I love pineapple on pizza != *I hate pineapple on pizza*: Stance-Aware Sentence Transformers for Opinion Mining

Vahid Ghafouri[1,2], Jose Such[3,4], Guillermo Suarez-Tangil[1]

[1]IMDEA Networks Institute, [2]Universidad Carlos III de Madrid, [3]King's College London, [4]VRAIN, Universitat Politecnica de Valencia

Correspondence: {vahid.ghafouri, guillermo.suarez-tangil}@imdea.org

Abstract

Sentence transformers excel at grouping topically similar texts, but struggle to differentiate opposing viewpoints on the same topic. This shortcoming hinders their utility in applications where understanding nuanced differences in opinion is essential, such as those related to social and political discourse analysis. This paper addresses this issue by fine-tuning sentence transformers with arguments for and against human-generated controversial claims. We demonstrate how our fine-tuned model enhances the utility of sentence transformers for social computing tasks such as opinion mining and stance detection. We elaborate that applying stance-aware sentence transformers to opinion mining is more computationally efficient than the classic classification-based approaches.

1 Introduction

Sentence transformers have become a cornerstone of Natural Language Processing (NLP), revolutionizing tasks like sentiment analysis, document retrieval, and text classification by capturing semantic meaning and contextual nuances. However, they grapple with a specific limitation that significantly impedes their utility in social computing — a critical domain where understanding sociopolitical stances is vital (e.g. Ghafouri et al. (2024)). In social computing, opinion mining and stance detection tasks demand the ability to discern between sentences expressing opposing stances on the same topic (Introne, 2023). Conventional sentence transformers often fall short in this regard, producing highly similar vectors even for sentences with contrasting opinions (Introne, 2023). For instance, the embeddings provided by the state-of-the-art sentence transformers for the sentences: "`The weather is good`" vs. "`The weather is NOT good`" manifest a high level of similarity in the embedding space, since both are talking about the quality of the weather, but with the exact opposite stance. In other words, they are *topically similar*, but *stance-wise dissimilar*.

This limitation is a major obstacle in tasks related to controversial sociopolitical topics where identifying differing perspectives is essential. Take, for instance, a situation where we want to automate the identification of the pro- and anti-abortion posts on Twitter through semantic search or semantic clustering of the sentence embeddings (Upadhyay et al., 2023). Using the default sentence transformers would group both pro- and anti-abortion tweets together since they are merely similar topic-wise. This disables the semantic method from detecting the stances of certain Twitter users with the automated and computationally cheap utilization of sentence transformers. An alternative, but computationally expensive, approach is to train a classifier capable of distinguishing the stances of pairs of statements (Küçük and Can, 2020; Sun et al., 2018; ALDayel and Magdy, 2021). However, this would require inputting pairs of sentences into the model at each point of pairwise comparison, with a subpar complexity in the order of $\binom{n}{2}$ times for pairwise comparison of n statements.

We address existing limitations by empowering sentence transformers, a computationally efficient method, with stance awareness. We extract and compose a rich dataset of supporting and opposing statements on controversial topics to fine-tune these models. Our objective is to lessen cosine similarities for statements representing opposing stances and increase similarities for congruent viewpoints. We perform this by fine-tuning a state-of-the-art sentence transformer with Siamese and Triplet networks using a contrastive and triplet loss function on top of the networks. These loss functions penalize the model for providing spatially close embeddings for contradictory, yet topically similar, pairs (triplets) of text.

In summary, our work makes the following con-

tributions:

1) Stance Awareness. We add stance awareness (§3) over topic-aware (§4) sentence transformers and verify its utility in opinion-mining tasks (§5).

2) Computational Efficiency. Classification-based stance-detection methods, require calling the model in the order of $\binom{n}{2}$ times for pairwise comparison of n sentences. We reduce this requirement to only n times (§5).

3) Experimental Insights. We gain several generalizable experimental insights (§5), including: i) Our novel *data-quality filtering* preprocessing step is useful for enhancing the model's quality and reducing the training workload. ii) The optimal value for $margin$ hyperparameters are moderate values. iii) *Parameter Efficient Fine-Tuning* minimizes the catastrophic forgetting, that in context, minimizes the fine-tuned model to forget the initial task of detecting *topic relevance*.

2 Motivation & Related Work

The main objective of this work is to enhance opinion mining and stance detection tasks. Thus, in this section, we motivate our work by examining the limitations of prior work.

Motivation: *Stance detection* is a vital task in social computing, aiming to identify an author's viewpoint (e.g., in favor, against, neutral) towards a specific topic (Biber and Finegan, 1988). Existing methods leverage state-of-the-art NLP architectures, such as BERT (Devlin et al., 2018), to classify the semantic relationship between a target sentence and a context sentence expressing a known stance.

Moreover, recent advancements in LLMs, have demonstrated significant potential in performing various NLP tasks, including stance detection, in a zero-shot setting without the need for fine-tuning (Qin et al., 2023).

However, both the *supervised classification-based* and the LLM-based approaches come with a significant *computational cost*. Since they involve feeding both the target sentence and the context sentence into the model simultaneously, for n pieces of text, they require calling the model $\binom{n}{2}$ times. This can be particularly problematic when dealing with large datasets or real-time analyses, such as analyzing stances in social media streams containing millions of posts. For instance, feeding dot-separated pairs of sentences to BERT-Base to predict their relationship (Devlin et al., 2018) (e.g., predicting similarity, predicting stance), would take an average inference time of 32ms per sentence pair on NVIDIA Tesla V100 GPU (Lamb et al., 2021). Comparing the stances of all the sentence pairs for 1,000 sentences will take $4.5\ hours \approx 32ms \times \binom{1000}{2}$.

Rise of Sentence Transformers: To address this problem of enormous computational workload for *sentence similarity tasks*, sentence transformers were introduced (Reimers and Gurevych, 2019). By fine-tuning BERT with Siamese networks, Reimers et al. proposed a way to generate semantically meaningful sentence embeddings that are spatially close for semantically similar sentences. These pre-generated embeddings removed the need for calling the models for every pairwise comparison, reducing the complexity to only n times for mapping the embeddings of n sentences; totaling: $32ms \times n$. Then, the similarity of every sentence pair is obtained by a swift calculation of the spatial distance of their pre-generated embeddings (approx 0.5ms per vector pair distance calculation). Thus, comparing all pairwise combinations for 1,000 sentences in terms of similarity would only take $4.5\ minutes \approx 32ms \times 1000 + 0.5ms \times \binom{1000}{2}$.

