TRoTR: A Framework for Evaluating the Recontextualization of Text

Francesco Periti
University of Milan
Via Celoria 18, 20133, Milan, Italy
francesco.periti@unimi.it

Pierluigi Cassotti
University of Gothenburg
Renströmsgatan 6, 40530 Gothenburg, Sweden
pierluigi.cassotti@gu.se

Stefano Montanelli
University of Milan
Via Celoria 18, 20133, Milan, Italy
stefano.montanelli@unimi.it

Nina Tahmasebi
University of Gothenburg
Renströmsgatan 6, 40530 Gothenburg, Sweden
nina.tahmasebi@gu.se

Dominik Schlechtweg
University of Stuttgart
Pfaffenwaldring 5 b, 70569 Stuttgart, Germany
dominik.schlechtweg@ims.uni-stuttgart.de

Abstract

Current approaches for *detecting* text reuse do not focus on *recontextualization*, i.e., how the new context(s) of a reused text differs from its original context(s). In this paper, we propose a novel framework called TRoTR that relies on the notion of topic relatedness for evaluating the diachronic change of context in which text is reused. TRoTR includes two NLP tasks: TRiC and TRaC. TRiC is designed to evaluate the topic relatedness between a pair of recontextualizations. TRaC is designed to evaluate the overall topic variation within a set of recontextualizations. We also provide a curated TRoTR benchmark of biblical text reuse, human-annotated with topic relatedness. The benchmark exhibits an inter-annotator agreement of .811. We evaluate multiple, established SBERT models on the TRoTR tasks and find that they exhibit greater sensitivity to textual similarity than topic relatedness. Our experiments show that fine-tuning these models can mitigate such a kind of sensitivity.

1 Introduction

As individuals, we often *reuse* someone else's words for diverse reasons and in various ways. This linguistic choice transcends cultural and temporal boundaries, representing an interesting phenomenon to study in Linguistics (Bois, 2014). For instance, linguistic scholars have investigated theories of Reception (Thompson, 1993; Hohendahl and Silberman, 1977) and Resonance (McDonnell et al., 2017; Dimock, 1997) to understand how individuals and communities interpret and reuse historical texts many years after they were written.

With the advent of digitization, recent years have seen a growing interest in computational methods for studying *text reuse*, i.e., "the reuse of existing written sources in the creation of a new text" (Clough et al., 2002). Existing methods focus on the main task of Text Reuse Detection (TRD). In TRD, text reuses are all assumed as "*topically related* to the source" (Hagen and Stein, 2011; Chiu et al., 2010), the boundaries of reused text are unknown, and the goal is to *detect* text reuse across a diachronic corpus (Seo and Croft, 2008). Whether and how the topic(s) or context(s) of a reused text differs from the source is generally overlooked. Thus, new methods are needed for modeling *recontextualization*, i.e., "the dynamic transfer-and-transformation of a text from one discourse/text-in-context to another" (Connolly, 2014; Linell, 1998).

In this paper, we propose a framework, called Topic Relatedness of Text Reuse (TRoTR), to evaluate computational methods for capturing the different recontextualizations of text reuse. In TRoTR, the boundaries of reused text are known and the goal is to distinguish reuses of the same text according to their different, latent (i.e., unlabeled) topics. As an example, consider three recontextualizations of the biblical passage *John 15:13* (in bold):

(1) It's the wonderful pride month!! ♥ ♥ ♥ ♥ ♥ ♥ Honestly pride is everyday! Love is love don't forget I love you ♥. Remember this! John 15:12-13: "My command is this: Love each other as I have loved you. **Greater love has no one than this: to lay down one's life for one's friends**"

(2) At a large Crimean event today Putin quoted the Bible to defend the special military operation in Ukraine which has killed thousands and displaced millions. His words "**There is no greater love than if someone gives soul for their friends**". And people were cheering him. Madness!!!

(3) "Freeing people from genocide is the reason, motive & goal of the military operation we started in the Donbas & Ukraine", Putin says, then quotes the Bible: "**There is no greater love than to lay down one's life for one's friends.**" It's like Billy Graham meets North Korea

In this example, the biblical passage is incorporated within three texts with different topic recontextualizations. In particular, the text (1) has a different topic with respect to text (2) and (3), while the texts (2) and (3) are topic related. In TRoTR, we

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 13972–13990

support the recognition of such a kind of recontextualizations by leveraging the notion of topic relatedness. TRoTR represents a new opportunity in Natural Language Processing (NLP) and can be used to distinguish recontextualizations of any kind of text reuse (e.g., proverbs, Ghosh and Srivastava, 2022), to investigate phenomena such as the use of misquotations (Porrino et al., 2008) and dogwhistles (Hertzberg et al., 2022), as well as to provide in-context interpretation to vague utterances, with special focus on enhancing the Large Language Models (LLMs)' capabilities to this end (DeVault and Stone, 2004).

Our original contribution.

- **We introduce a novel framework**, called TRoTR, with two NLP tasks called Text Reuse in-Context (TRiC) and Topic variation Ranking across Corpus (TRaC).
- **We provide TRoTR with a benchmark** containing gold labels derived by human judgements of topic relatedness. The judgements show an inter-annotator agreement of .811, calculated by the average pairwise correlation on assigned assessments.
- **We propose a novel annotation process** to model topics through topical relatedness in context pairs.
- **We evaluate 36 SBERT models** by considering 4 settings. Our results reveal that these models reach high performance (correlation of .600 – .800), but are more sensitive to textual similarity rather than topic relatedness.

2 Related work

Works related to TRoTR are about text reuse and recontextualization, semantic textual similarity and relatedness, and topic modeling and annotation.

Text reuse and recontextualization. Although multiple facets of text reuse have been investigated, such as historical (Büchler et al., 2014), cross-lingual (Muneer and Nawab, 2022), allusive (Manjavacas et al., 2019), explicit (Franzini et al., 2018), non-literal (Moritz et al., 2016), and local (Seo and Croft, 2008), computational approaches primarily focuses on *detecting* instances of text reuse. To the best of our knowledge, studies extending beyond mere TRD often leverage text metadata to analyze reuse within temporal and spatial graphs (Khritankov et al., 2015; Smith et al., 2013; Xu et al., 2014). However, these studies do not specifically focus on capturing how the reused text is recontextualized, thereby leaving a gap in the current literature.

Among recent advancements in NLP, some works are related to the recontextualization of text. Wilner et al. (2021) focus on Narrative Analysis by investigating how the recontextualization of events across whole stories impacts word embeddings. Ghosh and Srivastava (2022) introduce a benchmark for evaluating the LLMs' capability of generating proverbs in-context of narratives.

Over the past few years, there has been growing interest in quotations, i.e. "well known phrases or sentences that we use for various purposes such as emphasis, elaboration, and humor" (Lee et al., 2016). This interest extends to various forms of quotations spanning from epigraphs (Bond and Matthews, 2018) to biblical references (Moritz et al., 2016). In particular, there has been a surge of attention in recommendation systems that offers off-the-shelf quotations based on provided context (Wang et al., 2023, 2022, 2021).

