.

We make four contributions. We formalize this task of mapping issues/targets of disagreements in a corpus into a computational task which we call *corpus-oriented stance target extraction* (COSTEx). We then provide a metric for evaluating a method's performance on this task. We present a method which addresses the task, and show it outperforms existing methods on our task. Finally, we conduct a case study using our method, which shows that it can retrieve key issues (stance targets and stances) from a discussion represented by a corpus.

With the evaluation and development of a method that performs well on the task we outline here, we can unlock powerful insights in large-scale media corpora, giving us new tools to understand large-scale natural language behaviour such as polarization and public opinion. We release a library for this method at `https://github.com/bendavidsteel/stancemining`

2 Background

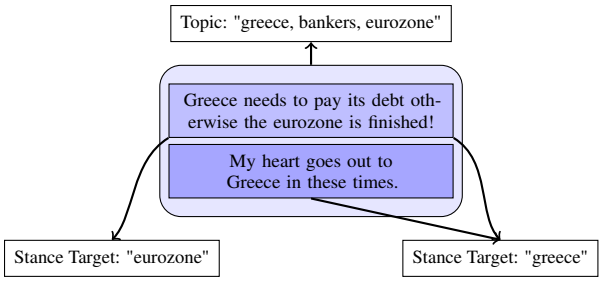
Subjectivity Detection The fields of stance detection, aspect-based sentiment detection, and argument mining have produced methods to identify targets of subjective perspective, and classifying the subjective judgement of documents towards those targets. Li et al. (2023) and Steel and Ruths (2024) look at stance target extraction, but both methods require a priori knowledge and manual target choice at a point in the method, and therefore do not fulfill feature 1. Akash et al. (2024), Irani et al. (2024) and Zhang et al. (2021), all look at open-target extraction (where there are no predefined targets) in stance detection, argument mining, and sentiment detection respectively. All three focus on inferring targets for documents in isolation and, as a result, none of these methods consider the multiple or hierarchical stance targets possible from a document (as represented in Fig. 1 and defined as a required feature 2), or the need for large stance target clusters —groups of documents mapped to the same stance target —if we want to aggregate the data for further analysis (feature 3). Nevertheless, we compare our developed method against WIBA (Irani et al., 2024) in this work.

Polarized and Controversial Topics Topic modelling derived methods are a common approach to this problem space, and naturally handle the desired aggregation process from feature 3. But converting topic clusters to stance target clusters is not trivial. Topic and stance targets clusters don't map neatly one-to-one, as demonstrated in Fig. 2a. Work on topic cluster representation, such as Pham et al. (2023) and Grootendorst (2022), uses large language models (LLMs) to improve the interpretability of cluster names, working towards feature 4. But mapping a topic cluster to a stance target is difficult, as it requires domain knowledge and reasoning to convert topic descriptions into a stance target (Fig. 2b).

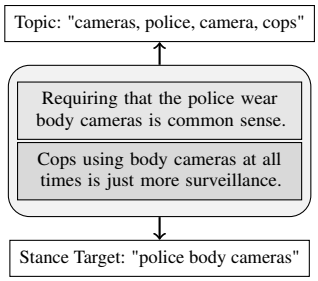
Fukuma et al. (2022) use a network method to find polarized topics, but this method is designed for X/Twitter specific features. Garimella et al. (2018) use hashtags to define conversational graphs, and find partitions in those graphs in order to find controversial topics. This method however relies on hashtags, limiting it to corpuses with heavy hashtag usage. Paschalides et al. (2021) and He et al. (2021) produce methods to find polarized topics, and we evaluate these methods in this work.

3 Problem Definition

Motivated by our desired features from Section 1, we define COSTEx as follows: given a corpus of documents, we seek to identify labeled clusters of those documents where all documents in a cluster share the same stance target, which is captured by the label of the cluster. Crucially, clusters can be overlapping, allowing a document to be assigned



(a) Same high-level topic, but different specific stance targets



(b) Stance targets need more specific names than topics.

Figure 2: Representation of the differences between stance target clusters and topic clusters, showing hierarchical relationships, one-to-many mappings, and different cluster naming requirements, as discovered in manual analysis.

more than one stance target. Formally, for a corpus of documents $D = \{d_1, ..., d_N\}$, we want to find a set stance target clusters, $C = \{c_1, c_2, ...c_M\}$ where $c_i \subset D$, and their corresponding stance targets $T = \{t_1, t_2, ..., t_M\}$. As stance detection has not previously considered corpus-aware methods, we propose new criteria that define success in COSTEx, that measure the extent to which a method that implements this task fulfills the desired features. As such, the COSTEx problem seeks C and T such that they reflect the following criteria:

1. **Clusters with Large Stance Variance:** Given the stance of each document on the stance target $stance(t_i, d_j) \rightarrow \{-1, 0, 1\}$, we want to find stance targets that maximize the stance variance for all related documents:

$$\frac{1}{|C|} \sum_{c_i \in C} Var(\{stance(t_i, d) : d \in c_i\})$$

   This is a metric for picking "controversial" stance targets. Intuitively, stance targets that no-one disagrees on are less interesting than stance targets that people disagree on.

2. **Stance Target Range and Relevance:** We want to find many stance targets that are relevant to the documents. We can measure relevance of targets by ensuring that the stance targets adhere to human judgments of stance targets, via comparison to labeled datasets and custom human annotation (as in Akash et al. (2024)), and we can measure 'many stance targets' by measuring the mean number of targets per document:

$$\frac{1}{|D|} \sum_{d \in D} |\{c_i \in C : d \in c_i\}|$$

3. **Balanced Stance Target Clusters:** We want to optimize for clusters of a range of sizes, including large clusters to allow for useful aggregations, that still capture clusters of meaningful grouping. To measure meaningful clusters, we will use human evaluation. And to measure cluster size, we will use the cluster size Shannon entropy multipled by normalized cluster size range, to ensure there are a balanced number of clusters of a range of sizes:

$$\frac{\max_i c_i - \min_i c_i}{\max_i c_i} \left(-\sum_{i=1}^{n} \frac{c_i}{C} \log_2 \frac{c_i}{C} \right)$$

   where $C = \sum_{k=1}^{n} c_k$

Naturally, in most situations it will be impossible to perfectly satisfy all of these. Solutions to this task will have to make careful trade-offs between these criteria. In practice, some of these metrics are trivially measurable, and some of them are much harder to measure (i.e. the ones requiring human evaluation). We will seek to do so via quantitative supervised and unsupervised metrics, and metrics from human evaluation tasks.

Finally, we must address the question of what we mean by *stance targets* in the formulation above. In the literature, it is common to define stance targets either as noun-phrases (e.g., "police body cameras"), or claims ("police should wear police body cameras") (Zhao and Caragea, 2024). A document assigned to this stance target contains content that takes a position on it. Note that, where stance targets are concerned, the problem definition requires only a means of scoring a document's position on a stance target (i.e., $stance(t_i, d)$). As a result, the problem admits either noun-phrases or claims as stance targets.