Need for Stance-Aware Sentence Transformers: The sentence transformers can solve the problem of computational inefficiency in sentence similarity measurement. Yet, if the task would be to compare the *stances* of sentence pairs on similar topics, current sentence transformers would perform far below ideal as they often confuse topic-wise similarity with stance-wise similarity; a limitation that has also been highlighted by previous work (Introne, 2023). This often results in assigning high similarity scores to statements that express opposing positions on the same topic. For example, "`I love pineapple on pizza`" and "`I hate pineapple on pizza`", two opposing stances on pizza, will be assigned a high similarity score as they are both talking about a taste towards the same food.

Another significant limitation of sentence transformers and similar models is their poor handling of negations and antonyms, as shown by recent research. Vahtola et al. (2022) demonstrate that sentence embeddings often fail to capture meaning-preserving transformations when one sentence includes a negated antonym of the other, such as "`I am not guilty`" and "`I am innocent.`" This deficiency further exacerbates the challenge of stance detection, where subtle shifts in meaning can completely reverse the stance.

Developing the ability to fine-tune sentence transformers for spatial dissimilarity in opposing viewpoints has the potential to significantly advance online opinion mining and stance detection. Take, as a running example, a case where we want to figure out the stances of several politicians on *abortion rights* using their Twitter timelines. A solution aided by sentence transformers, as we demonstrate in §5.3, can query anti- and pro-abortion statements such as "`abortion is murder`" and "`abortion is healthcare.`" Then, after embedding both queries and timelines into vectors using sentence transformers, we can systematically infer tweets with high spatial similarity to the pro (anti) abortion query and their stance. Another huge computational advantage of this approach is that the embeddings generated for the timelines can be saved and used for other queries in the future. For example, we can quickly generate a pair of queries representing pro- and anti-gun-carrying rights and run them on the same timelines that are already vectorized to mine the users' opinions on gun control.

Ideal Stance Detection Method: Based on the considerations above, in summary, an ideal stance detection method should satisfy three major requirements: *R1) Computational Efficiency* which is not addressed in classification-based methods, but it is in sentence transformers; *R2) Stance Awareness*, which is not addressed in sentence transformers yet, but can revolutionize stance detection methods if the following challenge was to be addressed properly; *R3) Maintaining Topic Awareness:* Crucially, when empowering sentence transformers with stance awareness, an important challenge would be to avoid *catastrophic forgetting*. This means that sentence transformers primarily pretrained to detect topically relevant texts should retain this primary functionality after being fine-tuned for stance awareness.

3 Methodology

In this section, we elaborate on the fine-tuning architecture and our experimental settings for strategizing the fine-tuning process. Figure 1 summarizes the entire pipeline of our approach, including fine-tuning (§3), data-preparation (§4), and the semantic-search application (§5).

3.1 Argument base: Anchor, Positive and Negative statements

The fine-tuning architecture for adding stance awareness requires pairs and triplets of statements with labels regarding their argumentative stance toward each other. Pairs are topically relevant statements that either Agree (Ag) or Oppose (Op) with each other whereas, every triplet, in the context of this task, is composed of an Anchor (An) which is an initial claim (parent claim), a Pro (P) argument that supports the parent claim, and a Con (C) argument that disagrees with the parent claim. We give grounded examples of such statements in our dataset (§4.1).

3.2 Architecture: Siamese and Triplet Model

Our approach leverages Siamese and Triplet network architectures, which are the underlying methods used to train sentence transformers. In this section, we briefly introduce both methods in the context of fine-tuning argumentative statements.

As both Siamese and Triplet architectures are well established in the literature, to avoid redundancy, we only introduce the main idea behind them here and detail their formulations in Appendix 3.3.

3.3 Siamese and Triplet Networks

Siamese Network with Contrastive Loss: A Siamese network (Koch, 2015) is a neural network consisting of two identical subnetworks, termed "*twins*," that share the same architecture and parameters. The Siamese network is specifically designed for tasks that involve comparing and contrasting pairs of input data.

In our case, the Siamese network takes pairs of arguments (supporting or contradictory) independently and computes their corresponding embeddings. These embeddings encapsulate the essential information of the arguments. Then, we use the contrastive loss function as in Eq. 3.3 to fine-tune the model such that produces close (distant) embeddings for aligning (contradictory) arguments.

$$\text{Contrastive Loss} = y_i \times D(E_i^1, E_i^2) + (1 - y_i) \times \max(margin - D(E_i^1, E_i^2), 0)$$

E_i^1 and E_i^2 are embeddings, i.e.: the outputs of the model which denote the projection of statement pairs into the embedding space. $D(E_i^1, E_i^2)$ is a distance metric, often the Euclidean or cosine distance, which measures the dissimilarity between

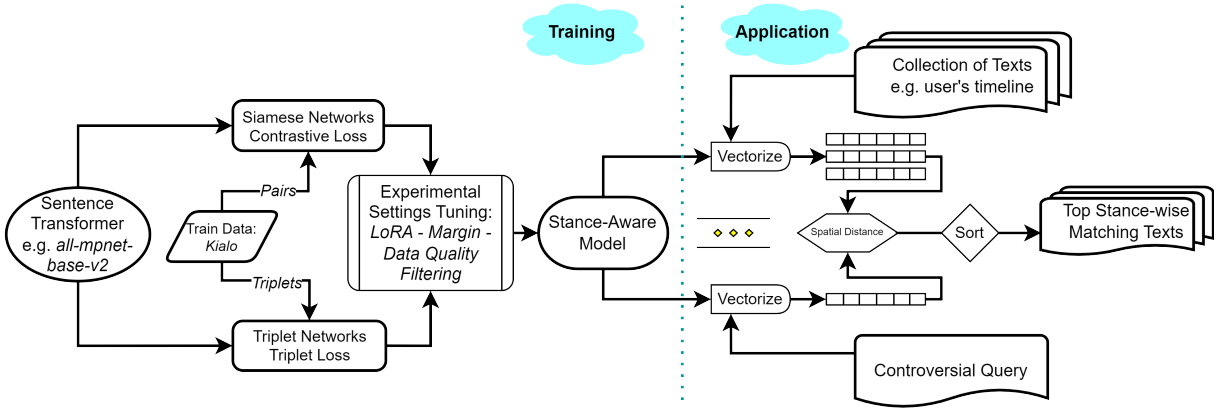


Figure 1: Our methodological pipeline and its application process.

