Corpus-Oriented Stance Target Extraction

Anonymous ACL submission

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 discussion. Making decisions requires identifying disagreements, reaching consensus involves compromising in disagreements, and convincing others requires negotiating disagreement. This means finding disagreement is a necessary part of understanding discussion. As more public discussion moves online (Gottfried, 2024), the scale of these discussions grow, as does the potential for harm to come to them (Saurwein and Spencer-Smith, 2021; Goldstein et al., 2023; Commission, 2024). Understanding and ensuring the ongoing health of these conversations requires robust methods to measure them.

The constituent posts of social media meander and mix different related issues, topics, and contexts. They approach the same issue in many different forms. Any effort to map the contours of disagreements in discussion must contend with these and other factors. While the field of stance detection has made significant progress in determining stance in discussions on known issues (i.e. stance targets), comparatively little attention has been paid to automatically mapping the issues themselves. Improving our methods here allows us to map the myriad disagreements in a varied discussion as represented in a social media corpus. We propose that such a method needs four key features in order to faithfully and clearly capture disagreements in the underlying discussion:

1. The issues need not be known a priori to the researcher - avoiding both human bias in issue selection, and improving scalability.
2. A single post can articulate a position on multiple, or hierarchical, issues - which happens abundantly in the real-world - and as such, the method should map the post to these issues.
3. Issues should be determined in their context - meaning both that the discussion as a whole aids the inference of the issues of a post, and that posts should be clustered to issues to allow aggregation for downstream applications.
4. Documents should be mapped to clear representations of these issues expressed as stance targets.

Existing work has several limitations that do not address these requirements. Previous stance target extraction work has done natural language generation of 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). Other methods do consider a corpus as a whole when determining disagreement (Paschalides et al., 2021; He et al., 2021), but do not produce a clear mapping of documents to stance targets.

In this paper, we make four contributions. First, we formalize this corpus issue disagreement map-

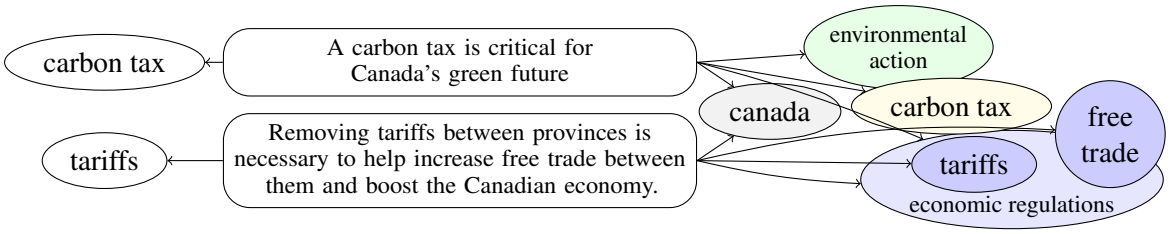


Figure 1: Comparison of assigning a single stance target to each document (left), versus assigning multiple hierarchical stance targets that overlap with other texts as proposed here (right).

ping activity 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 their corresponding stance) 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 `anonymous.4open.science/r/stancemining-E77B`

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) look at stance target extraction, but the focus is on mapping documents to a set of predefined stance targets. 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 hierarchical nature of stance targets, or the need for stance target clusters if we want to aggregate the data for further analysis. Having said that, we will compare our developed method against WIBA (Irani et al., 2024) in this work. Steel and Ruths (2024) use a combination of text embedding based clustering and manual domain knowledge to assign stance targets to documents, but the reliance on domain knowledge reduces the scalability of this method.

Polarized and Controversial Topics Work on topic cluster representation, such as Pham et al. (2023) and Grootendorst (2022), uses large language models (LLMs) to label clusters with an interpretable description, rather than the typical bag-of-words method. However, these methods are only designed for topic clusters, not stance target clusters. 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

Our effort to map the contours of disagreements in a corpus of social media posts motivates the COSTEx problem, which we conceive of as follows: given a corpus of documents, we seek 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 more than one stance target. Here we formalize this computational task.