Semantic textual similarity and relatedness. In NLP, a possible option for assessing text recontextualization is to use *semantic* (textual) *similarity*. However, semantic similarity is traditionally used as a metric to assess paraphrases or entailment equivalence between two texts (Hercig and Kral, 2021; Konopík et al., 2017; Cer et al., 2017; Agirre et al., 2016, 2015, 2014, 2013, 2012); thus, it is not suitable for TRoTR. *Semantic* (textual) *relatedness* has been long recognized as a core aspect in understanding the meaning of texts (Miller and Charles, 1991; Halliday and Hasan, 2014), and encompasses a multitude of intricate relationships, such as sharing a common *topic*, expressing similar viewpoints, or originating from the same temporal period (Abdalla et al., 2023). However, there is no universally accepted linguistic theory or set of guidelines for evaluating relatedness. Its assessment is inherently more complex than semantic similarity, as two texts may lack semantic similarity but still be semantically related through some textual relationship.

Topic modeling and annotation. An alternative method for assessing text recontextualization is by analyzing topics where text is reused (Jin and Spence, 2021; Kim et al., 2018). Topic models can be useful tools to discover latent topics in collections of documents (Abdelrazek et al., 2023),

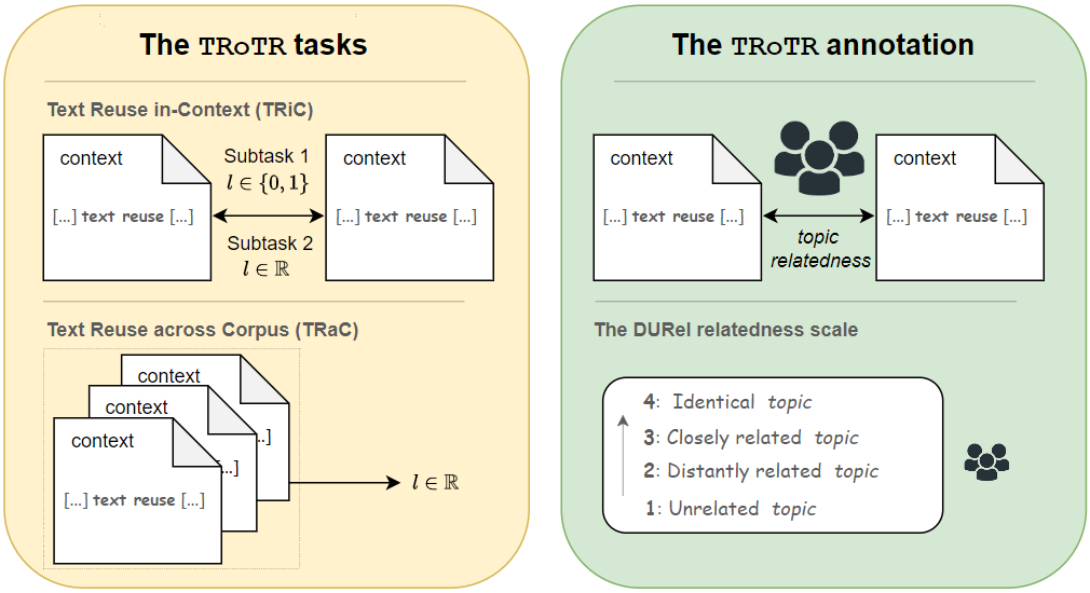

Figure 1: The TRoTR framework consists of two tasks, called Text Reuse in-Context (TRiC) and Topic variation Ranking across Corpus (TRaC), along with a corresponding annotation process. We use [. . .] to denote the left and right context of a target text-reuse excerpt.

either as probability distributions like LDA (Blei et al., 2003) or clustering of embeddings like BERTopic (Grootendorst, 2022). When applied, the derived topics need to be carefully evaluated against benchmarks containing manually derived ground truth. As topics represent vague concepts, different guidelines for deriving ground truth use different topic definitions tailored to the specific interests of analysis (Orita et al., 2014). Generally, these guidelines result in manual annotations of topic labels that typically differ across annotators and thus require post-processing techniques to be uniform and standardized (Poursabzi-Sangdeh and Boyd-Graber, 2015). For example, annotators can use different wording to express the same concept.

As a result, there is no well-established guideline for annotating topics. However, common to different guidelines is a definition of topic that relies on the notion *what the text is about* (Bauwelinck and Lefever, 2020; Hovy and Lin, 1998).

3 The TRoTR framework

The TRoTR framework consists of two tasks, called Text Reuse in-Context (TRiC) and Topic variation Ranking across Corpus (TRaC), along with a corresponding annotation process (see Figure 1). TRiC and TRaC are grounded on human judgments of a specific facet of semantic relatedness that considers the extent to which two texts share a common *topic*. We call this facet **topic relatedness** (see Table 1 for an example). In our study, the definition of topic follows the popular notion of *what the text is about*.

When dealing with complex problems, such as recontextualization, a general approach involves starting with a smaller sub-problem to establish a focused foundation before further expanding. Thus, we first present TRiC as a *context-pair level* task. Then, we present TRaC as a more complex *corpus-level* task that must be addressed to identify potential varying targets for real, in-depth analysis.

3.1 Tasks

In the TRoTR tasks, instances of text reuse are presented within different contexts, each representing a new recontextualization of the original text.

Text Reuse in-Context frames a text reuse t within two different contexts c_1 and c_2. The goal is to assess the topic relatedness of c_1 and c_2. TRiC includes two subtasks, namely *binary classification* and *ranking*. These subtasks resemble the structure of the Word-in-Context task (Pilehvar and Camacho-Collados, 2019) and the Graded Word Similarity in Context task (Armendariz et al., 2020), respectively. However, while they focus on distinguishing the different meanings words can have in different contexts, TRiC focuses on distinguishing different topics in which texts are reused.

Each TRiC instance is associated with a binary label $l \in \{0, 1\}$ and a continuous score $1 \leq s \leq 4$.