the two embeddings. Smaller $D(E_i^1, E_i^2)$ indicates greater similarity. Next, $margin$ is a hyperparameter that defines the separation margin. If the distance between similar samples $D(E_i^1, E_i^2)$ for the opposing statements ($y_i = 0$) is smaller than the $margin$, the loss function incurs a penalty. On the other hand, where E_i^1 and E_i^2 agree with each other ($y_i = 1$), the spatial distance between E_i^1 and E_i^2 incurs penalty in loss function.

Triplet Network with Triplet Loss: The *Triplet network* (Hoffer and Ailon, 2015) extends the idea of shared parameterization so that the model focuses on the relationships among triplets of inputs, adding more context to the samples. Our architecture uses argument-base statements as defined in §3.1 to form triplets. *Triplet loss* on top of the Triplet architecture is designed to enforce a specific learning objective: the model is trained to minimize the distance between the anchor (parent claim) and the positive example (Pro argument) while maximizing the distance between the anchor and the negative example (Con argument). This is formulated in Eq. 3.3:

$$\text{Triplet Loss} = \sum_{i=1}^{N} \max(D(E_i^{An}, E_i^{P}) - D(E_i^{An}, E_i^{C}) + \text{margin}, 0)$$

where E_i^a, E_i^p, and E_i^c, denote the embeddings of the parent claim (anchor), supporting argument (pro), and opposing argument (con).

Hybrid: In our work, we also test the Siamese and the Triplet networks together, which we call *Hybrid* throughout this paper. We arrange this by fine-tuning the model with the Triplet network for half of the epochs and then fine-tuning with the Siamese network on top of it for the other half of the epochs. Our hypothesis is that this setting can combine the contextualization strengths of triplets while maintaining the direct comparison between data pairs from the Siamese network.

3.4 Fine-tuning Strategy

We next describe the strategy we use to optimize our fine-tuning task, detailing how we iterate over different values of key hyperparameters and experimental settings. For our base model, we use a light-weight (420MB) state-of-the-art[1] pretrained sentence transformer model "*all-mpnet-base-v2*"[2] that is widely used in previous computational social science literature. Details on training costs and utilized packages can be found in Appendix A.3. There are also newer generations of heavy-weight LLM-based text embedders available online, yet, since this paper is oriented toward demonstrating the *feasibility* of obtaining a stance-aware sentence transformer, a light-weight sentence transformer with competitive performance would suffice for answering our main research question. In any case, we also show in Appendix A.4 that LLM-based text embedders would face the same issues.

Margin: A larger $margin$, both in contrastive and triplet loss, enforces a greater separation between contrasting stances, potentially enhancing stance discrimination but risking over-separation where nuanced differences are overlooked. Our experimentation involves finding the optimal $margin$ that balances precision and recall in the training. We tune this hyperparameter with a grid search over the range (0.1, 1, step = 0.1).

Data Quality Filtering: This step aims at filtering noisy and low-quality inputs to the model from

[1] www.sbert.net/docs/pretrained_models.html

[2] huggingface.co/sentencetransformers/all-mpnet-base-v2

opposing statements. Take for instance the following two statements extracted from two posts with opposing views around abortion: 1) "`Abortion is murder`" (A) and "`I disagree`" (B). In the absence of comprehensive context and background information, these two sentences alone may not represent genuine opposing stances. Sentence B is not particularly an anti-abortion statement in its nature unless one is aware of the context in which it has been used. Yet, we are training the model to be used for converting short phrases into vectors independent of their context. Hence, compelling the model to represent statements A and B as contrasting statements could introduce noise and hinder overall model performance.

The data quality filtering step that we introduce, seeks to address this concern by prioritizing relevant and contextually meaningful instances during training. We initially employ the "*all-mpnet-base-v2*" model to compute the cosine similarity between instances (pro-con pairs) in the training set and filter out statements that are lower than a threshold. For triplet networks, we filter out instances where the lowest pair-wise cosine similarity between all three sentences is lower than the threshold. We experimentally try different thresholds and retain 50% and 30% for contrastive and triplet networks respectively based on the major gaps in the frequency histogram of the training data.

Parameter Efficient Fine-Tuning with LoRA: We employ Low-Rank Adaptation (LoRA) (Hu et al., 2021) which is designed for computationally efficient fine-tuning of large language models, while also mitigating the risk of catastrophic forgetting. Traditional fine-tuning can be computationally expensive, especially during hyperparameter experimentation. LoRA addresses this challenge by introducing trainable adapter modules into specific layers, allowing targeted adjustments to the pre-trained model without modifying all the weights. We specifically target attention layers with a rank of 32 (Wang et al., 2023), reducing computational costs compared to full tuning.

To reduce the training workload, we only apply our iterative grid-search over other experimental settings with LoRA and select the best experimental setting for a round of full training as well.

4 Datasets

4.1 Training Data: Kialo

We use the Kialo platform (*www.kialo.com*) to create pairs and triplets of agreeing and opposing arguments on certain topics which are the essential inputs of the Siamese and Triplet networks (cf. §3.2). Kialo is an online debate platform where users create and discuss controversial topics. Each debate on Kialo is formatted in a tree structure, where the root/parent node is the main topic (initial thesis) of the debate and the branch/child nodes are the arguments that support or oppose the main topic. Furthermore, each of the branch/child arguments can turn into parent/root arguments to subsequent branch/child arguments supporting or opposing them. Figure 2 shows a sample Kialo discussion on "`whether Ukraine should surrender to Russia or not.`"

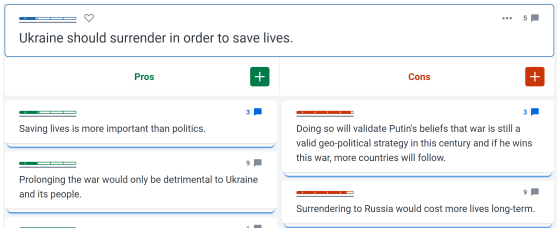


Figure 2: Sample discussion on Kialo website.