4 Methods

Here, we propose our method that fulfills all the features in Section 1. Rather than clustering documents directly (which conflates topics with stance

targets as discussed in Section 2), we first extract multiple stance targets (fulfilling feature 2), then cluster those targets. This lets us find large stance target clusters, where documents can naturally belong to multiple large clusters. We then generate stance targets for each of those clusters to find higher-level stance targets (i.e. targets that are hierarchically more abstract, or more general to the cluster). Collecting all these stance targets together for each document, we then have small, specific stance target clusters, and larger, high level stance target clusters. We call this method *ExtractCluster* (EC), formally define it in Algorithm 3, and show a simplified system diagram of it in Fig. 4. Our method aims to meaningfully achieve each criteria from Section 3.

The base stance targets are produced using an LLM fine-tuned on document - stance target pairs, using diverse beam search (Vijayakumar et al., 2016) to generate multiple targets. We cluster the targets using BERTopic (Grootendorst, 2022), which provides an easy-to-use and configurable topic modelling solution. The default clustering configuration of BERTopic gives us one layer of clusters, meaning there are two hierarchical levels of stance targets. The higher-level stance targets are generated using an LLM with a few-shot prompt (shown in Appendix B.6). To avoid producing stance targets for each document that are paraphrases of each other, we remove stance targets where their sentence embeddings have high cosine similarity based on a configurable threshold (Reimers and Gurevych, 2019) (detailed in Appendix B.4).

function EXTRACTCLUSTER(D)
  for each document $d \in D$ **do**
    $T_d \leftarrow$ ExtractStanceTargets(d)
    $T_d \leftarrow$ RemoveSimilarTargets(T_d)
  end for
  $C \leftarrow$ TopicModelTargets(T)
  for each cluster $c \in C$ **do**
    $T_c \leftarrow$ GenerateHigherLevelTargets(c)
    $T_c \leftarrow$ RemoveSimilarTargets(T_c)
    for each $d : \exists t \in T_d : t \in c$ **do**
      $T_d \leftarrow T_d \cup T_c$
    end for
  end for
  for each document $d \in D$ **do**
    for each target $t \in T_d$ **do**
      $S_{d,t} \leftarrow$ ClassifyStance(d, t)
    end for
  end for
  return D, T, S
end function

Figure 3: Algorithm used by *EC*. Topic modelling is done on the flat list of stance targets using BERTopic. Removal of similar targets is based on high cosine similarity between stance target sentence embeddings.

4.1 Comparison Methods

We selected three methods to compare to EC on our task COSTEx. Although these methods do not fully address our proposed task, they address it sufficiently to warrant evaluation.

POLAR (Paschalides et al., 2021) uses entity extraction and network methods to find polarized topics. While this method is designed to find polarized topics, we apply it here to the similar but more general COSTEx task. Though the method does not explicitly map documents to stance targets, we extend it to use any entities or noun phrases that are tagged as part of a polarized topic as stance targets for their respective documents.

PaCTE (He et al., 2021) combines topic modeling and a partisanship classification model to find topics of partisan disagreement. We adapt it here to finding targets of stance disagreement.

WIBA (Irani et al., 2024) uses three fine-tuned LLMs to determine whether a document features an argument, extracts the claim topic of the argument, then determines the stance of the document on that argument. In this application we remove the argument detection step, instead relying on the neutral label in stance classification. While this method is defined for argument detection, it maps neatly to stance detection. Although a more stance detection-centric method is now available (Akash et al., 2024), we use Irani et al. (2024) because it was available with an implementation at the time of this work's inception. However, the two methods are functionally similar enough as to be interchangeable in this context.

Comparison To summarize, these three methods from the literature fulfill different features of the COSTEx task as defined in Section 1. We summarize the ways in which the representative methods —which we will evaluate here —fulfill those requirements in Table 1. As shown, none of the methods achieve all of the necessary attributes, but they

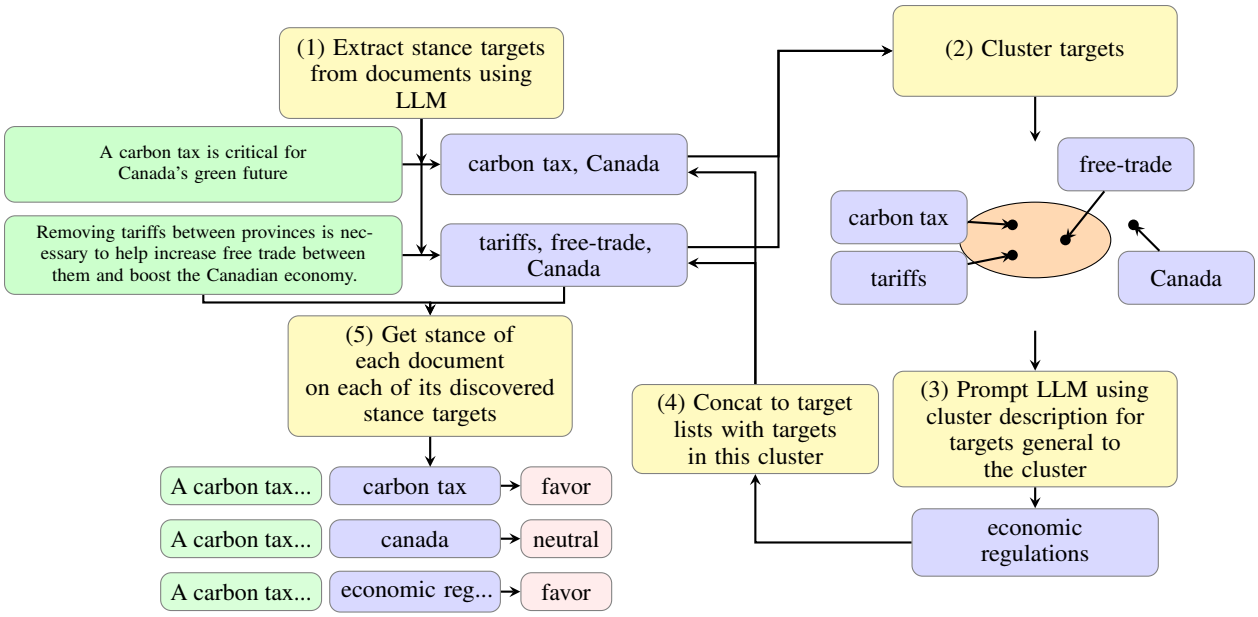


Figure 4: Simplified system diagram of the *ExtractCluster* (EC) method. Green boxes represent documents, yellow boxes represent system component steps, blue boxes represent stance targets, the orange circle represents a cluster, and pink boxes represent stance classifications. Numbers in each component step indicate the sequence of operations. We have excluded the stance target de-duplication step for brevity.

each achieve most aspects of the desired method.

5 Experiments

With our method in hand, we now want to see to what extent it fulfills COSTEx by testing it using metrics and human evaluation methods derived from our formulation, and comparing it to our comparison methods.

Datasets We use two large stance detection datasets to evaluate the methods, VAST (Allaway and McKeown, 2020) and EZ-STANCE (Zhao and Caragea, 2024). These datasets come from two domains, New York Times comments and Twitter respectively, enabling testing across diverse text types. Importantly, both datasets derive their stance targets from each document —as opposed to a dataset designed around a specifically chosen set of stance targets —allowing us to grade the produced stance targets against the annotated stance targets from the datasets. We report statistics from the datasets (Tab. 2).