Text 1	Text 2	Semantic Textual Similarity	Semantic Textual Relatedness	Semantic Textual *Topic* Relatedness
It's the wonderful pride month!! ♥ ♥ ♥ ♥ ♥ ♥ Honestly pride is everyday! Love is love don't forget I love you ♥. Remember this! John 15:12-13: "My command is this: Love each other as I have loved you. **Greater love has no one than this: to lay down one's life for one's friends**"	Happy Pride Month! ♥ Remember, pride isn't just for a month—it's a daily celebration! Love knows no boundaries, and I want you to know that I cherish you every single day. ♥ Let's always remember these powerful words from John 15:12-13: "My command is this: Love each other as I have loved you. **Greater love has no one than this: to lay down one's life for one's friends**"	✓ paraphrase	✓ related in some aspects	✓ related in topic
"Freeing people from genocide is the reason, motive & goal of the military operation we started in the Donbas & Ukraine", Putin says, then quotes the Bible: "**There is no greater love than to lay down one's life for one's friends.**" It's like Billy Graham meets North Korea	At a large Crimean event today Putin quoted the Bible to defend the special military operation in Ukraine which has killed thousands and displaced millions. His words "**There is no greater love than if someone gives soul for their friends**". And people were cheering him. Madness!!!	× neither paraphrases nor entailment	✓ related in some aspects	✓ related in topic
It's the wonderful pride month!! ♥ ♥ ♥ ♥ ♥ ♥ Honestly pride is everyday! Love is love don't forget I love you ♥. Remember this! John 15:12-13: "My command is this: Love each other as I have loved you. **Greater love has no one than this: to lay down one's life for one's friends**"	At a large Crimean event today Putin quoted the Bible to defend the special military operation in Ukraine which has killed thousands and displaced millions. His words "**There is no greater love than if someone gives soul for their friends**". And people were cheering him. Madness!!!	× neither paraphrases nor entailment	✓ related in some aspects	× unrelated in topic
You are altogether beautiful, my darling; there is no flaw in you. Charm is deceitful, and beauty is vain, but a woman who fears the Lord is to be praised	At a large Crimean event today Putin quoted the Bible to defend the special military operation in Ukraine which has killed thousands and displaced millions. His words "**There is no greater love than if someone gives soul for their friends**". And people were cheering him. Madness!!!	× neither paraphrases nor entailment	× unrelated in any aspects	× unrelated in topic

Table 1: Examples of *semantic textual similarity*, *semantic textual relatedness*, and *topic relatedness*. The first and last pair of sentences are examples of paraphrases and semantically unrelated content, respectively. Most people will agree that the second pair of sentences is more related in topic than the third pair of sentences. However, some people may still consider the third pair as semantically related due to the presence of the same quotation.

- Subtask 1 - *binary classification*: the task is to identify, for each instance, whether the contexts c_1 and c_2 share roughly the same topic (i.e., $l = 1$) or not (i.e., $l = 0$).
- Subtask 2 - *ranking*: the task is to rank the TRiC instances according to the degree of topic relatedness s of the contexts c_1 and c_2.

Topic variation Ranking across Corpus frames a text reuse t within a corpus C that includes various contexts c_i where t occurs. TRaC resembles the structure of the Lexical Semantic Change (LSC) detection task defined by Schlechtweg et al. (2018); Kutuzov and Pivovarova (2021). However, while this focuses on assessing the semantic change of a word, TRaC focuses on assessing the *topic variation* of a reused text.

Each TRaC instance is associated with a continuous score $s \in [0, 1]$ of topic variation that indicates the variability in topic usages for a target text reuse t across the corpus C. Specifically, a score of 1 indicates that a target is associated with a high number of topics, while a score of 0 indicates that a target is associated with a single topic.

Given a set of target text reuses $t \in T$, the task is to rank the text reuses by the degree of topic variation across the corpus C.

3.2 Annotation process

The TRoTR annotation process is enforced to collect human judgments of topic relatedness. In our study, we sidestep the need for annotating topics explicitly using a well-established paradigm adopted for modeling word meaning. Our intuition is that annotating topic relatedness, instead of relying on explicit topic labels, closely mirrors recent work

exemplified in the Word-in-Context task (Pilehvar and Camacho-Collados, 2019), which relies on annotating meaning relatedness rather than explicit sense labels.

Annotators are asked to evaluate the *topic relatedness* of different text reuse instances $\langle t, c_1, c_2 \rangle$, where t is a target text reuse, and c_1 and c_2 are two different contexts in which t occurs.[1]

The topic relatedness is evaluated by utilizing the four-point DURel relatedness scale (Schlechtweg et al., 2024), with annotators following instructions inspired by the guidelines from Erk et al. (2013), as well as those provided for SemEval-2020 Task 1 (Schlechtweg et al., 2020) and the PLATOS project (Bauwelinck and Lefever, 2020). See Appendix G for our guidelines. [2]

4 The TRoTR benchmark

The TRoTR benchmark is composed of human-annotated instances of text reuse. Specifically, we first manually collected and curated tweets from Twitter (now X) containing text reuse instances. We then incorporated gold labels derived by human annotations.

4.1 Data

Inspired by Moritz et al. (2016); Büchler et al. (2014), we focus on text reuse of biblical passages because they typically show high context variety (Greenough, 2021; Cheong, 2014), the degree of which we aim to study. Moreover, they are frequently and explicitly mentioned *in-context*, often with an identifying reference (e.g., *John 15:13*). Tweets were collected through a manual search process, thus allowing us to avoid a TRD phase and its validation.

For a set of 42 target passages we collected 30 tweets each. These were curated by experts by removing minor word variations in phrasing that can stem from the use of e.g., different Bible versions.

4.2 Human judgments

We collected judgments according to the procedure outlined in Section 3.2. Specifically, we recruited four native English speakers as annotators. Annotators were trained and tested on a small set of instances in an online tutorial.

For each target passage t, we generate all possible context pairs where the contexts are chosen from the 30 tweets. We then randomly sampled 150 context pairs. These were presented to annotators in randomized order to be judged for topic relatedness. Each context pair received at least two judgements, although the majority received three.

The outcome of our annotation pipeline is a dataset of 6,300 annotated context pairs. We measured inter-annotator agreement on judgments using Krippendorff's α coefficient (Krippendorff, 2018) and the weighted mean of Spearman correlations (Spearman, 1904) between annotator pairs. Table 5 in Appendix provides a summary of our agreement scores. Similar to previous studies that reported Krippendorff's α of .439 (Loureiro et al., 2022) and weighted mean of Spearman correlation between annotator judgments ranging from .550 to .680 (Erk et al., 2013; Schlechtweg et al., 2018), we obtained a comparable Krippendorff's α score of .420 and Spearman correlation of .506.

4.3 Deriving gold labels

Following Loureiro et al. (2022), we employ filtering criteria for the annotation instances to reduce uncertainty and ensure a more controlled setting.

For TRiC, we first filtered out all instances with high disagreement[3], e.g. an instance with three different judgments where it is unclear which the gold label could be. We also enforce a clear-cut separation by filtering out all the instances where the average judgment score is between 2 and 3. This filtering results in a more refined dataset of 3,821 annotated context pairs, characterized by a Krippendorff's α agreement of .709 and a weighted average pairwise Spearman agreement of .811.

For TRaC, we adopted a different filtering approach at the level of targets to ensure a comparable number of instance pairs when deriving the gold labels. Specifically, we filtered out the targets t where the weighted average pairwise Spearman agreement is below .150 leading to the exclusion of 2 targets.

TRiC labels. For each instance, we aggregate the judgments of all annotators by averaging. We then directly use the average judgment s of each instance to derive binary labels and continuous scores for Subtask 1 and Subtask 2.

[1] The data was annotated in the PhiTag annotation system: https://phitag.ims.uni-stuttgart.de/.

[2] The annotation guidelines for TRoTR, along with its benchmark, and our code, are available at https://github.com/FrancescoPeriti/TRoTR.

[3] We consider high disagreement to be a difference between the maximum and the minimum judgment of 2 or 3.