The raw tree-formatted data of Kialo was collected by (Ghafouri et al., 2023). This dataset contains a collection of discussion trees for a variety of controversial topics such as "`Should animal testing be banned?`", "`Should the government provide free healthcare?`", "`Should the death penalty be abolished?`", etc. The dataset has 5,631 discussions with 430,034 arguments in total and a balanced proportion of supporting arguments and counter-arguments.

We make a 9:1 train-test split of the discussions. Table 1 reports the number of generated pair and triplet samples (detailed description of the procedure can be found in Appendix A.1). Note that our split is based on the entire discussion trees, not the individual arguments, i.e.: the sampled pairs or triplets in the test set do not originate from the same discussion as in the training set. This ensures the test set assesses the performance in challenging scenarios where the supporting or contradicting pairs of arguments are from topics not seen by the model before.

Data	Train (90%)	Test (10%)
Discussion Topics	4430	493
Generated Pairs	972395	112724
Generated Triplets	303081	34453

Table 1: Kialo dataset's size.

4.2 Baseline Data: STS-B

As with every other fine-tuning, our task is also subject to the risk of *catastrophic forgetting* which refers to the cases where after fine-tuning, as a result of over-training on the newer task, the model forgets its ability to perform the older task it was initially trained to do (McCloskey and Cohen, 1989). In this context, the primary task of sentence transformers was to detect semantic similarity (regardless of stance). Thus, we need a separate validation on a dataset annotated for semantic similarity to assess how far fine-tuning the models for stance-awareness, would forget this primary task.

The Semantic Textual Similarity Baseline (STS-B) dataset is a widely recognized benchmark designed to assess the ability to compute semantic similarities between pairs of sentences. It comprises pairs of sentences with similarity scores ranging from 0 (no semantic overlap) to 5 (semantic equivalence). We only use the test set which consists of 1,379 pairs. These pairs span over diverse topics, including news headlines, forum discussions, and product reviews.

4.3 Out of Distribution Data: SemEval-2014

As our out-of-distribution test data, we look into the "*SemEval-2014: Task 1*" dataset, a widely used contradiction detection dataset that does not overlap with Kialo. The dataset contains a variety of sentence pairs annotated as *Neutral* (5611), *Entailment* (2857), and *Contradiction* (1459). The *Entailment* and *Contradiction* pairs are relevant topic-wise but are aligned or contradictory stance-wise, yet the *Neutral* pairs can either be topically relevant or be totally irrelevant statements.

4.4 Application Data

Finally, to demonstrate the applicability of our model to semantic search of controversial statements, which is one of the main motivations for our work, we use a publicly available dataset of tweets from congresspeople.[3] The dataset contains the timeline of 564 congresspeople (*Democrats*: 292, *Republicans*: 270, *Independent*: 2). In total 2.3M tweets (*Democrats*: 1.4M, *Republicans*: 840K, *Independent*: 9K) of the congresspeople are collected.

[3] https://github.com/alexlitel/congresstweets/tree/master

5 Experiments, Results, & Observations

We next describe our experiments and results after applying our method to fine-tune the sentence transformer. We first test the performance of all the fine-tuned models on a test set from the Kialo and STS-B datasets (§5.1 and §5.2). Using the best-performing model, we evaluate how the learning transfers to another dataset (§A.5). Finally, we showcase its application on semantic search for opinion mining (§5.3).

5.1 Validation on Kialo

As the first step of the validation, we create frequency plots of cosine similarities over the 10% test-set of the Kialo dataset. Figure 3a reveals that the original model struggles to distinguish stances, as the pro (green) and con (red) distribution curves align closely. The green and red frequency distribution curves represent the cosine similarities between pro and con statement pairs. The alignment of the curves shows that the original model does not effectively differentiate between pro and con statement pairs. Figure 5 shows that even the recent state-of-the-art heavy-weight LLM-based text embedders suffer from the same limitation.

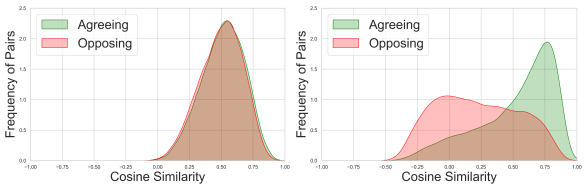
On the other hand, Figure 3b shows the same curves for one of our best (settings: *Hybrid*, *margin = 0.4*, *LoRA*) fine-tuned versions of the model. We see a notable shift in the distribution of pro statements (green) to the right side and a corresponding shift in the distribution of con statements (red) to the left side.

> **Observation:** This significant shift indicates that our fine-tuned model has become stance aware, effectively separating pro and con statements even on previously unseen topics, partly fulfilling requirement R2 as in §2.

To quantify the performance of this separation we calculate the KL-Divergence between the cosine similarity distributions of Opposing pairs and cosine similarity distributions of Agreeing pairs. A higher amount of KL-Divergence translates into a desirable higher separation between Agreeing and Opposing statements by the model. Table 2 reports

			Margin									
Model Type	**Filtering**	**LoRA**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
Siamese	None	yes	0.03	0.21	0.41	0.37	0.37	0.34	0.34	0.36	0.35	0.36
Siamese	$< 50\%$	yes	0.01	0.24	0.38	0.44	0.34	0.31	0.38	0.37	0.37	0.38
Triplet	None	yes	0.31	0.36	0.39	0.40	0.39	0.40	0.37	0.36	0.35	0.33
Triplet	$< 30\%$	yes	0.26	0.37	0.42	0.42	0.41	0.39	0.38	0.36	0.36	0.34
Hybrid	$< 30\%$ & $< 50\%$	yes	0.23	0.35	0.44	0.44	0.44	0.45	0.41	0.39	0.38	0.36
Hybrid	$< 30\%$ & $< 50\%$	no	0.66	0.72	0.67	0.71	0.69	0.63	0.66	0.62	0.61	0.59
Original "all-mpnet-base-v2"			0.004									

Table 2: KL Divergence Between Agreeing and Opposing statements' distributions in Kialo Test Set.