Configuration We fine-tune a stance target extraction model and a stance detection model, on both VAST and EZ-STANCE for both tasks, using *Llama-3.2-1B-Instruct* as a base model (an open-weight 1B parameter model) (Meta, 2024). We use these fine-tuned models for both EC and WIBA. For diverse beam generations in EC, we sample 3 generations, as this is the ceiling integer above the highest mean number of targets in each dataset (Tab. 2). We use a cosine similarity value of 0.8 for EC, as through manual validation, this de-duplicates stance targets that are functionally identical. We use *phi-3.5-mini-instruct* (Abdin et al., 2024) to generate higher-level stance targets for EC, a 4B model suitable for few-shot prompting. We list all other experimental implementation details of each comparison method in Appendix B.

5.1 Automated Evaluation

Metrics As previously highlighted (Sec. 3), some of the outcomes that we want to optimize in our method are trivially measurable, and some are much more difficult to measure. We therefore propose a set of metrics that assess the extent to which the method outputs optimize for the objectives defined above. While our method does produce hierarchical stance targets, in evaluation we will treat them as a flat list, while maintaining the valuable property of higher-level stance targets aggregating across more documents.

- **Target F1:** The BERTScore F1 (Zhang et al., 2019) of the discovered targets, compared to the annotated dataset, as in (Akash et al., 2024). As we have a set of annotated stance targets for each document in our labelled dataset, we compute the precision by comparing each predicted stance target to all gold

Feature	PaCTE	POLAR	WIBA	EC
Stance target discovery through aggregation	✓	✓	✗	✓
Multiple stance targets per document	✓	✓	✗	✓
Map documents to stance targets	✗	✗	✓	✓

Table 1: Comparison of different methods against our method, *EC*, for each of features 3, 2, and 4, as defined in Section 1. All of the methods fulfill feature 1.

Dataset	Num. Ex.	Mean. Num. Targets	Stance Split (F/N/A)	Lang.
VAST	784	2.45	0.47/0.02/0.51	en
EZ-STANCE	1561	1.71	0.36/0.35/0.29	en

Table 2: Statistics from the datasets used for testing.

stance targets, and the recall by comparing each gold stance target to all predicted stance targets, and compute the F1 from the resulting precision and recall, as defined in Appendix E.1. This metric measures adherence to Criterium 2.

- **Stance Retrieval F1:** The F1 of the discovered stance of the documents, compared to a labeled dataset. Seeing as we have a potentially different set of predicted stance targets as the gold stance targets, we create a mapping of predicted stance targets to gold stance targets where the sentence embedding cosine similarity is greater than 0.9, then compute the precision by comparing each predicted stance to all gold stances, the recall by comparing each gold stance to all predicted stances, and the F1 score from the precision and recall, as defined in Appendix E.2.
- **Stance Variance:** See Criterium 1.
- **Mean Num. of Targets:** See Criterium 2
- **Balanced Cluster Sizes:** See Criterium 3
- **Walltime:** Method run-time duration (s).

The supervised metrics, the target F1 and stance retrieval F1, are measuring the adherence of the method to a typical stance detection dataset. However, we also want to optimize for multi-target, hierarchical, and clustered stance targets. Optimizing for metrics that measure these aspects will reduce our target F1 score, as the stance targets will be further from the stance targets given in the base datasets. We need to assess our results holistically, and consider that, as part of our task formalization, any solution to this problem is making a trade-off between objectives. We will therefore determine the overall ranking of the methods via a summed rank order: we find the rank of each method on every metric, sum all the ranks for each method, and the lowest summed rank order is the best method.

Results We report the supervised and unsupervised metrics from the mean of 5 runs for each method on each dataset (Tab. 3). EC generally outperforms other methods, except on stance target F1 and precision, and wall-time. Stance retrieval rankings are robust to varying cosine similarity, see Appendix E.2.1. We conducted ablations of cosine similarity threshold and number of beam generations and confirm that our chosen values yield the best results, see Appendix C. We also tested a version of the method that removes lower variance stance targets and find that it achieves higher mean stance variance but worse on all other metrics (See Appendix C.1), showing our method could not be trivially improved in this manner.

Human Evaluation We created two human evaluation tasks to evaluate the method outputs. The first task presents a triad of documents (a base document, another from the same cluster, and one from a different cluster) and has the annotator select which two documents go in the same stance target cluster. We measure how often the annotators agree with the document clustering chosen by each method. A second task presents a base document, and two stance target sets provided by two different methods, and a prompt asks the labeler to choose between the two stance target sets, or neither if neither are suitable. Each annotator received an evaluation guide prior to their evaluation task to explain the concepts in use in the task. We obtained 483 and 492 annotations for each task respectively, from 6 annotators who were students in the authors' lab. We show the prompts and evaluation guide given to annotators, and example generation process in Appendix D.

To ensure there was agreement between annotators, we had two annotators evaluate the same set of 20 examples from each task. The Fleiss' Kappa (Fleiss et al., 1981) of the stance target cluster task was 0.53, and for the stance target task it

Method	Target F1 ↑	Target P. ↑	Target R. ↑	Stance F1 ↑	Stance P. ↑	Stance R. ↑	Mean Num. Targets ↑	Stance Variance ↑	B. Cluster Sizes ↑	Wall time ↓
VAST										
PaCTE	0.775	0.779	0.771	0.000	0.000	0.000	1.212	**0.226**	2.314	**155.2**
POLAR	0.512	0.524	0.501	-	-	-	2.140	-	3.028	327.7
WIBA	**0.910**	**0.930**	0.891	0.116	0.190	0.089	1.000	0.108	7.753	246.9
EC	0.897	0.889	**0.907**	**0.143**	**0.210**	**0.119**	**3.190**	0.136	**8.031**	1569.1
EZSTANCE										
PaCTE	0.766	0.768	0.763	0.000	0.000	0.000	1.038	**0.208**	4.311	**213.7**
POLAR	0.038	0.038	0.037	-	-	-	0.218	-	0.168	582.7
WIBA	**0.884**	**0.899**	**0.871**	0.145	0.200	0.120	1.000	0.019	9.495	766.4
EC	0.859	0.851	0.867	**0.158**	**0.202**	**0.141**	**3.380**	0.039	**9.520**	3349.8

Table 3: Metrics comparison across datasets and methods averaged across 5 runs for each method and dataset. Best metrics are indicated with arrows. P. and R. stand for precision and recall respectively. We do not include stance results for POLAR as it does not assign stance to individual documents. Bold numbers indicate the best performance, underline indicates second best.

Method	LSR	Agree Pct.	Example Output
PaCTE	-2.23	0.19	school,health,covid...
POLAR	-2.79	0.00	anyone
WIBA	1.51	**0.62**	medical law
EC	**2.23**	0.34	jerusalem

Table 4: Luce Spectral Ranking (LSR) pairwise comparison score, calculated by comparing different methods' stance target sets for each document, alongside an example stance target output from each method for reference (PaCTE example shown truncated). And percentage of examples where annotators agreed with the clustering of a document triad, for each method.