For Subtask 1, we binarize s as 1 if $s \geq 2.5$ or as 0 if $s < 2.5$ and associate each instance with the corresponding binary label. A threshold of 2.5 is a midpoint split on the judgment scale. It follows that the 0 label consists of Unrelated and Distantly related annotations, while label 1 consists of Identical and Closely related annotations. Overall, our benchmark includes a total of 2,621 examples with label 0 and a total of 1,200 examples with label 1.

For Subtask 2, we directly utilize the continuous score s for each instance.

TRaC labels. For each target, we use a judgment summary measure similar to the DURel EARLIER/LATER measures introduced by Schlechtweg et al. (2018) in the field of LSC (Periti and Montanelli, 2024; Tahmasebi et al., 2021). This involves computing the average of annotator judgments over all instances for a target. Lower scores correspond to greater topic variation, while greater scores (i.e., more Identical annotations) are associated with less topic variation.

5 Evaluation setup

We use the TRoTR tasks and benchmarks to evaluate the ability of sequence-level models to capture topic relatedness and variation in different text recontextualizations to set baselines for the tasks.

Because Sentence-BERT (SBERT) models are recognized to be the state-of-the-art architecture for addressing sequence-level tasks (Reimers and Gurevych, 2019), we choose a range of different SBERT models tailored for sequence-level embeddings and textual similarity.

5.1 SBERT models

We consider 36 SBERT models trained on a wide range of tasks including Paraphrasis, Semantic Similarity, and Question Answering. Specifically, we evaluate all the (non-image based) pre-trained models available at `https://www.sbert.net/index.html`. We evaluate each SBERT model in its pre-trained version (base) and three different settings, namely:

- *+MASK*: given an instance $\langle t, c_1, c_2 \rangle$, we mask the text-reuse excerpt t in the contexts c_1 and c_2 to prevent that the topic estimate of topic relatedness is influenced by the common t in c_1 and c_2. To this end, we replace t in c_1 and c_2 with a dash (i.e., "-");
- *+FT*: we fine-tune the pre-trained model on TRiC instances using the *contrastive loss* (Hadsell et al., 2006). This loss minimizes the distance between embeddings of similar sentences and maximizes the distance for dissimilar sentences;
- *+FT+MASK*: we combine both the +FT and +MASK settings, meaning that we fine-tune the model and then evaluate it by considering contexts where targets are masked.

SBERT architectures. Each SBERT model has been pre-trained using one of two architectures:

- *Bi-Encoder* models are designed to produce a sequence embedding for an input text sequence. Given an instance $\langle t, c_1, c_2 \rangle$, we independently feed a Bi-Encoder model with the sequence c_1 and c_2 to obtain the corresponding sequence embeddings u and v. Similar to Abdalla et al. (2023), we use the cosine similarity between u and v as an estimate of the topic relatedness between c_1 and c_2.
- *Cross-Encoder* models are designed to produce an output value that indicates the similarity of two input sequences. Thus, given an instance $\langle t, c_1, c_2 \rangle$, we simultaneously pass the sequences c_1 and c_2 to the Cross-Encoder model and use the output value as an estimate of the topic relatedness between c_1 and c_2.

5.2 TRiC evaluation

Similar to the WiC tasks (e.g., Pilehvar and Camacho-Collados, 2019), we split the TRoTR benchmark into three distinct partitions, namely training set (Train), development set (Dev), and test set (Test), comprising approximately 80%, 10%, and 10% of the instances, respectively. To strengthen the robustness of the evaluation, ten randomized Train-Dev-Test splits were generated (see Appendix A). The average performance across all the splits is used as reference for comparison.

Additionally, inspired by Raganato et al. (2020), we include the evaluation of target text reuse t that are unseen during fine-tuning. The goal is to evaluate the ability of models to generalize the assessment of topic relatedness. Specifically, we fine-tune each considered model on the Train set and we evaluate it on two different Test sets: i) the standard Test set, containing instances $\langle t, c_1, c_2 \rangle$ whose target t was either seen or unseen during fine-tuning; and ii) the **Out-of-Vocabulary** (OOV)

Test set, containing only instances $\langle t, c_1, c_2 \rangle$ whose target t was not seen during fine-tuning. OOV Test set represents half of the Standard Test set.

For TRiC Subtask 1, we need to define a threshold to determine instances $\langle t, c_1, c_2 \rangle$ where c_1 and c_2 share roughly the same topic or not. Thus, given a model, we tune a threshold-based classifier on the Dev set. Specifically, for each instance $\langle t, c_1, c_2 \rangle$ in Dev, we use the model to predict the topic relatedness between c_1 and c_2. Then, we determine the optimal threshold that maximized the Weighted F1 (Harbecke et al., 2022) score over the Dev set. Finally, we apply this threshold to both the Train and Test sets. Due to the unbalanced distribution of gold binary labels, we evaluate models using the F1 metric. Precision (PR) and Recall (RE) for each individual class are also reported for completeness.

For TRiC Subtask 2, given a model, we directly use its predictions as estimates of topic relatedness. Then, we evaluate the model using Spearman correlation (SP) with continuous gold scores.

5.3 TRaC evaluation

Similar to the LSC tasks (e.g., Schlechtweg et al., 2020), we consider an *unsupervised* scenario. In particular, motivated by the limited number of targets (i.e., 42), we do not split the benchmark into Train-Dev-Test partitions with the aim to mitigate the potential evaluation impact of a small Test set. Without training instances, the configurations with +FT and +FT+MASK are not applicable to TRaC.

To quantify the topic variation of a target, we adopted the same approach used for determining the gold scores. Thus, given a model, the topic variation of a target t is calculated as the average prediction of topic relatedness across all the annotated $\langle t, c_1, c_2 \rangle$ pairs. We then evaluate models using Spearman correlation (SP) with gold scores.

6 Evaluation results

First, we evaluated an extensive set of pre-trained SBERT models on the TRiC task (see Table 6 in Appendix). Then, for simplicity, we opted to consider and fine-tune a smaller set of models, precisely the top-five models by SP over the Train sets. Since we did not perform any training over the models, the Train sets act as a larger sets for testing the models. Specifically, we chose: all-distilroberta-v1 **(ADR)**, distiluse-base-multilingual-cased-v1 **(DBM)**, paraphrase-multilingual-MiniLM-L12-v2 **(PAM)**, paraphrase-multilingual-mpnet-base-v2 **(PAR)**, and multi-qa-mpnet-base-cos-v1 **(MQA)**. In particular, ADR and DBM are Bi-Encoders for English. PAM and PAR are multilingual Bi-Encoders fine-tuned on paraphrase pairs. Similarly, MQA is a multilingual Bi-Encoder fine-tuned on question-answer pairs.

As a general remark on our initial evaluation, we note that Bi-Encoder models consistently exhibit superior performance compared to Cross-Encoder models in both TRiC Subtask 1 and Subtask 2. This finding aligns with the recent comparisons by Ishihara and Shirai (2022) and Cassotti et al. (2023) for News Article Similarity and LSC, challenging the idea that the use of cross-attention benefits Cross-Encoder architectures in sequence-level tasks (Lee et al., 2023; Thakur et al., 2021). In the following, we first present the results of our evaluation by comparing the use of pre-trained and fine-tuned models (+FT); then, we discuss the results in the masking settings (+MASK, +FT+MASK). We report in Table 2 and 3 the overall results for TRiC and TRaC, respectively.