(a) Original Model. (b) Fine-Tuned Model.

Figure 3: Comparison of Model Distributions.

results for different combinations of the experimental settings. The data quality filtering threshold is set to *None*, below 50% for pairs in the Siamese network, and below 30% for minimum pairwise similarity in any pairs of a triplet in the Triplet network. Recall that we apply LoRA to all models and we experiment with further fine-tuning over the best-performing configuration (the last Hybrid row in this case). Finally, the $margin$ hyperparameter is iterated over in steps of 0.1 to obtain the best combination.

Observation: Our fine-tuning approach yielded significant performance leap, with all fine-tuned models outperforming the original model by a substantial gap. Hybrid narrowly wins among LoRA models while the fully fine-tuned model outperforms all. LoRA being an efficient transformer, significantly contributes towards requirement R1.

5.2 Sentence Similarity Baseline

Next to the model's performance on the task for which the model had been trained (*primary task*), we assess the amount of catastrophic forgetting introduced when fine-tuning. Table 3 reports the models' performance on the STS-B dataset, the primary task. For this, we use the Spearman correlation between two cosine similarities: 1) over sentence pairs provided by the model (predicted values), and 2) over pairs annotated by humans (true values ranging [0, 5]). Higher cosine similarity values indicate better model performance in capturing semantic similarity between sentence pairs, a proxy for low catastrophic forgetting.

We see that the performance of the base model has a strong correlation of 0.83, which means that it performs well with the primary task. While, as expected, none of the fine-tuned models outperforms the base model in the primary task, we see comparative performances (also at 0.83) of some LoRA fine-tuned models, especially for lower $margin$s in the range [0.1, 0.4]. However, the base model shows a very poor performance in the new task (0.004 divergence, as shown in the previous section). Conversely, the fully fine-tuned model (LoRA = no) shows subpar performance in the primary task. This is because catastrophic forgetting is higher in fully fine-tuned models, as expected when dealing with parameter-efficient fine-tuning as identified by prior work (cf. §3.4).

While fine-tuning creates a tension between the objective of the *primary* and the *new task*, our LoRA models significantly reduce this tension by eliminating catastrophic forgetting, unlike the base model, while maintaining comparable results when compared to the base model in the primary task. This demonstrates the model's robustness in adapting to a new task while retaining previously learned knowledge, satisfying R3.

For selecting the best model and parameters, we consider the trade-off between its performance on the two tasks (new vs. the primary task) as discussed above. As mentioned, Table 2 represents the stance-aware results, i.e.: the new task, where the best margins here are in the range [0.4, 0.7]. Instead, in the primary task, lower margins in the range [0.1, 0.4] cause the least catastrophic forgetting (as we observe in Table 3). Thus, we select 0.4, where the two ranges meet, and the LoRA fine-tuned version of Hybrid in what follows.

			Margin									
Model Type	**Filtering**	**LoRA**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
Siamese	None	yes	0.73	0.77	0.78	0.79	0.81	0.82	0.82	0.82	0.81	0.79
Siamese	$< 50\%$	yes	0.77	0.79	0.79	0.80	0.82	0.83	0.82	0.81	0.80	0.79
Triplet	None	yes	0.83	0.83	0.82	0.81	0.80	0.79	0.78	0.76	0.73	0.71
Triplet	$< 30\%$	yes	0.83	0.83	0.82	0.81	0.81	0.80	0.78	0.77	0.75	0.73
Hybrid	$< 30\%$ & $< 50\%$	yes	0.83	0.83	0.81	0.80	0.79	0.78	0.77	0.76	0.74	0.72
Hybrid	$< 30\%$ & $< 50\%$	no	0.72	0.71	0.68	0.63	0.59	0.53	0.51	0.49	0.47	0.45
Original "all-mpnet-base-v2"			0.83									

Table 3: Performance of models on STS-B test set (Spearman correlation).

5.3 Application: Semantic Search

Once demonstrated the performance of our models, we showcase the practical implications of performing stance identification and its potential to enhance social computing tasks. A practical use-case of the stance-aware model is retrieving text with certain stances in corpora through the use of semantic search.

We generate two controversial statements with the exact opposite viewpoints on abortion: "*Abortion is healthcare*" and "*Abortion is murder*." Then, we query these two statements from the 2.3M tweets of the congresspeople dataset (cf. §4.4). As it is typically done in semantic search with S-BERT, we first convert each tweet and query into vectors; separately using the original and the fine-tuned model. We then compute the cosine similarity between the query embeddings and tweet embeddings, applying similarity thresholds, suggested by (Iqbal et al., 2023), to filter out less relevant tweets. The more aligned the stances of the remaining tweets with the query, the better the model is in stance awareness.

Table 4 shows the results of the alignments, and Table 5 offers an excerpt of the top matching results (highest *cosims*) with the pro-abortion query.Looking at the summary of our results in Table 4, we see that when we shift from the original model to the fine-tuned one, the alignment precision of the model from Twitter rises from 76% to 91% for pro-abortion (Democrat) and from 67% to 80% for the anti-abortion (Republican) query. This means that desirably 91% (80%) of the top similar results for a Democrat (Republican) query has correctly matched with the tweets of Democrat (Republican) congresspeople. This experiment shows that our method can be utilized to perform robust and efficient opinion mining.

These results are the demo results for one of our best model settings (*Hybrid* architecture, *margin = 0.4*, *LoRA*). To see the results of alignment precision for other fine-tuned models, see Table 8 in Appendix A.6.

Disclaimer: Despite §5.1, §5.2, and §A.5, the main objective of this section was not to ***evaluate*** the stance awareness of the fine-tuned model, but to elaborate ***how*** such a stance-aware language model can be used in practice to improve opinion mining tasks. That's why we focused on a case study of *abortion-related tweets*. More experiments can be done around other controversial topics in real-world applications of the model.