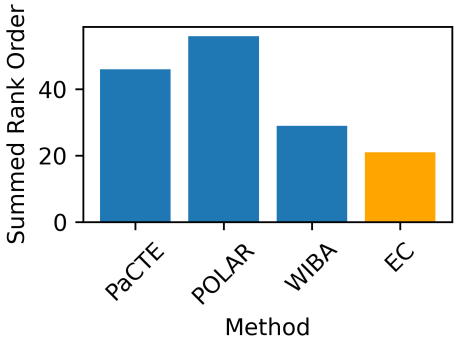


Figure 5: Summed rank order across all metrics for each method. *EC* outperforms the other methods we trial from the literature across our metrics.

was 0.83, indicating inter-annotator agreement. For the stance target set comparison task, we use the Luce spectral ranking (LSR) (Maystre and Grossglauser, 2015) (via Choix [1]) to determine the output stance targets most preferred by human annotators. EC and WIBA are rated the highest, with POLAR and PaCTE rated poorly (Tab. 4). For stance target cluster agreement scores, we simply record the number of times the human evaluator agreed with the method. WIBA, EC, and PaCTE obtain the best results for cluster evaluation, and POLAR obtains no agreement from evaluators (Tab. 4).

Summary We show the summed rank order of each method, for each metric, in Fig. 5. This demonstrates the overall rank of the methods on the COSTEx task we introduce in this work.

6 Discussion

POLAR needs to find many named entities to find polarized topics (being designed for news articles), and as such performs poorly on the short text datasets used here, especially the EZ-STANCE dataset (Tab. 3). We observe poor evaluations of naming and clustering performance (Tab. 4).

PaCTE's use of LDA topic modeling and a small classifier model means that it can quickly find large stance target clusters with high stance variance (Tab. 3). However, the naming of clusters with topic keywords results in a low evaluation score (Tab. 4), and the stance target clusters are only moderately agreed with (Tab. 4).

WIBA's stance target extraction produces good stance targets (Tab. 3 and 4), and performs highest on our cluster agreement evaluation (Tab. 4). But the small stance target clusters it produces —due to only producing one stance target per document —result in lower stance retrieval F1, and low stance variance and cluster size (Tab. 3).

EC outperforms WIBA in stance target cluster size, stance variance, stance retrieval, and stance

[1] github.com/lucasmaystre/choix

target set preference (Tab. 4). However, we also see that it under-performs WIBA on our cluster agreement evaluation (Tab. 4), and stance target precision. We infer that as EC maps higher-level stance targets to each document —that have no parallel in the annotated datasets we use —which results in large clusters defined by abstract stance targets that are too general for annotators to spot in our cluster agreement exercise. Nonetheless, from the summed rank order (Fig. 5), EC is the most effective method tested here.

7 Case Study

Having empirically shown our method outperforms other methods from the literature, we chose to assess its effectiveness at identifying key characteristics of a discourse under real-world conditions. Topic modelling is frequently used for exploratory analysis of discourse corpuses (Hobson et al., 2024; Falkenberg et al., 2022). Although it does not disaggregate expressed valence —a key part of separating discourse (Ghafouri et al., 2024)—the previously missing step of stance detection —stance target discovery —makes it labour-intensive to run as a go-to exploratory step. Crucially, both EC and BERTopic (Grootendorst, 2022) require no notable parameter tuning and, so, are of equal complexity for a domain researcher to use.

We assumed the role of a researcher studying the political views present in a social media dataset. We chose a 2024 Twitter dataset consisting of 1.4m tweets (81% English, 9% French, 10% other languages) from 1.9k prominent Canadian media accounts (Pehlivan et al., 2025). See Appendix F for implementation details. Table 5 shows the largest stance target vs. topic clusters.

Meaningful clusters. Both stance target and topic modeling methods can produce nonsensical clusters. How do we quickly remove the noise? In topic modeling, this is messy: as seen in Table 5, some of the largest topic clusters are meaningless (e.g., "shes, shell, shed, quelle"). In contrast, with EC, an easy way of filtering weak stance targets is by simply dropping small stance clusters, with the intuition being that modal stance targets are more frequently good stance targets. In this case study, we found removing stance targets with less than 50 data-points to be a good level. At this border, there are some good stance targets ('Organic Food Movement', and 'US Col. Lawrence Wilkerson') but also many non-specific or nonsensical stance targets ('which will', 'candidate nomination').

Cluster informativeness. Table 5 highlights the informativeness of EC clusters in several ways. First, stance target clusters capture more of the documents than the largest topics, due to EC allowing documents to belong to multiple stance target clusters. If we kept just the first extracted stance target (as previously (Akash et al., 2024)), the 'trudeau' stance target would only be assigned to 22k documents, with our method allowing us to know the stance of more documents on 'trudeau' where he may be referred to implicitly. Second, the stance targets capture the large ongoing issues of Canadian public discourse (Canada, Justin Trudeau, the Liberal Party), and topical issues (Donald Trump's presidency, the B.C. election, the Israel-Palestine conflict), where these large issues are missed by the topics - instead emphasizing smaller topics like the Olympics. Even for topic clusters that are not "noise", the stance target names are consistently more specific, and therefore more usable for further analysis. However, EC needs improved stance target de-duplication, as shown by the presence of 'j. trudeau' and 'trudeau'.

Understanding stance on the target clusters. We show a map of the 30 largest stance target clusters in Fig. 6. Having stance classifications on so many targets surfaces key aspects of the discourse: allowing us to compare mean stance on party leaders (-0.57 for Trudeau vs. -0.44 for Poilievre), parties (-0.45 for the NDP vs. -0.62 for the Liberal Party), and foreign policy issues (-0.46 for Israel vs. -0.79 for Hamas) with one method application (where we have substituted 'favor' for 1, 'neutral' for 0, and 'against' for -1).

This case study highlights how EC gave the researcher a larger and more detailed map of the discussion in our dataset, alongside more specific and understandable cluster names.

8 Conclusion

We have motivated and conceptualized the task of *COSTEx*, and shown that our new method for this task, *EC*, outperforms previous methods for similar tasks. We then used a large-scale real-world dataset to demonstrate that our method reliably captures and represents clusters of stance target discussion. We hope that this method can aid practitioners in quickly understanding discourse in large and wide-ranging real-world datasets, and help improve understanding of complex behaviors such as polarization and public opinion in our quickly changing information environments.

Stance Targets		Topics	
Name	Count	Name	Count
canada	76k	gaza, israel, israeli, hamas	22k
j. trudeau	54k	olympics, game, olympic, athletes	15k
trudeau	39k	hes, guy, coyne, mrstache9	10k
trump presidency	29k	url, juliemarienolke, thejagmeetsingh, saudet80	9k
liberal party	22k	healthcare, nurses, doctors, doctor	9k
israeli	17k	shes, shell, shed, quelle	8k
trump	17k	housing, rent, rental, homes	7k
b.c. ndp	16k	trudeau, justin, trudeaus, resign	6k

Table 5: Comparing largest stance target clusters to largest topic clusters.