Formally, for a corpus of documents $D = \{d_1, ..., d_N\}$, we want to find a set stance target clusters, $\hat{C} = \{\hat{c}_1, \hat{c}_2, ...\hat{c}_M\}$ where $\hat{c}_i \subset D$, and their corresponding stance targets $T = \{t_1, t_2, ..., t_M\}$. The COSTEx problem seeks $\mathcal{C}$ 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:

$$\sigma_S^2 = \frac{1}{|C|} \sum_{c_i \in C} Var(\{stance(t_i, d) : d \in c_i\})$$

This serves as 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 relevancy of targets by ensuring that the stance targets adhere to human judgments of stance targets, via comparison to labeled datasets and custom human annotation, and we can measure 'many stance targets' by measuring the mean number of targets per document:

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

3. **Large, Meaningful, Stance Target Clusters:** We want to optimize for large cluster sizes, 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 can compare mean cluster size, as a ratio of dataset size (to allow comparisons across datasets):

$$\mu_C = \frac{1}{|D||C|} \sum_{c \in C} |c|$$

Naturally, in most situations it will be impossible to perfectly satisfy all of these. As a result, 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 of these formulations.

Problems with Existing Approaches We can group most existing methods for finding issues of discussion in a corpus into two approaches: generating stance targets from each document using natural language generation (Irani et al., 2024; Akash et al., 2024; Zhang et al., 2021), or topic modeling a corpus, and inferring subjectivity/stance for each topic (He et al., 2021; Steel and Ruths, 2024; Rashed et al., 2021). However, each of these approaches have issues that must be addressed.

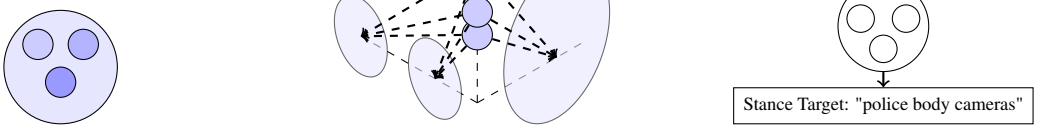
Our divergence from the first approach is as represented in Figure 1. Previous approaches have mapped one document to one stance target. For starters, this approach misses that documents can have multiple distinct stance targets. Furthermore, stance targets can be hierarchical (stance on 'free trade' implies stance on 'economic regulations'). And finally, if we want to compare stance across documents in aggregations, it is useful to select stance targets that are common across the corpus.

The second approach, topic modeling, naturally handles the desired aggregation process, and via hierarchical soft topic modeling, can approach a solution to the multiple stance target per document issue. 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 Figures 2a and 2b. And as shown in Figure 2c, 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.

4 Methods

Here, we propose our method that fulfills all the requirements we have detailed. The key idea of our method is that we want to cluster the documents in some fashion, in order to find large semantically related stance target clusters. But we don't want to directly cluster the documents, for reasons already stated in Section 3. If we first extract multiple stance targets from each document, and then cluster those stance targets, we can find large stance target clusters, where documents can naturally belong to multiple large clusters. By then using an LLM to assign a stance target to this cluster, using cluster information as the input, we can 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), and define it in Algorithm 3. Our method, by construction, attempts to achieve each criteria from Section 3 to a non-trivial degree.

The base stance targets are produced using an LLM fine-tuned on document - stance target pairs,



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

(b) Documents can map to one topic, but multiple stance targets.

(c) 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.

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*.

using diverse beam search (Vijayakumar et al., 2016) to generate mutiple targets. We cluster the targets using BERTopic (Grootendorst, 2022). The higher level stance targets are produced using an LLM with a few-shot prompt (shown in Appendix B.5). To avoid producing stance targets for each document that are paraphrases of each other, we remove stance targets based on having a high cosine similarity of sentence embedding (Reimers and Gurevych, 2019) (detailed in Appendix B.4). EC represents the best of two methods we designed and experimented with, as detailed in Section A.

4.1 Comparison Methods

We selected three methods to compare to EC on our task COSTEx. While there are substantial ways in which these methods do not address our proposed task, they do address it in other ways sufficiently that, for each of them, we consider it necessary to evaluate them against our task here.