6 Discussion

This work tackles the critical challenge of balancing three essential requirements in NLP tasks: computational efficiency (R1), stance awareness (R2), and maintaining topic awareness (R3). We address these challenges by proposing a novel approach that leverages fine-tuning while mitigating its drawbacks. We reviewed how *prior work* fails to meet these three requirements together in §2 and we showed how our work (§3) addresses them (§5), we next summarize the main findings of our paper a discuss their implications and limitations.

Computational Efficiency. Our approach makes opinion mining efficient, only needing to call the model n times for mapping the embeddings of n sentences, that is, linear with the number of sentences. A limitation may arise in how much a single statement used as a query might encompass all variations of the stance on a certain topic. An important consideration is to maintain sufficient diversity in query selection to account for all parts of the spectrum of opinions.

A balance is feasible. our work demonstrates the feasibility of achieving a balance between efficiency, stance awareness, and topic coherence through careful fine-tuning strategies. This approach can be further explored and adapted for various NLP applications, particularly those requir-

Model	Query	Affiliation	Cosim Threshold	R	D	Alignment Precision
Original	"Abortion is healthcare."	Democrat	0.70	31 ✗	98 ✓	**76%**
Fine-Tuned	"Abortion is healthcare."	Democrat	0.70	4 ✗	43 ✓	**91%**
Original	"Abortion is murder."	Republican	0.60	95 ✓	46 ✗	**67%**
Fine-Tuned	"Abortion is murder."	Republican	0.60	12 ✓	3 ✗	**80%**

Table 4: Alignment Precision for semantic search on congresspeople tweets with abortion-related queries. D: Democrat alignment, R: Republican alignment.

	Text (Query/Tweet)	Party	Aligned?
	QUERY: "Abortion is healthcare."	Dem	
Original	In case anyone forgot – abortion is NOT healthcare.	Rep	✗
Original	Reminder: abortion is health care.	Dem	✓
Original	Stop pretending abortion is healthcare...	Rep	✗
Original	... I have to say this once again, but abortion is NOT healthcare. #ProLife	Rep	✗
Original	... A procedure where a successful outcome is the death of a living human is not healthcare.	Rep	✗
FineTuned	Just a reminder: abortion is healthcare. #SOTU	Dem	✓
FineTuned	... EVERY woman has the constitutional authority to make decisions about their own body ...	Dem	✓
FineTuned	Reminder: abortion is health care.	Dem	✓
FineTuned	... Roe v. Wade is the law of the land and we have to ensure it will stay that way...	Dem	✓
FineTuned	Reproductive care is health care...	Dem	✓

Table 5: Most similar semantic search results for a pro-abortion query for the Original and Fine-Tuned models.

ing robust stance-aware analysis on large datasets.

7 Conclusion

Overall, our work paves the way for stance-aware sentence transformers, offering a powerful tool for social computing tasks like opinion mining. Our work demonstrably surpasses the state-of-the-art in *stance awareness of sentence transformers*, achieving significant improvements in distinguishing stances across in-distribution (Kialo test-set) and out-of-distribution (SemEval 2014 and Twitter) datasets. By designing an innovative *model architecture*, we observed a measurable improvement of results with the Hybrid (combination of Siamese and Triplet) model. We implemented a *data filtering* approach by removing low cosine similarity pairs, which probed a unique experimental contribution that effectively mitigated the impact of "low-quality" human-generated data within the training set. This also resulted in an improvement of the model performance, while significantly reducing the train-set size and thus the training time.

Two main future steps in this direction can significantly improve the quality of the task: 1) Improving general-purpose sentence transformers using (LLMs) and extensive datasets, such as recently developed Open AI's text embedders[4]; 2) Developing dedicated datasets tailored to social media platforms like Twitter and Mastodon and fine-tuning the general-purpose sentence transformer on such datasets. This will enable the model to learn stance awareness in the context of the targeted social networks of analysis. Nevertheless, our model, which is fine-tuned on Kialo arguments also demonstrated a promising performance on the Twitter data. This forecasts an even brighter future for models that are specifically fine-tuned on online social media data for the same task.

Reproducibility: We open-source both code and models to foster reproducibility.[5]

Acknowledgments

This project was funded by TED2021-132900A-I00, from the Spanish Ministry of Science and Innovation, with funds from MCIN/AEI/10.13039/501100011033, and the European Union-NextGenerationEU/PRTR; and supported by UKRI through REPHRAIN (EP/V011189/1), the UK's Research Centre on Privacy, Harm Reduction and Adversarial Influence online, as part of its PROM project. G. Suarez-Tangil has been appointed as 2019 Ramon y Cajal fellow (RYC-2020-029401-I) funded by MCIN/AEI/10.13039/-501100011033 and ESF Investing in your future. The authors also thank Dr. Mansooreh Karami for her invaluable advice.

[4] https://platform.openai.com/docs/guides/embeddings

[5] https://github.com/vahidthegreat/StanceAware_SBERT

Limitations

The main goal of this paper was to demonstrate the feasibility of obtaining stance awareness in sentence transformers. Thus, the language model of analysis in this paper is merely limited to "*all-mpnet-base-v2*", the widely used state-of-the-art sentence transformer in SBERT leaderboard list [6] which is *light-weight* and suitable for the purpose of our experiments. Yet, more heavy-weight LLM-based text-embedders are not explored in this paper. We nevertheless, report the stance unawareness of "*NV-Embed-v1*", the best performing *Massive Text Embedder* in MTEB leaderboard [7], in §A.4 but do not apply our fine-tuning experiments as the lighter model we use satisfies our main goal (demonstrating the feasibility of obtaining stance awareness) with a significantly lower computational cost. Yet, for those interested in improving the quality of the model and the task, it is possible to fine-tune any state-of-the-art text embedder by a simple replication of our experimental pipeline using the code that we make publicly available (see Reproducibility above).

Another limitation of our paper is in the scope we demonstrated the application of the model in §5.3. We only showcased the application of the finetuned model on semantic search over tweets related to *abortion*. The reason is that the main purpose of §5.3 was not to validate the model like §5.1,§A.5, and §5.2 but to explain *how* the model can be used in opinion mining and computational social science tasks. Similar experiments on other controversial topics such as gun-control, war on Ukraine, etc. are left for future works.