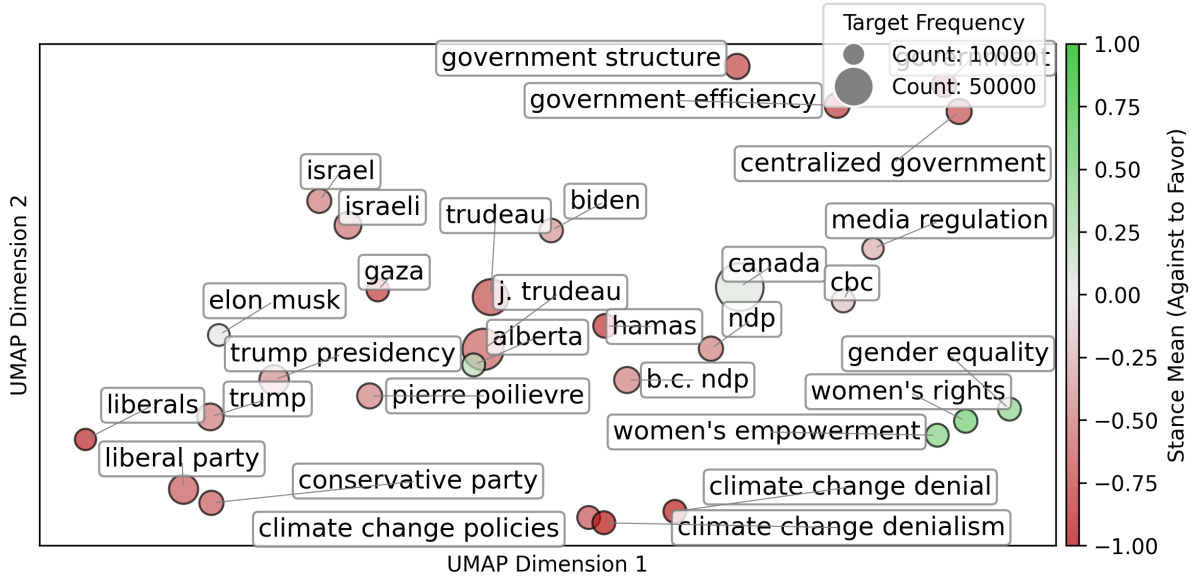


Figure 6: Map of top stance targets, sized by frequency, coloured by average stance. In general, the major issues that Canadian social media users tend to have attitudes on are represented. However, we can also see that improved stance target de-duplication is necessary, along stance target clarity (‘climate change denialism’).

9 Limitations

Datasets We found that the lack of hierarchical stance targets in the stance-detection datasets used in this work made it difficult to evaluate the ability of the methods to find a full breadth of hierarchical, clustered stance targets for each document. We can only use these datasets to assess the extent to which the method found the base stance targets for each document. Future work should develop new datasets to evaluate higher-level stance targets.

Methods Stance target de-duplication became an issue when we applied our method to a larger corpus. We experimented with using DBSCAN to some success, but de-duplicating different ways of spelling names (‘j. trudeau’, ‘trudeau’) while avoiding false positives requires a carefully set distance threshold between embeddings. Additionally, our method of using diverse generation to generate multiple stance targets for each document —while not requiring re-training of our stance target generation model —could be made faster and more flexible by generating targets as a list.

Task Formulation Optimizing for stance variance deprioritizes stance targets that are generally agreed upon, but when disagreed upon, are interesting, such as conspiracy theories, so optimizing for this metric is a trade-off.

Another key limitation was a lack of a principled framework for defining a hierarchy of targets. In practice, LLM prompting produced sufficiently useful results here, but a more well-defined definition could produce stronger results.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat

Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Abu Ubaida Akash, Ahmed Fahmy, and Amine Trabelsi. 2024. Can large language models address open-target stance detection? *arXiv preprint arXiv:2409.00222*.

Emily Allaway and Kathleen McKeown. 2020. Zero-shot stance detection: A dataset and model using generalized topic representations. *arXiv preprint arXiv:2010.03640*.

European Commission. 2024. Commission opens formal proceedings against tiktok on election risks under the digital services act.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Max Falkenberg, Alessandro Galeazzi, Maddalena Torricelli, Niccolò Di Marco, Francesca Larosa, Madalina Sas, Amin Mekacher, Warren Pearce, Fabiana Zollo, Walter Quattrociocchi, et al. 2022. Growing polarization around climate change on social media. *Nature Climate Change*, 12(12):1114–1121.

Joseph L Fleiss, Bruce Levin, Myunghee Cho Paik, et al. 1981. The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2(212-236):22–23.

Tomoki Fukuma, Koki Noda, Hiroki Kumagai, Hiroki Yamamoto, Yoshiharu Ichikawa, Kyosuke Kambe, Yu Maubuchi, and Fujio Toriumi. 2022. How many tweets dowe need?: Efficient mining of short-term polarized topics on twitter: A case study from japan. *arXiv preprint arXiv:2211.16305*.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1):1–27.

Vahid Ghafouri, Jose Such, Guillermo Suarez-Tangil, et al. 2024. I love pineapple on pizza!= i hate pineapple on pizza: Stance-aware sentence transformers for opinion mining. In *Empirical Methods in Natural Language Processing*.

Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*.

Jeffrey Gottfried. 2024. Americans' social media use. *Pew Research Center*, 31.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Zihao He, Negar Mokhberian, António Câmara, Andrés Abeliuk, and Kristina Lerman. 2021. Detecting polarized topics using partisanship-aware contextualized topic embeddings. *arXiv preprint arXiv:2104.07814*.

David Hobson, Haiqi Zhou, Derek Ruths, and Andrew Piper. 2024. Story morals: Surfacing value-driven narrative schemas using large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12998–13032.

Matthew Hoffman, Francis Bach, and David Blei. 2010. Online learning for latent dirichlet allocation. *advances in neural information processing systems*, 23.

Arman Irani, Ju Yeon Park, Kevin Esterling, and Michalis Faloutsos. 2024. Wiba: What is being argued? a comprehensive approach to argument mining. *arXiv preprint arXiv:2405.00828*.

Neel Jain, Ping-yeh Chiang, Yuxin Wen, John Kirchenbauer, Hong-Min Chu, Gowthami Somepalli, Brian R Bartoldson, Bhavya Kailkhura, Avi Schwarzschild, Aniruddha Saha, et al. 2023. Neftune: Noisy embeddings improve instruction finetuning. *arXiv preprint arXiv:2310.05914*.

Yingjie Li, Krishna Garg, and Cornelia Caragea. 2023. A new direction in stance detection: Target-stance extraction in the wild. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10071–10085.

Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Lucas Maystre and Matthias Grossglauser. 2015. Fast and accurate inference of plackett–luce models. *Advances in neural information processing systems*, 28.

Meta. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41.

Demetris Paschalides, George Pallis, and Marios D Dikaiakos. 2021. Polar: a holistic framework for the modelling of polarization and identification of polarizing topics in news media. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 348–355.

Zeynep Pehlivan, Saewon Park, Alexei Sisulu Abrahams, Mika Jacques Patel Desblancs, Benjamin David Steel, and Aengus Bridgman. 2025. Can-polnews: A multi-platform dataset of political discourse in canada. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 2550–2559.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2023. Topicgpt: A prompt-based topic modeling framework. *arXiv preprint arXiv:2311.01449*.

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.

Florian Saurwein and Charlotte Spencer-Smith. 2021. Automated trouble: The role of algorithmic selection in harms on social media platforms. *Media and Communication*, 9(4):222–233.