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 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 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

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.

from our formulation, and comparing it to our comparison methods. We use the same fine-tuned models for EC and WIBA. We list all other experimental implementation details of each comparison method in Appendix B.

Datasets We use two large stance detection datasets to evaluate our 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, allowing us to test 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.

5.1 Automated Evaluation

Metrics As previously mentioned in our task formulation, 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 attempts to assess the extent to which the method outputs optimize for the objectives defined above.

- **Target F1:** This is the BERTScore F1 (Zhang et al., 2019) of the discovered targets, compared to the annotated dataset. 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 stance targets, and the recall by comparing each gold stance target to all predicted stance targets, and compute the F1 as normal, as defined in Appendix D.1. This metric measures adherence to Criterium 2.
- **Stance Retrieval F1:** This is 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 as usual, as defined in Appendix D.2.
- **Stance Variance:** Defined in Criterium 1.
- **Mean Number of Targets:** Defined in Criterium 2
- **Mean Cluster Size:** Defined in Criterium 3

Note that 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 likely reduce our target F1 score, as the stance targets will be further from the stance targets given in the base datasets. We therefore 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 a few objectives.

Results We report the supervised metrics from the mean of 5 runs for each method on each dataset in Table 2. We see that EC generally outperforms other methods, except on stance target F1 and precision, where WIBA outperforms it. We report the unsupervised metrics from the mean of 5 runs for each method on each dataset in Table 3.

Here we see that the methods that approach the task through aggregations (PaCTE, POLAR, and EC) tend to outperform WIBA through producing large average cluster sizes, and finding more stance targets in the corpus. PaCTE and EC are also able to find stance targets with a higher stance variance.

5.2 Human Evaluation

We created two human evaluation tasks to evaluate the method outputs. The first task presents a triad of documents, and has the annotator select which two documents go in the same stance target cluster. We measure how often the annotators agree with the clusters produced 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.

Method	Target			Stance Retrieval		
	F1 ↑	Prec. ↑	Recall ↑	F1 ↑	Prec. ↑	Recall ↑
VAST						
PaCTE	0.775	0.779	0.771	0.000	0.000	0.000
POLAR	0.512	0.524	0.501	-	-	-
WIBA	**0.910**	**0.930**	0.892	0.115	0.189	0.088
EC	0.898	0.889	**0.908**	**0.143**	**0.209**	**0.119**
EZ-STANCE						
PaCTE	0.766	0.768	0.763	0.000	0.000	0.000
POLAR	0.038	0.038	0.037	-	-	-
WIBA	**0.884**	**0.899**	**0.871**	0.145	0.200	0.120
EC	0.859	0.851	0.867	**0.157**	**0.201**	**0.140**

Table 2: Supervised metrics comparison across datasets and methods averaged across 5 runs for each method and dataset. We do not include stance results for POLAR as it does not assign stance to individual documents.

Method	Mean Num. Targets ↑	Stance Variance ↑	Cluster Size ↑
VAST			
PaCTE	1.212	**0.226**	0.088
POLAR	2.140	-	**0.124**
WIBA	1.000	0.108	0.002
EC	**3.190**	0.136	0.004
EZSTANCE			
PaCTE	1.038	**0.208**	0.021
POLAR	0.218	-	**0.038**
WIBA	1.000	0.019	0.001
EC	**3.380**	0.039	0.002

Table 3: Unsupervised metrics comparison across datasets and methods averaged across 5 runs for each method and dataset. Best metrics are indicated with arrows. Stance variance is absent for POLAR because the method does not assign a stance to documents.

Method	LSR Score	Example Output
PaCTE	-2.23	school,health,covid...
POLAR	-2.79	anyone
WIBA	1.51	medical law
EC	**2.23**	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).

Method	Agreement Pct.
PaCTE	0.19
POLAR	0.00
WIBA	**0.62**
EC	0.34

Table 5: Percentage of examples where annotators agreed with the clustering of a document triad, for each method.

We obtained 483 and 492 annotations for each task respectively, from 6 annotators. We show the full prompts and data generation given to annotators, and data generation process in Appendix C.