Ethical Considerations

The datasets in §4.3, §sec:datasets:sts, and §4.4 are publicly available. The Kialo dataset in §4.1 is obtained from another published work by (Ghafouri et al., 2023) and will be available only to researchers upon request. All the datasets are anonymized and no personal or private information is handled. All the results are honestly reported, with their source code available on GitHub for replication and validation. Moreover, Gemini, a generative language model, has been utilized in all the sections solely for the task of polishing and summarizing the text.

[6] www.sbert.net/docs/sentence_transformer/pretrained_models.html

[7] huggingface.co/spaces/mteb/leaderboard

References

Abeer ALDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Douglas Biber and Edward Finegan. 1988. Adverbial stance types in english. *Discourse Processes*, 11(1):1–34.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Vahid Ghafouri, Vibhor Agarwal, Yong Zhang, Nishanth Sastry, Jose Such, and Guillermo Suarez-Tangil. 2023. Ai in the gray: Exploring moderation policies in dialogic large language models vs. human answers in controversial topics. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 556–565.

Vahid Ghafouri, Faisal Alatawi, Mansooreh Karami, Jose Such, and Guillermo Suarez-Tangil. 2024. Transformer-based quantification of the echo chamber effect in online communities. In *ACM Conference on Computer-Supported Cooperative Work and Social Computing*, CSCW2 '24.

Elad Hoffer and Nir Ailon. 2015. Deep metric learning using triplet network. In *Similarity-Based Pattern Recognition*, pages 84–92, Cham. Springer International Publishing.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Joshua Introne. 2023. Measuring belief dynamics on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):387–398.

Waleed Iqbal, Vahid Ghafouri, Gareth Tyson, Guillermo Suarez-Tangil, and Ignacio Castro. 2023. Lady and the tramp nextdoor: Online manifestations of real-world inequalities in the nextdoor social network. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):399–410.

Gregory R. Koch. 2015. Siamese neural networks for one-shot image recognition.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Comput. Surv.*, 53(1).

Alex Lamb, Di He, Anirudh Goyal, Guolin Ke, Chien-Feng Liao, Mirco Ravanelli, and Yoshua Bengio. 2021. Transformers with competitive ensembles of independent mechanisms.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384, Singapore. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. 2018. Stance detection with hierarchical attention network. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2399–2409, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Rishabh Upadhyay, Gabriella Pasi, and Marco Viviani. 2023. A passage retrieval transformer-based reranking model for truthful consumer health search. In *Machine Learning and Knowledge Discovery in Databases: Research Track*, pages 355–371, Cham. Springer Nature Switzerland.

Teemu Vahtola, Mathias Creutz, and Jörg Tiedemann. 2022. It is not easy to detect paraphrases: Analysing semantic similarity with antonyms and negation using the new SemAntoNeg benchmark. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 249–262, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yiming Wang, Yu Lin, Xiaodong Zeng, and Guannan Zhang. 2023. Multilora: Democratizing lora for better multi-task learning.

A Appendix

A.1 Generating Training Pairs and Triplets

To form pairs for the Siamese Networks (see §3.2), we choose to use a combination of child-to-parent and child-to-child pairs of arguments from the Kialo dataset. Child-to-parent pairs are pairs consisting of a child's argument versus its parent's argument with which it is agreeing or disagreeing. Child-to-child pairs are pairs where both arguments are children of a unique parent argument with which they agree or disagree. Table 6 illustrates samples of child-to-child and child-to-parent pair generation from the example discussion in Figure 2; **i.e., two *cons* of a unique parent will also be labeled as *Agreeing* to each other when paired together.** After forming all the possible sentence pairs, we obtain 420,838 child-to-parent pairs and 713,725 child-to-child pairs, a total of 1,134,663 argument pairs.

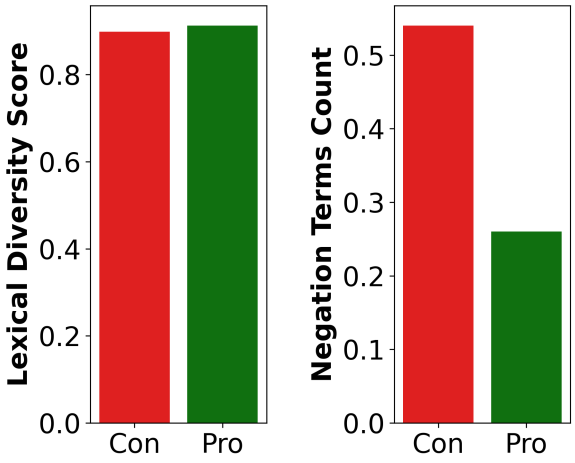
For the Triplet networks, our samples are composed of triplets of statements. Each triplet consists of an anchor statement (parent claim), a supporting statement (a child "pro" argument) that agrees with the anchor, and an opposing statement (a child "con" argument) that disagrees with the anchor. We derive the triplet samples by iterating over every parent claim and sampling every possible pairwise combination of its pro and con child arguments. Table 7 shows a sample triplet from the Kialo discussion depicted in Figure 2.

A.2 Semantic Stats on Train Data

Lexical Diversity: Lexical diversity is a measure of the richness of the vocabulary used in a sentence. It is calculated as the ratio of unique words to the total number of words in a sentence. A higher lexical diversity score indicates a more varied vocabulary. The average lexical diversity for the *Pro* and *Con* statements is illustrated in Figure 4a.

Negation Terms:

Negation terms ("*no*", "*not*", "*never*", "*none*", "*cannot*", "*n't*", "*neither*", "*nor*") and similar words are critical in determining the stance of a sentence. Figure 4b shows the average number of negation terms used in arguments labeled as *Pro* and *Con* in the Kialo train set.



(a) Average Lexical Diversity by Stance

(b) Average Negation Terms Count by Stance

Figure 4: Comparison of Lexical Diversity and Negation Count by Stance. The green bars represent the *Pro* stance, while the red bars represent the *Con* stance.