6.1 TRiC: pre-trained vs. fine-tuned

Across the overall *standard* Test sets, when *pre-trained* models are used for Subtask 1, we observe high precision (PR) values, ranging from .93 to .96, and low recall (RE) values ranging from .21 to .47 for label 0 (i.e., different topics). Conversely, for label 1 (i.e., roughly identical topics), we observe an inverse trend of performance, with PR values ranging from .31 to .42 and RE values ranging from .93 to .97. Such results suggest that SBERT models face difficulties in distinguishing different recontextualization. For Subtask 1, we observe a moderate F1-score (F1) ranging from .43 to .61; for Subtask 2, we observe only moderate Spearman correlation coefficients (SP) ranging from .54 to .58.

Additional results for the *OOV* Test sets are reported in Table 2. We note that the results for the OOV Test sets are lower in performance while being associated to higher standard deviations. For pre-trained models, we attributed this drop to (1) the unbalance number of instances and labels available for each target; (2) that the inter-annotator agreements differ between targets. If target words with small number of instances or lower inter-annotator agreement fall in the OOV Test sets, then the performance will be much lower. Finally, (3) the size of the OOV Test sets is smaller because it splits the standard Test sets in two halves.

	Standard Test set								Out-of-vocabulary (OOV) Test set							
	Label 0			*Label 1*			***All***		*Label 0*			*Label 1*			***All***	
Models	PR	RE	F1	PR	RE	F1	F1	SP	PR	RE	F1	PR	RE	F1	F1	SP
ADR	**.95±.03**	.47±.13	.62±.11	.42±.11	.93±.04	.57±.10	.61±.10	.55±.09	**.94±.07**	.45±.20	.58±.20	.38±.19	**.93±.06**	.51±.18	.58±.16	.48±.20
+FT	**.95±.03**	**.61±.15**	**.73±.11**	**.50±.14**	**.93±.03**	**.64±.10**	**.71±.10**	**.66±.07**	.91±.12	**.49±.24**	**.61±.22**	**.40±.21**	.91±.06	**.52±.18**	**.61±.18**	**.51±.22**
+MASK	*.89±.05*	*.87±.07*	*.87±.03*	*.70±.14*	*.72±.12*	*.69±.07*	*.82±.03*	*.67±.06*	*.90±.07*	*.85±.10*	*.87±.05*	*.62±.21*	*.71±.18*	*.63±.14*	*.82±.05*	*.62±.15*
+FT+MASK	*.90±.07*	*.89±.07*	*.89±.03*	*.75±.12*	*.76±.12*	*.74±.05*	*.85±.04*	*.71±.05*	*.87±.11*	*.88±.09*	*.87±.06*	*.66±.20*	*.70±.15*	*.65±.09*	*.82±.06*	*.63±.15*
DBM	.96±.02	.26±.12	.40±.14	.35±.09	**.97±.03**	.51±.09	.43±.12	.54±.09	**.96±.08**	.21±.19	.31±.23	.31±.14	**.97±.05**	.45±.16	.38±.18	.44±.23
+FT	**.97±.02**	**.46±.17**	**.60±.15**	**.43±.10**	.96±.03	**.58±.09**	**.61±.13**	**.64±.07**	.93±.15	**.34±.23**	**.46±.26**	**.34±.14**	.95±.05	**.49±.15**	**.50±.19**	**.48±.29**
+MASK	*.87±.07*	*.88±.07*	*.87±.03*	*.72±.14*	*.66±.16*	*.66±.09*	*.81±.03*	*.64±.04*	*.88±.09*	*.88±.09*	*.87±.05*	*.66±.23*	*.64±.25*	*.58±.19*	*.82±.04*	*.58±.12*
+FT+MASK	*.88±.06*	*.89±.07*	*.88±.04*	*.74±.11*	*.70±.13*	*.70±.04*	*.83±.03*	*.66±.04*	*.85±.12*	*.87±.09*	*.85±.08*	*.63±.19*	*.58±.20*	*.57±.13*	*.80±.08*	*.58±.14*
PAM	**.96±.02**	.46±.09	.61±.08	.41±.09	**.96±.02**	.57±.08	.61±.07	.58±.08	**.96±.04**	.43±.17	**.57±.16**	**.37±.15**	**.95±.05**	**.52±.15**	**.59±.12**	.49±.22
+FT	.95±.03	**.59±.12**	**.72±.11**	**.48±.09**	.92±.04	**.63±.08**	**.70±.09**	**.66±.06**	.90±.18	**.45±.21**	**.57±.23**	**.37±.13**	.92±.06	.51±.13	**.59±.17**	**.51±.22**
+MASK	*.89±.05*	*.88±.06*	*.88±.03*	*.71±.10*	*.72±.10*	*.70±.05*	*.83±.03*	*.67±.04*	*.89±.09*	*.86±.09*	*.87±.06*	*.65±.19*	*.71±.18*	*.65±.12*	*.83±.05*	*.60±.13*
+FT+MASK	*.90±.05*	*.90±.03*	*.90±.03*	*.76±.07*	*.77±.06*	*.76±.03*	*.86±.03*	*.69±.04*	*.88±.10*	*.89±.05*	*.88±.06*	*.68±.13*	*.73±.11*	*.69±.07*	*.84±.06*	*.60±.12*
PAR	**.95±.03**	.40±.10	.56±.09	.39±.09	**.95±.04**	.55±.08	.56±.07	.56±.09	**.93±.11**	.35±.18	.49±.19	.34±.15	**.95±.06**	.49±.16	.52±.15	.47±.25
+FT	**.95±.05**	**.60±.10**	**.73±.08**	**.49±.10**	.93±.05	**.63±.08**	**.71±.07**	**.66±.06**	.91±.17	**.46±.21**	**.58±.21**	**.38±.16**	.91±.08	**.51±.15**	**.59±.18**	**.53±.24**
+MASK	*.89±.05*	*.85±.07*	*.87±.04*	*.69±.10*	*.75±.11*	*.70±.05*	*.83±.03*	*.68±.03*	*.90±.08*	*.83±.13*	*.86±.07*	*.63±.19*	*.75±.17*	*.65±.10*	*.82±.05*	*.62±.11*
+FT+MASK	*.89±.06*	*.91±.05*	*.90±.03*	*.78±.09*	*.73±.11*	*.74±.05*	*.86±.03*	*.70±.04*	*.87±.11*	*.90±.07*	*.88±.06*	*.68±.16*	*.66±.18*	*.64±.11*	*.83±.07*	*.61±.14*
MQA	.94±.03	.42±.11	.58±.11	.40±.10	**.94±.03**	.55±.09	.58±.09	.55±.09	**.94±.09**	.39±.19	.53±.20	.36±.19	**.96±.03**	.50±.18	.55±.16	.49±.21
+FT	**.96±.03**	**.61±.13**	**.74±.10**	**.50±.10**	**.94±.04**	**.65±.08**	**.72±.09**	**.68±.06**	.92±.15	**.47±.22**	**.60±.24**	**.39±.16**	.94±.05	**.53±.15**	**.61±.19**	**.54±.21**
+MASK	*.88±.05*	*.87±.07*	*.88±.04*	*.71±.10*	*.71±.12*	*.69±.06*	*.83±.04*	*.68±.05*	*.89±.07*	*.86±.10*	*.87±.06*	*.63±.18*	*.69±.16*	*.63±.13*	*.83±.05*	*.62±.13*
+FT+MASK	*.90±.05*	*.91±.04*	*.90±.03*	*.77±.08*	*.76±.09*	*.76±.05*	*.86±.03*	*.72±.04*	*.88±.10*	*.90±.04*	*.88±.06*	*.67±.16*	*.69±.16*	*.65±.11*	*.84±.06*	*.63±.13*