Benjamin Steel and Derek Ruths. 2024. Multi-target user stance discovery on reddit. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 200–214.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 504–510.

Chenye Zhao and Cornelia Caragea. 2024. Ez-stance: A large dataset for english zero-shot stance detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15697–15714.

A Methods

In addition to the method we propose in this work, we also trialled a method we call *ClusterExtract*, inspired by PaCTE. It starts by finding hierarchical topics in the corpus using BERTopic (Grootendorst, 2022), then assigns stance targets to each topic. It is described in Algorithm 1. However, we found that it produced inferior results to *EC*, and so do not detail it in the main results of the work.

B Implementations

B.1 POLAR

We used all of the default parameter settings and models for POLAR for the VAST dataset, but for EZ-STANCE, we reduce the noun phrase clustering threshold from 0.8 to 0.6, as the default value was resulting in no found clusters given that the EZ-STANCE dataset is composed of low word count tweets, which have low entity mention counts.

In adapting this method, we need to extend it by mapping the chosen polarized topics back to the documents, to allow our metrics to be applied to the results. We do so by considering a document to be in a stance target cluster when it features a polarized entity, and the discovered noun phrases as the stance targets.

B.2 PaCTE

We train the PaCTE BERT model (Devlin, 2018) using the combined training sets from VAST and EZ-STANCE, removing all neutral examples as the original implementation was only trained on partisan news.

We use online latent dirichlet allocation (LDA) (Hoffman et al., 2010) as a drop in method speed-up, instead of the original single-core method. Other implementation details are all the same as the original implementation.

Algorithm 1 Algorithm used by *ClusterExtract*.

Require: Documents D

function CLUSTEREXTRACT(D)
  $C \leftarrow$ TopicModelDocs(D)
  ▷ Handle outlier documents (Topic = -1)
  $D_{out} \leftarrow$ FilterOutliers(D, C)
  for each document $d \in D_{out}$ **do**
    $T_d \leftarrow$ ExtractStanceTargets(d)
    $T_d \leftarrow$ RemoveSimilarTargets(T_d)
  end for
  ▷ Handle non-outlier documents
  for each cluster $c \in C$ **do**
    $T_c \leftarrow$ ExtractClusterStanceTargets(c)
    $T_c \leftarrow$ RemoveSimilarTargets(T_c)
  end for
  ▷ Generate hierarchical topic targets
  $H \leftarrow$ GetHierarchicalTopics(T)
  for each parent cluster $c \in H$ **do**
    $C_p \leftarrow$ GetChildTopics(c)
    $T_p \leftarrow$ AggregateChildTargets(C_p)
    $T_p \leftarrow$ RemoveSimilarTargets(T_p)
  end for
  ▷ Combine targets and remove duplicates
  for each document $d \in D$ **do**
    if $d \notin D_{out}$ **then**
      $c \leftarrow$ GetDocumentCluster(d)
      $p \leftarrow$ GetParentCluster(c)
      $T_d \leftarrow T_c \cup T_p$
      $T_d \leftarrow$ RemoveSimilarTargets(T_d)
    end if
    for each target $t \in T_d$ **do**
      $S_{d,t} \leftarrow$ DetermineStance(d, t)
    end for
  end for
  return D, T, S
end function

B.3 WIBA

We used Llama 3.2 1B (Meta, 2024) as the base LLM for our implementation of WIBA, for its trade-off of performance with small size. Training used the combined VAST and EZ-STANCE train/validation sets. On the combined test sets, it achieved a stance detection F1 of 71.5%, and for stance target extraction it obtained a BERTScore of 90.3%, comparable with the metrics achieved in the original work.

We replaced the system and instruction tuning tokens with a chat template as appropriate for the Llama model. We used a cosine learning rate with warmup that increments every step (Loshchilov and Hutter, 2016), and NEFTune to improve fine-tuned accuracy (Jain et al., 2023). We trained on a 24GB NVIDIA GPU, training took roughly 8 hours.

B.4 EC

For diverse generation, we generate 3 return sequences, by exploring 3 beam groups using 6 beams, with a diversity penalty of 10.0. We use a no repeat n-gram size of 2 to prevent repetition.

We use the *paraphrase-MiniLM-L6-v2* sentence transformer model (Reimers and Gurevych, 2019) to embed candidate stance targets, and remove a target from pairs that have a cosine similarity of higher than 0.8.

We run ablation experiments on the number of beam groups and the cosine similarity threshold used for stance target de-duplication in Section C.

B.5 Datasets

When using VAST as a comparison dataset for the methods, we remove the synthetic neutral examples, as these targets aren't specific for each document. We do however use the synthetic neutral examples to train our stance detection model.

B.6 Prompts

We include the few-shot prompt used for stance target extraction from topic clusters in Prompt 1:

Prompt 1: Prompt used for extracting stance targets from a topic cluster.

System:
You are an expert at analyzing discussions across multiple documents.

Human:
Your task is to identify a common stance target that multiple documents are expressing opinions about.
Instructions:
1. Read all provided documents
2. Identify topics that appear across multiple documents
3. Determine if there is a shared target that documents are taking stances on
4. Express the target as a clear noun phrase
Input:
Documents: [list of texts]
Output:
Stance target: [noun phrase or "None"]
Reasoning: [2-3 sentences explaining the choice]
Examples:
Example 1:
Documents:
"The council's new parking fees are excessive. Downtown businesses will suffer as shoppers avoid the area."
"Increased parking rates will encourage public transit use. This is exactly what our city needs."
"Local restaurant owners report 20% fewer customers since the parking fee increase."
Output:
Stance target: downtown parking fees
Reasoning: All three documents discuss the impact of new parking fees, though from different angles. The documents show varying stances on this policy change's effects on business and transportation behavior.""",
Example 2:
Documents:
"Beijing saw clear skies yesterday as wind cleared the air." "Traffic was unusually light on Monday due to the holiday." "New subway line construction continues on schedule."
Output:
Stance target: None
Reasoning: While all documents relate to urban conditions, they discuss different aspects with no common target for stance-taking. The texts are primarily descriptive rather than expressing stances.
Example 3:
Documents:
"AI art tools make creativity accessible to everyone."
"Generated images lack the soul of human-made art."
"Artists demand proper attribution when AI models use their work."
Output:
Stance target: AI-generated art
Reasoning: The documents all address AI's role in art creation, discussing its benefits, limitations, and ethical implications. While covering different aspects, they all take stances on AI's place in artistic creation.
—
Documents:
{formatted_docs}

Assistant:
Output:
Stance target:

We include the few-shot prompt used for aggregating stance targets in Prompt 2:

Prompt 2: 3-shot in-context prompt for aggregating stance target clusters.

System:
You are an expert at analyzing and categorizing topics.