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 was 0.83, indicating inter-annotator agreement. For the stance target set comparison task, we use the Luce Spectral Ranking (LSR) inference algorithm (Maystre and Grossglauser, 2015) via the Choix library [1]. We report the results in Table 4. EC and WIBA are rated the highest, with POLAR and PaCTE rated poorly. For stance target cluster agreement scores, we simply record the number of times the human evaluator agreed with the method, which we report in Table 5. We see that WIBA, EC, and PaCTE obtain the best results for cluster evaluation, and POLAR obtains no agreement from evaluators.

5.3 Summary

We show the summed rank order of each method, for each metric, in Figure 4. This demonstrates the overall rank of the methods on the COSTEx task we introduce in this work.

[1] github.com/lucasmaystre/choix

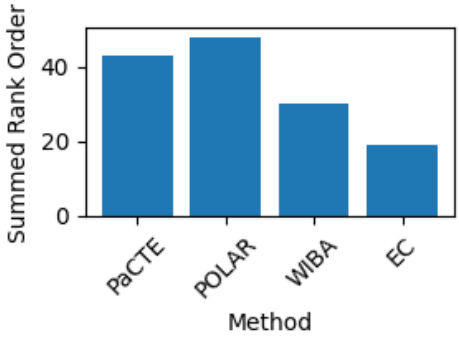


Figure 4: Summed rank order across all metrics for each method. We see that *EC* outperforms the other methods we trial from the literature across our metrics.

6 Discussion

POLAR needs to find many named entities to find polarized topics (being designed for news articles), and as such performs poorly on our chosen short text datasets, especially on the EZ-STANCE tweet dataset, as seen in Table 2. We also see poor evaluations of naming and clustering performance (Table 4 and 5). POLAR has multiple tunable parameters to potentially allow stronger performance, but this makes it less easily applicable.

PaCTE's use of LDA topic modeling and a small classifier model mean that it can quickly find large stance target clusters with seemingly high stance variance (Table 3). It produces reasonable stance target clusters, as shown by the cluster agreement percentage in Table 5. However, the naming of clusters with topic keywords results in a low evaluation score in Table 4.

WIBA's stance target extraction approach produces good stance targets, as shown in Tables 2 and 4. WIBA performs highest on our cluster agreement evaluation, as shown in Table 5. 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 (Tables 2 and 3).

EC improves over WIBA in stance target cluster size, stance variance, stance retrieval, and stance target set preference (Table 4). However, we also see that it under-performs WIBA on our cluster agreement evaluation (Table 5), and stance target precision. We infer that this is because EC applies more general, higher level, stance targets to each document, that have no parallel in the annotated datasets we use, and result in larger clusters that are linked by expressing a stance on more general issues that are sometimes too general to spot in our cluster agreement exercise. Overall we consider EC to be the most effective method tested here, as quantitatively confirmed by Figure 4.

However, there are key issues to further address with this method. As indicated by the lower agreement percentage in the cluster agreement task, some larger stance target clusters created by the method may not be coherent or intuitive when compared to manual stance target groupings. Another clear issue is the use of diverse beam search for stance target generation, which limits the maximum number of initial stance targets to the number of diverse generations compared to generating stance targets as a list output from the LLM. We will leave these issues for future work.

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. Our objective was to determine its usefulness relative to a topic modeling method (Grootendorst, 2022), as topic modeling is often used as an initial step in exploratory analysis of a dataset (e.g., Hobson et al. (2024), Falkenberg et al. (2022)).

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 1.4 million tweets from 1,907 prominent accounts in the Canadian media sphere (Bridgman et al., 2024). See Appendix E for implementation details.

Crucially, both EC and BERTopic (Grootendorst, 2022) require no notable parameter tuning and, so, are of equal complexity for a domain researcher (political scientist, in our case study) to use. Table 6 shows the largest stance target vs. topic clusters. From this list, alone, several analytical advantages of EC are clear.