Table 2: **TRiC evaluation** on Subtask 1 and Subtask 2 for both Test and OOV Test sets. For Subtask 1, precision (PR), recall (RE), and Weighted -F1 scores (F1) are reported for both label 0 (i.e., different topics) and label 1 (i.e., roughly identical topics). For Subtask 2, Spearman correlation (SP) is reported on the overall set of instances. Standard deviations (±) across the 10 Test splits are presented for comparative analysis. For each metric, the best performance of the comparison between pre-trained/fine-tuned models is highlighted in **bold**. Results for masking settings are reported in *italic*.

Models	ADR *+MASK*	DBM *+MASK*	PAM *+MASK*	PAR *+MASK*	MQA *+MASK*
Spearman	.72 *.84*	.66 *.80*	.66 *.81*	.73 *.76*	.65 *.80*

Table 3: **TRaC evaluation** using the pre-trained models alone and in the +MASK setting (*italic*).

Fine-tuning: When the pre-trained models are *fine-tuned* on TRiC instances (i.e., +FT), we observe a significant improvement in performance for both Subtask 1 and Subtask 2 on both the standard Test set and the OOV Test set. This observation indicates that fine-tuning SBERT models on TRiC instances enhances their capability to contextualize a sequence *in-context*. In particular, the improvement is more pronounced on the standard Test sets than on the OOV Test sets. We attribute this discrepancy to the limited size of our benchmark that includes a small number of target quotations sufficient for testing purposes. A larger number of targets will further improve the models' generalization capability. For Subtask 1, we observe a F1 ranging from **.61** to **.72** (standard) and from **.50** to **.61** (OOV); for Subtask 2, we observe SP coefficients ranging from **.64** to **.68** (standard) and **.51** to **.54** (OOV).

6.2 TRiC and TRaC: masking settings

When pre-trained and fine-tuned models are used in the masking settings (i.e., +MASK and +FT+MASK), we observe a significant improvement in performance for both TRiC and TRaC. Notably, this improvement for TRiC is substantially larger compared to the one observed in the prior comparison (pre-trained vs. fine-tuned), with +FT+MASK exhibiting slightly superior performance to +MASK. We attribute this improvement to the fact that, in the masking settings, models are compelled to pay more attention to the surrounding contexts of reused texts, thereby fostering a more comprehensive understanding of topic relatedness.

For TRiC, we observe the following performance. For Subtasks 1, we observe a F1 ranging from .81 to .83 and from **.82** to **.86** for +MASK and +FT+MASK, respectively. For Subtask 2, we observe a SP coefficients ranging from .60 to .68 and from **.60** to **.72** for +MASK and +FT+MASK, respectively.

For TRaC, we observe SP coefficients ranging from **.65** to **.73**. Conversely, when pre-trained models are used in the +MASK setting, SP coefficients exhibit a substantial improvement, ranging from **.76** to **.84**.

6.3 Discussion

The results found in our experiments underscore the difficulty of SBERT models in distinguishing different text recontextualizations. This aligns with the work by MacLaughlin et al. (2021), where the performance of (off-the-shelf) SBERT for standard text reuse detection was underwhelming in comparison to lexical overlap baselines. As a matter of fact, pre-trained models exhibit a bias towards their typ-

ical pre-training focus, namely *semantic similarity*, while demonstrating only a superficial understanding of *topic relatedness*. Although the masking settings seem to offer a valuable workaround to sidestep the problem, we claim that their use is generally undesirable in real scenario involving text reuse. First, because masking may disrupt the natural flow of sentences precluding to obtain optimal performance. Second, because the boundaries of text reuse are often nuanced or unbalanced in different recontextualizations, when considering a form of text reuse broader than explicit quotation that implicitly reuses text *in-context*. In such cases, masking may result in the removal of crucial contextual information.

Consequently, to provide a more accurate modeling of text-reuse *in-context*, we argue that there is a clear imperative to develop or fine-tune novel models specifically tailored on topic relatedness. In this regard, `TRoTR` represents a valuable framework for evaluating language models that extend existing benchmarks on sentence-pair regression tasks, such as Semantic Textual Similarity (Agirre et al., 2012) and Semantic Textual Relatedness (Abdalla et al., 2023). While current benchmarks rely on a notion of *similarity* or *relatedness*, they overlook the potential impact of shared substrings, such as text-reuse excerpts, on computational estimates.

7 Concluding remarks and future work

To the best of our knowledge, this work represents a first pioneering effort in the computational modeling of *recontextualization*. We relied on the notion of *topic relatedness* to introduce a novel framework named Topic Relatedness of Text Reuse (`TRoTR`) with two tasks: Text Reuse in-Context (TRiC) and Topic variation Ranking across Corpus (TRaC). The tasks are inherently difficult as topic relatedness is under-defined, and under-researched, therefore this paper presents important steps forward.

First, we presented a human annotated benchmark of text reuse instances extracted from Twitter. This benchmark can be used to support Linguistic Recycling and Reception studies, ranging from misuse and dog whistles to the study of author influence. Using the framework, the benchmark can easily be extended in future work to cover more diverse sets of text reuse from other sources, e.g., literature and political text.

Next, we comprehensively evaluate SBERT models on the TRiC and TRaC tasks. We find that the Bi-Encoder model outperform the Cross-Encoder models. Additionally, we evaluate the considered models by masking the occurrences of text reuse and find that the models exhibit a greater sensitivity to semantic similarity rather than topic relatedness. These results now constitute a *baseline* for continued research and can be used as comparison for improved models and architectures.

Future work: Text reuse is inherently *diachronic* and can take place both over short and long time spans. The `TRoTR` framework is applicable to address the recontextualization problem across time, space, or domain. In our ongoing work, we will extend the TRoTR benchmark by annotating historical text and explicitly modeling change in topical variation over time. This will allow us to track the evolution of a quote like `To be or not to be` where Hamlet originally reflected on the struggles of existence and the fear of the unknown. Over the centuries, the phrase has become deeply embedded in various languages and cultures, often improperly referenced, quoted, and parodied in diverse literary works, contexts, and topics (Bate, 1985).