Human:
Your task is to generate a generalized stance target that best represents a cluster of related specific stance targets.
Instructions:
1. Review the provided stance targets and keywords that characterize the topic cluster
2. Identify the common theme or broader issue these targets relate to
3. Generate a concise noun phrase that:
- Captures the core concept shared across the targets
- Is general enough to encompass the specific instances
- Is specific enough to be meaningful for stance analysis
Input:
Representative stance targets: [list of stance targets]
Top keywords: [list of high tf-idf terms]
Output format:
Generalized target: [noun phrase]
Reasoning: [1-2 sentences explaining why this generalization fits]
Examples:
Input:
Representative stance targets: ["vaccine mandates", "mandatory covid shots", "required immunization for schools"]
Top keywords: ["mandatory", "requirement", "public health", "immunization", "vaccination"]
Output:
Generalized target: vaccination requirements
Reasoning: This captures the common theme of mandatory immunization policies while being broad enough to cover various contexts (workplace, school, public spaces).
Input:
Representative stance targets: ["EVs in cities", "gas car phase-out", "zero emission zones"]
Top keywords: ["emissions", "vehicles", "transportation", "electric", "fossil-fuel"]
Output:
Generalized target: vehicle electrification
Reasoning: This encompasses various aspects of transitioning from gas to electric vehicles, including both the technology and policy dimensions.
Input:
Representative stance targets: ["content moderation", "online censorship", "platform guidelines"]
Top keywords: ["social media", "guidelines", "content", "moderation", "posts"]
Output:
Generalized target: social media content control
Reasoning: This captures the broader issue of managing online content while remaining neutral on the specific approach or implementation.
—
Representative stance targets: {repr_docs}
Top keywords: {keywords}

Assistant:
Output:
Generalized target:

C Ablations

We re-ran EC for varying values of the number of beam generations and cosine similarity threshold for stance target de-duplication to explore the impact it had on method outputs. Figs. 7 and 8 shows that 3 generated beam groups produces consistently the best results on our automated metrics (other the number of beam groups and wall-time being linearly directly correlated), out of 2, 3, and 5 as possible values. Figs. 9 and 10 shows that varying the cosine similarity threshold between values of 0.8, 0.9, and 0.95 has minimal effect on the final metrics.

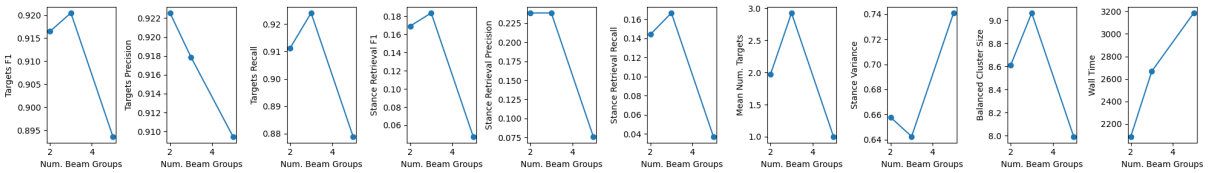


Figure 7: Ablation of the number of beam group generations used for EC for VAST.

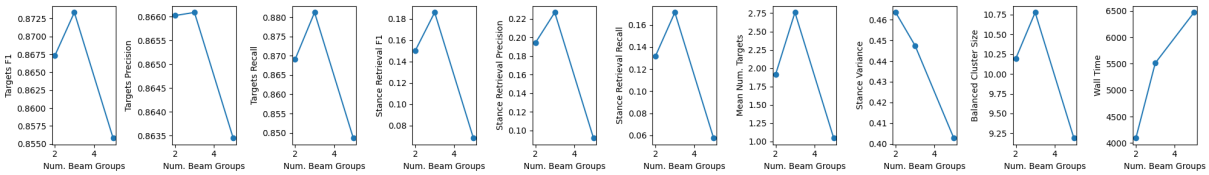


Figure 8: Ablation of the number of beam group generations used for EC for EZ-STANCE.

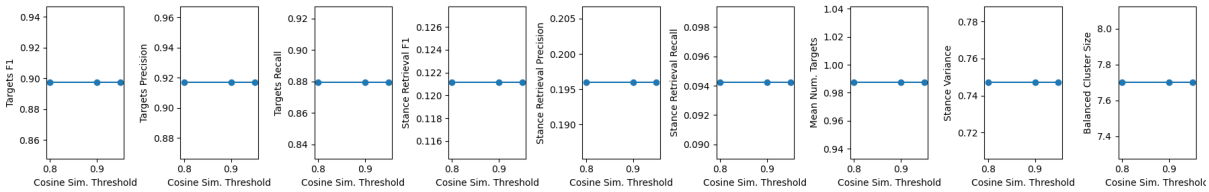


Figure 9: Ablation of the cosine similarity threshold used for stance target de-duplication in EC for VAST.

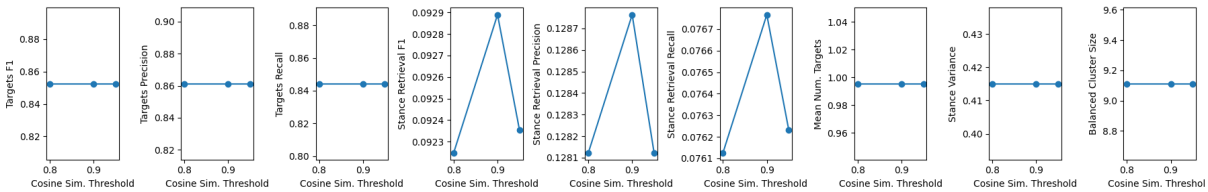


Figure 10: Ablation of the cosine similarity threshold used for stance target de-duplication in EC for EZ-STANCE.

C.1 Stance Variance

Our stance variance metric could be optimized to through de-generate solutions. We wanted to determine the effect that solely optimizing for this metric would have on other metrics. We ran an experiment where we kept only stance targets that had over the 75^{th} and 90^{th} percentiles. Keeping only stance targets over the 75^{th} stance variance percentile did not impact the mean stance variance (0%) change, and reduced stance retrieval F1 by 4.5%. Keeping only stance targets over the 90^{th} stance variance percentile increased mean stance variance by 33%, but decreased stance retrieval F1 (-58%), stance target F1 (-37%), balanced cluster size (-27%), and the mean number of targets per document (-53%).

D Human Evaluation

Human evaluators were fellow students from the authors' lab.

We provided each annotator this explanatory document prior to their evaluation task to help them understand the concepts in Prompt 3.

Prompt 3: Evaluation Guide

What is a stance target? A stance target is a concept that one can have an opinion on. While one can technically have an opinion on almost anything (i.e. one can technically be for or against atoms, but we generally do not consider atoms to be an issue that one is for or against), there are a more constrained set of concepts that we generally put forth opinions, or stances, on.
What is a topic? In computer science, defined as a set of frequently co-occurring words. More generally, synonymous with a theme, or subject that a document can reference or be about. A document can have multiple topics. It is an abstract concept.
Stance Targets So for example, the text: 'I discussed my preference for tariffs over free trade while playing golf today at mar-a-lago' There are 4 prominent concepts: tariffs, free trade, golf, and mar-a-lago. Two topics for this text would be trade policy (tariffs, free trade), and golf (golf, mar-a-lago), as these are frequently co-occurring words/concepts. The two prominent stance targets are tariffs and free-trade, as they are discussed in the context of having a position on them, and are things that one generally has a stance on.
Golf could also be considered a stance target in this context, but is discussed with less emphasis on stance.
Stance targets can also exist at a higher conceptual level. For example, here the author is expressing not only their preference for tariffs, but economic regulation, and protectionism. In this way, the most representative set of stance targets for this text would be 'tariffs', 'free trade', 'economic regulations', and 'protectionism'
One can discuss a stance target while staying

neutral. For example: 'I read about the idea of tariffs recently. Undecided on whether or not they're effective'. This author is neutral on the stance target of tariffs.
Stance Target Clusters Two documents fit in the same stance target cluster, if they discuss the same stance target, whatever the conceptual level of that stance target. The two documents may both be favoring the stance target, on opposing sides of the stance target, or both neutral on the stance target.
For example, the texts: 'I think tariffs are a terrible idea' 'Taxes should be much higher!' Are not in the same 'tariffs' stance target cluster, but are in a stance target cluster: 'economic regulations'

The exact text prompt given to human evaluators for the stance target cluster comparison task is shown in Prompt 4:

Prompt 4: Stance target cluster comparison prompt

Which document discusses a stance target that the base document is also discussing? If both documents discuss completely different stance targets from the base document, choose neither.