Selecting 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 6, even 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 stance targets that have a small number of documents corresponding to them, with the intuition being that modal stance targets are more frequently good stance targets. In practice, for our dataset of 1.4 million documents, we find completely ignoring stance tar-

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 6: Comparing largest stance target clusters to largest topic clusters.

gets 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 6 highlights the informativeness of EC clusters in several ways. First, stance target clusters capture more of the posts than the largest topics, due to EC allowing posts to belong to multiple stance target clusters. Therefore, stance target clustering leaves out less of the discussion than the topic modeling in this corpus. Second, the stance targets capture the large ongoing issues of Canadian public discourse (Canada, Justin Trudeau, the Liberal Party), alongside more topical issues (Donald Trump's presidency, the B.C. election, the Israel-Palestine conflict), whereas many of these large ongoing 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. On the flip side, we see that EC needs improved stance target deduplication, 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 Figure 5. Having stance classifications on so many targets immediately surfaces key aspects of the discourse to the researcher: 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.

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.

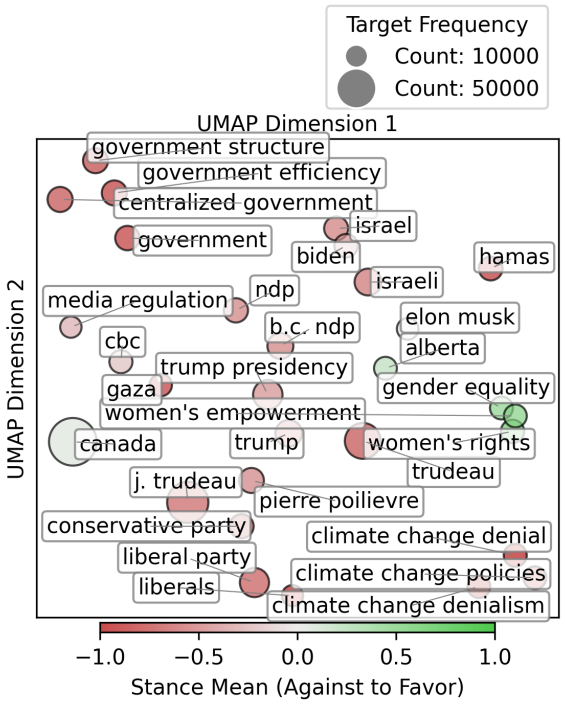


Figure 5: Map of top stance targets, sized by frequency, coloured by average stance.

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 more quickly understanding the discussion space of large and wide-ranging real world datasets, and put it to work aiding understanding of complex behaviors such as polarization and public opinion in our quickly changing information environments.

9 Limitations

Datasets In the course of using stance detection datasets for this work, we realized that the lack of high level stance targets in the datasets 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 the datasets used in the work to assess the extent to which we've found the lowest level stance target for each document. Future work would ideally find new datasets to evaluate stance targets in a supervised manner.

Methods Stance target de-duplication became apparent as 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', 'justin trudeau', 'trudeau') while avoiding false positives requires a carefully set distance threshold between embeddings. We will continue to iterate on this issue as we improve this method.

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 more accurate and faster by generating targets as a list. We are currently experimenting with this change, but did not have time to re-do our human evaluations for this work.

Task Formulation We have not addressed the issue of quantifying the extent to which a document maps to a stance target through entailment, instead mapping all stance targets equally to a document. We will leave this for future work.

Another issue with the task formulation that become apparent was the nature of optimizing for stance variance. This objective 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.

References

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*.

Aengus Bridgman, Alexei Abrahams, Thomas Bergeron, Blake Lee-Whiting, Haaya Naushan, Jennie Phillips, Zeynep Pehlivan, Saewon Park, Sarah Parker, Benjamin Steel, et al. 2024. The canadian information ecosystem.

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.

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.

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.

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*.

Ammar Rashed, Mucahid Kutlu, Kareem Darwish, Tamer Elsayed, and Cansın Bayrak. 2021. Embeddings-based clustering for target specific stances: The case of a polarized turkey. In *Proceedings of the International AAAI Conference on web and social media*, volume 15, pages 537–548.

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

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

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.

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 Method

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.

B.5 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 Human Evaluation

Human evaluators were fellow students from the authors' lab.

The exact text prompt given to human evaluators for the stance target cluster comparison task is shown in Prompt 3:

Prompt 3: 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 4:

Prompt 4: 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.

D Metrics

D.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:

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

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

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

D.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 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