Acknowledgements

This work has in part been funded by the research program *Change is Key!* supported by Riksbankens Jubileumsfond (under reference number M21-0021). The computational resources were provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

8 Limitations

The main limitations of this work pertain to the benchmark, including the data collection and processing:

- *Manual tweet search*: we conducted a manual search of tweets by leveraging the Twitter search bar. This allowed us to sidestep a Text Reuse Detection phase and its validation. However, manually checking the suitability of retrieved tweets is extremely time consuming, thus limiting our ability to collect a large amount of tweets. Moreover, due to the Twitter ranking of matching results, the topic distribution of recontextualizations may be biased.

- *Randomization of the annotation instances*: in generating the pairs of tweets to compare for human judgement, we randomized the order of $\langle t, c_1, c_2 \rangle$ instances. However, we did not randomize the order of the two contexts within a pair. The ordering of c_1 and c_2 in $\langle t, c_1, c_2 \rangle$ was fixed and determined by their IDs. If item order influences annotator judgments, this may have created a bias towards certain orderings.

- *Human judgments*: we discarded some of judgments from human annotators to ensure high-quality of annotation results. This implied a high degree of imbalance in the distribution of TRiC labels for Subtask 1. We addressed and discussed this imbalance in the experimental results (see Section 5.2 and Appendix A).

As a further limitation, the TRoTR benchmark contains English tweets only with literal text reuse (i.e., explicit quotations). However, the benchmark can be extended to consider multi-language corpora and implicit text reuse.

As this work is the first of its kind to phrase a new problem, recontextualization of text-reuse, create a human-annotated benchmark, and attempt to solve the problem using computational tools, we do not claim our work to be exhaustive.

9 Ethical considerations

The authors have carefully considered the ethics associated with the TRoTR benchmark. The benchmark data, extracted from Twitter (now X), and annotations have been used while respecting the privacy and confidentiality of both users and annotators. For users, we made an effort to anonymize publicly available tweets' content by removing tweet mentions and users. For human annotators, we explicitly notified them prior to the annotation that some instances of text reuse might encompass discriminatory language against people or communities. We encourage the research community to approach our benchmark with a critical perspective, recognizing the potential ethical implications of working with data from social media platforms.

The annotation campaign was conducted with Native English speakers who were reached through email broadcasts. Compensation details, set in advance, were based on an hourly rate of €12. Each annotator spent a total of 53 hours on the annotation process, resulting in an overall compensation of €636. This fixed compensation was determined according to our time estimation. As per our contract terms, annotators received payment at the conclusion of the annotation campaign.

References

Mohamed Abdalla, Krishnapriya Vishnubhotla, and Saif Mohammad. 2023. What Makes Sentences Semantically Related? A Textual Relatedness Dataset and Empirical Study. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 782–796, Dubrovnik, Croatia. Association for Computational Linguistics.

Aly Abdelrazek, Yomna Eid, Eman Gawish, Walaa Medhat, and Ahmed Hassan. 2023. Topic Modeling Algorithms and Applications: A Survey. *Information Systems*, 112:102131.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. SEM 2013 shared

task: Semantic Textual Similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA. Association for Computational Linguistics.

Carlos Santos Armendariz, Matthew Purver, Matej Ulčar, Senja Pollak, Nikola Ljubešić, and Mark Granroth-Wilding. 2020. CoSimLex: A Resource for Evaluating Graded Word Similarity in Context. In *Proc. of LREC*, pages 5878–5886, Marseille, France. ELRA.

Jonathan Bate. 1985. Parodies of Shakespeare. *Journal of Popular Culture*, 19(1):75.

Nina Bauwelinck and Els Lefever. 2020. Annotating Topics, Stance, Argumentativeness and Claims in Dutch Social Media Comments: A Pilot Study. In *Proceedings of the 7th Workshop on Argument Mining*, pages 8–18, Online. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

John W. Du Bois. 2014. Towards a dialogic syntax. *Cognitive Linguistics*, 25(3):359–410.

Francis Bond and Graham Matthews. 2018. Toward An Epic Epigraph Graph. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Marco Büchler, Philip R. Burns, Martin Müller, Emily Franzini, and Greta Franzini. 2014. *Towards a Historical Text Re-use Detection*, pages 221–238. Springer International Publishing, Cham.

Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Pauline Hope Cheong. 2014. Tweet the message? religious authority and social media innovation. *Journal of Religion, Media and Digital Culture*, 3(3):1 – 19.

Stanford Chiu, Ibrahim Uysal, and W. Bruce Croft. 2010. Evaluating Text Reuse Discovery on the Web. In *Proceedings of the Third Symposium on Information Interaction in Context*, IIiX '10, page 299–304, New Brunswick, New Jersey, USA. Association for Computing Machinery.

Paul Clough, Robert Gaizauskas, Scott SL Piao, and Yorick Wilks. 2002. Measuring text reuse. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 152–159.

John H. Connolly. 2014. Recontextualisation, Resemiotisation and Their Analysis in Terms of an FDG-based Framework. *Pragmatics*, 24(2):377–397.

David DeVault and Matthew Stone. 2004. Interpreting Vague Utterances in Context. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1247–1253, Geneva, Switzerland. COLING.

Wai Chee Dimock. 1997. A Theory of Resonance. *PMLA*, 112(5):1060–1071.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2013. Measuring Word Meaning in Context. *Computational Linguistics*, 39(3):511–554.

Greta Franzini, Marco Passarotti, Maria Moritz, and Marco Büchler. 2018. Using and Evaluating TRACER for an Index Fontium Computatus of the Summa contra Gentiles of Thomas Aquinas. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018*, volume 2253 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Sayan Ghosh and Shashank Srivastava. 2022. ePiC: Employing Proverbs in Context as a Benchmark for Abstract Language Understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3989–4004, Dublin, Ireland. Association for Computational Linguistics.

Chris Greenough. 2021. From Biblical Text to Twitter: Teaching Biblical Studies in the Zeitgeist of# MeToo. *Journal of Feminist Studies in Religion*, 37(1):133–135.

Maarten Grootendorst. 2022. BERTopic: Neural Topic Modeling with a Class-based TF-IDF Procedure.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, pages 1735–1742. IEEE Computer Society.

Matthias Hagen and Benno Stein. 2011. Candidate Document Retrieval for Web-Scale Text Reuse Detection. In *String Processing and Information Retrieval*, pages 356–367, Berlin, Heidelberg. Springer Berlin Heidelberg.

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. 2014. *Cohesion in English*. 9. Routledge.

David Harbecke, Yuxuan Chen, Leonhard Hennig, and Christoph Alt. 2022. Why only Micro-F1? Class Weighting of Measures for Relation Classification. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 32–41, Dublin, Ireland. Association for Computational Linguistics.

Tomáš Hercig and Pavel Kral. 2021. Evaluation Datasets for Cross-lingual Semantic Textual Similarity. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 524–529, Held Online. INCOMA Ltd.

Niclas Hertzberg, Robin Cooper, Elina Lindgren, Björn Rönnerstrand, Gregor Rettenegger, Ellen Breitholtz, and Asad Sayeed. 2022. Distributional properties of political dogwhistle representations in Swedish BERT. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 170–175, Seattle, Washington (Hybrid). Association for Computational Linguistics (ACL).