To generate triads, for each method and document from both datasets, we randomly sample a document that is a stance target cluster that the base document is also in, and randomly sample a document that is not in any of the same stance target clusters. If the method does not place the base document in a stance target cluster with any other document, then two documents that are not in the same stance cluster are sampled. The order of the two comparison documents is randomly swapped to prevent the chosen document being inferred from the order. We then simply check if the annotator agrees with the method.

The exact text prompt given to human evaluators for the stance target comparison task is shown in Prompt 5:

Prompt 5: Stance target comparison prompt

Compare the two sets of stance targets, and choose the set that better covers the stance targets the document discusses. If neither sets fit at all, choose neither.

We sample comparisons from the set of all pairwise stance target set comparisons between methods for all documents from both methods. We randomly swap the order of these sets to ensure the same method does not always appear on the same side.

E Metrics

E.1 Stance Target F1

For the stance target BERTScore, given a set of documents D where each document d has predicted targets P_d and gold targets G_d, we compute the precision, recall and F1 as:

$$P = \frac{1}{|D|}\sum^{D}\frac{\sum^{P_d}\max_{g\in G_d}\text{BERTScore}(p,g)}{|P_d|}$$

$$R = \frac{1}{|D|}\sum^{D}\frac{\sum^{G_d}\max_{p\in P_d}\text{BERTScore}(g,p)}{|G_d|}$$

$$F1 = \frac{2\cdot P\cdot R}{P+R}$$

E.2 Stance Retrieval F1

Given a set of documents D, where each document d has predicted target-stance pairs $P_d = \{(t,s)\}$, and gold target-stance pairs $G_d = \{(t,s)\}$, where stance can be any of $\{favor, against, neutral\}$.

We define a mapping between predicted stance targets and gold stance targets, where stance targets are only mapped to each other if their sentence embedding cosine similarity is higher than $\theta = 0.9$:

$$M = \{(t_p, t_g) : \max_{t'\in G}\text{sim}(t_p,t')\wedge\text{sim}(t_p,t_g)\geq\theta\}$$

For each document d, define the set of correct predictions:

$$C_d = \{(t_p,s)\in P_d : \exists(t_g,s)\in G_d, (t_p,t_g)\in M\}$$

Then:

$$\text{P} = \frac{1}{|D|}\sum_{d\in D}\frac{|C_d|}{|P_d|}$$

$$\text{R} = \frac{1}{|D|}\sum_{d\in D}\frac{|C_d|}{|G_d|}$$

$$\text{F1} = \frac{2\text{PR}}{\text{P}+\text{R}}$$

E.2.1 Threshold Sensitivity

We looked at the sensitivity of our stance retrieval metrics to the chosen cosine similarity parameter, as seen in Fig. 11. The rankings of the method are robust to varying values of the chosen cosine similarity.

F Case Study Implementation

When deploying EC at scale in the case study, we use smaller models: *SmolLM2-360M-Instruct*[2] to generate the base targets, and *SmolLM2-135M-Instruct*[3] to classify stance. Although this makes applying this method to large datasets more tractable, it occasionally results in poor stance targets. This problem is alleviated by using a strong model for the higher-level stance target generation (*huggingface.co/microsoft/Phi-4-mini-instruct*).

[2] huggingface.co/HuggingFaceTB/SmolLM2-360M-Instruct

[3] huggingface.co/HuggingFaceTB/SmolLM2-135M-Instruct

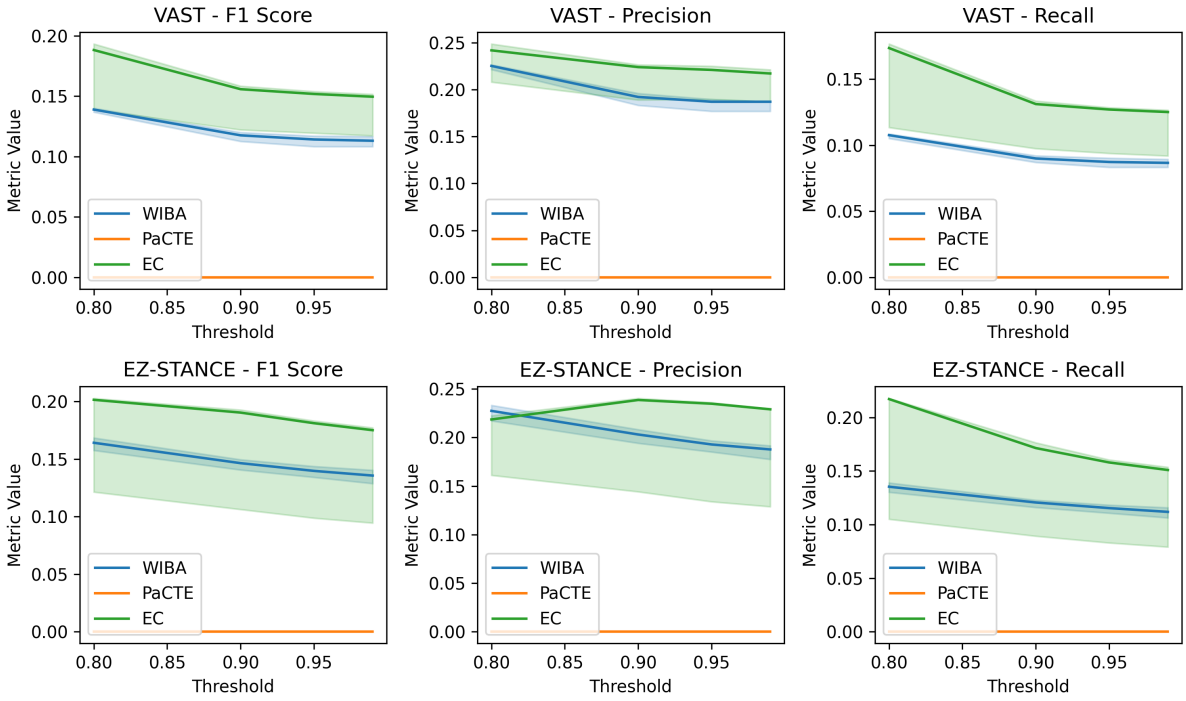


Figure 11: Varying values of the cosine similarity parameter used for calculating stance retrieval against the final value of the metric. Method ranking remains robust to the varying parameter. Shown are the median and quartiles for 5 outputs of each method.