A.3 Training Cost and Packages

In our experiments, we utilized the "*all-mpnet-base-v2*" model for fine-tuning. This model, which has a size of approximately 420 MB, contains

Child-to-Parent Sample Pairs	Child-to-Child Sample Pairs
(*Saving lives is more important than politics*, *Ukraine shall surrender to save lives*) **Pair Label =** Agreeing	(*Saving lives is more important than politics*, *Surrendering to Russia costs more lives long-term*) **Pair Label =** Opposing

Table 6: Example of argument pair creation.

Anchor	Pro	Con
Ukraine should surrender in order to save lives	*Saving lives is more important than politics*	*Surrendering to Russia would cost more lives long-term*

Table 7: Example of triplet creation.

a total of 111,845,760 parameters. To optimize the training efficiency, we employed Low-Rank Adaptation (LoRA), which allowed us to significantly reduce the number of trainable parameters to 2,359,296, representing only 2.11% of the total parameters.

The training was conducted over 4 epochs. For Siamese networks, each epoch required approximately 2 hours, whereas for Triplet networks, each epoch took around 1 hour. This difference in training time is attributed to the distinct architectural and computational requirements of Siamese and triplet networks.

The computational resources used for training included NVIDIA A100 80GB PCIe GPUs. The coding was done in *Python* using *PyTorch* and *PEFT* libraries.

A.4 Results for other Text-Embedding Models

Figure 5 shows the poor performance of *NV-Embed-v1*, the current best LLM-based (29GB) text embedder[8], in differentiating between opposing vs. supporting statements in terms of spatial distance.

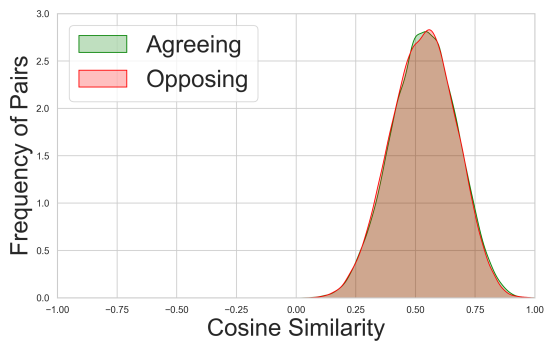


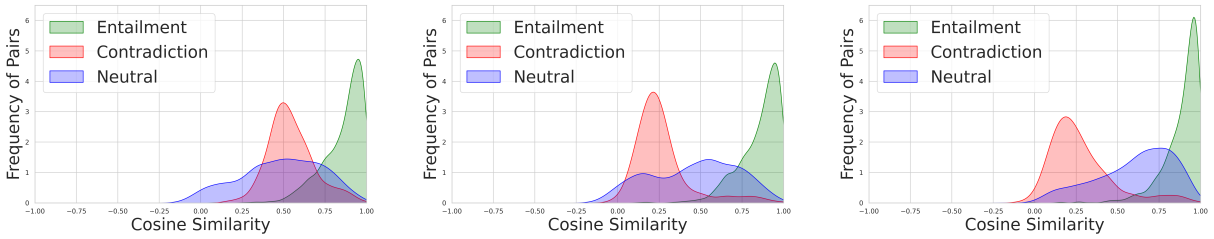
Figure 5: Performance of *NV-Embed-v1* on Kialo Test-Set.

A.5 Out of Distribution Validation

Figure 6 depicts the distributions of cosine similarities provided by the original and the fine-tuned models (LoRA and fully fine-tuned) for the three categories of pairwise relationships in the dataset: *Neutral*, *Entailment*, and *Contradiction*. Ideally, in the fine-tuned model, we would desire to witness: 1) a further shift for the contradictory pairs' distribution (red curve) to the left side, 2) while the distribution of the entailing pairs (green curve) peaking near the right side, and 3) *Neutral* pairs (blue curve) maintaining a relatively more uniform distribution across the x-axis as it includes both topically relevant (majority) and irrelevant (minority) pairs of statements. Moreover, we expect the peak of the *Neutral* pairs' curve to stand in between the former two so that when it comes to sentence pair similarity, our fine-tuned model preserves the ascending order of: 1) topically relevant but contradictory, 2) topically relevant but neutral, and 3) topically relevant and entailing.

Across Figures 6a, 6b, and 6c, we observe a progression in stance detection abilities. Initially, the original *all-mpnet-base-v2* model can also distinguish *Entailment* from *Contradiction* (Fig. 6a), suggesting that the contradictions in this dataset are less subtle than in the Kialo test set (Figure 3a). Yet, our LoRA fine-tuned model significantly improves differentiation, correctly shifting *Contradiction* pairs leftwards, and maintaining an appropriate balance between *Neutral* and *Entailment* pairs — desirably forcing the topically relevant *Neutrals* peak to stand between the peaks of *Contradiction* and *Entailment* curves (Fig. 6b). However, full fine-tuning (Fig. 6c) manifests its catastrophic forgetting — while the gap between the distributions of *Contradiction* and *Entailment* is also enhanced when compared to the original model, *Neutral* pairs are undesirably shifted towards the *Entailment*.

[8] https://huggingface.co/spaces/mteb/leaderboard



(a) Original KL-Div. for green vs. red = 3.5.

(b) LoRA fine-tuned. KL-Div. = 5.7.

(c) Fully fine-tuned KL-Div. = 5.4

Figure 6: Distributions of cosine similarities of pairs in *SemEval 2014* dataset.

Query	**Original**	**Hybrid, margin = 0.4**	**Siamese, margin = 0.4**	**Triplet, margin = 0.4**
"Abortion is healthcare."	76%	91%	84%	94%
"Abortion is murder."	67%	80%	64%	79%

Table 8: Alignment Precision for semantic search on congresspeople tweets with abortion-related queries.

This highlights the advantage of the LoRA fine-tuned model in achieving both stance-awareness and preserving prior knowledge, underscoring its value in fine-tuning for stance-aware sentence embeddings.

Observation: Our fine-tuned models exhibit an increase in stance awareness compared to the original model, which possessed some limited understanding of stances in a different dataset, i.e.: *SemEval-2014*, contributing to R2.

A.6 Semantic Search with other Models

Table 8 extend the results of Table 4 to other best-performing models from different architectures (Siamese and Triplet).