Peter Uwe Hohendahl and Marc Silberman. 1977. Introduction to Reception Aesthetics. *New German Critique*, 10:29–63.

Eduard Hovy and Chin-Yew Lin. 1998. Automated Text Summarization and the Summarist System. In *TIPSTER TEXT PROGRAM PHASE III: Proceedings of a Workshop held at Baltimore, Maryland, October 13-15, 1998*, pages 197–214, Baltimore, Maryland, USA. Association for Computational Linguistics.

Shotaro Ishihara and Hono Shirai. 2022. Nikkei at SemEval-2022 task 8: Exploring BERT-based bi-encoder approach for pairwise multilingual news article similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1208–1214, Seattle, United States. Association for Computational Linguistics.

Xianlin Jin and Patric R. Spence. 2021. Understanding crisis communication on social media with CERC: topic model analysis of tweets about Hurricane Maria. *Journal of Risk Research*, 24(10):1266–1287.

Anton S. Khritankov, Pavel V. Botov, Nikolay S. Surovenko, Sergey V. Tsarkov, Dmitriy V. Viuchnov, and Yuri V. Chekhovich. 2015. Discovering text reuse in large collections of documents: A study of theses in history sciences. In *2015 Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT)*, pages 26–32.

Munui Kim, Injun Baek, and Min Song. 2018. Topic Diffusion Analysis of a Weighted Citation Network in Biomedical Literature. *Journal of the Association for Information Science and Technology*, 69(2):329–342.

Miloslav Konopík, Ondřej Pražák, and David Steinberger. 2017. Czech Dataset for Semantic Similarity and Relatedness. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 401–406, Varna, Bulgaria. INCOMA Ltd.

Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications.

Andrey Kutuzov and Lidia Pivovarova. 2021. Rushifteval: a shared task on semantic shift detection for russian. *Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialog conference*.

Hanbit Lee, Yeonchan Ahn, Haejun Lee, Seungdo Ha, and Sang-goo Lee. 2016. Quote recommendation in dialogue using deep neural network. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 957–960, New York, NY, USA. Association for Computing Machinery.

Hyun Seung Lee, Seungtaek Choi, Yunsung Lee, Hyeongdon Moon, Shinhyeok Oh, Myeongho Jeong, Hyojun Go, and Christian Wallraven. 2023. Cross encoding as augmentation: Towards effective educational text classification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2184–2195, Toronto, Canada. Association for Computational Linguistics.

Per Linell. 1998. *Approaching Dialogue: Talk, Interaction and Contexts in Dialogical Perspectives*. John Benjamins.

Daniel Loureiro, Aminette D'Souza, Areej Nasser Muhajab, Isabella A. White, Gabriel Wong, Luis Espinosa-Anke, Leonardo Neves, Francesco Barbieri, and Jose Camacho-Collados. 2022. TempoWiC: An Evaluation Benchmark for Detecting Meaning Shift in Social Media. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3353–3359, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ansel MacLaughlin, Shaobin Xu, and David A. Smith. 2021. Recovering Lexically and Semantically Reused Texts. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 52–66, Online. Association for Computational Linguistics.

Enrique Manjavacas, Brian Long, and Mike Kestemont. 2019. On the Feasibility of Automated Detection of Allusive Text Reuse.

Terence E. McDonnell, Christopher A. Bail, and Iddo Tavory. 2017. A Theory of Resonance. *Sociological Theory*, 35(1):1–14.

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Maria Moritz, Andreas Wiederhold, Barbara Pavlek, Yuri Bizzoni, and Marco Büchler. 2016. Non-Literal Text Reuse in Historical Texts: An Approach to Identify Reuse Transformations and its Application to

Bible Reuse. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1849–1859, Austin, Texas. Association for Computational Linguistics.

Iqra Muneer and Rao Muhammad Adeel Nawab. 2022. Cross-Lingual Text Reuse Detection at sentence level for English–Urdu language pair. *Computer Speech & Language*, 75:101381.

Naho Orita, Naomi Feldman, Jordan Boyd-Graber, and Eliana Vornov. 2014. Quantifying the Role of Discourse Topicality in Speakers' Choices of Referring Expressions. In *Proceedings of the Fifth Workshop on Cognitive Modeling and Computational Linguistics*, pages 63–70, Baltimore, Maryland, USA. Association for Computational Linguistics.

Francesco Periti and Stefano Montanelli. 2024. Lexical Semantic Change through Large Language Models: a Survey. *ACM Comput. Surv.*, 56(11).

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Jack A. Porrino, Virak Tan, and Aaron Daluiski. 2008. Misquotation of a Commonly Referenced Hand Surgery Study. *The Journal of Hand Surgery*, 33(1):2.e1–2.e9.

Forough Poursabzi-Sangdeh and Jordan Boyd-Graber. 2015. Speeding Document Annotation with Topic Models. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 126–132, Denver, Colorado. Association for Computational Linguistics.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics.

Dominik Schlechtweg, Shafqat Mumtaz Virk, Pauline Sander, Emma Sköldberg, Lukas Theuer Linke, Tuo Zhang, Nina Tahmasebi, Jonas Kuhn, and Sabine Schulte im Walde. 2024. The DURel annotation tool: Human and computational measurement of semantic proximity, sense clusters and semantic change. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 137–149, St. Julians, Malta. Association for Computational Linguistics.

Jangwon Seo and W. Bruce Croft. 2008. Local Text Reuse Detection. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, page 571–578, New York, NY, USA. Association for Computing Machinery.

David A. Smith, Ryan Cordell, and Elizabeth Maddock Dillon. 2013. Infectious texts: Modeling text reuse in nineteenth-century newspapers. In *2013 IEEE International Conference on Big Data*, pages 86–94.

Charles Spearman. 1904. The proof and measurement of association between two things. *American Journal of Psychology*, 15:88–103.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2021. Survey of Computational Approaches to Lexical Semantic Change Detection. In *Computational approaches to semantic change*. Language Science Press.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.

Martyn P. Thompson. 1993. Reception Theory and the Interpretation of Historical Meaning. *History and Theory*, 32(3):248–272.

Lingzhi Wang, Xingshan Zeng, and Kam-Fai Wong. 2021. Quotation Recommendation and Interpretation Based on Transformation from Queries to Quotations. 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 754–758, Online. Association for Computational Linguistics.

Lingzhi Wang, Xingshan Zeng, and Kam-Fai Wong. 2022. Learning when and what to quote: A quotation recommender system with mutual promotion of recommendation and generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3094–3105, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lingzhi Wang, Xingshan Zeng, and Kam-Fai Wong. 2023. Quotation Recommendation for Multi-Party Online Conversations Based on Semantic and Topic Fusion. *ACM Trans. Inf. Syst.*, 41(4).

Sean Wilner, Daniel Woolridge, and Madeleine Glick. 2021. Narrative Embedding: Re-Contextualization Through Attention. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1405, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shaobin Xu, David Smith, Abigail Mullen, and Ryan Cordell. 2014. Detecting and Evaluating Local Text Reuse in Social Networks. In *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, pages 50–57, Baltimore, Maryland. Association for Computational Linguistics.

Appendix

A Train-Dev-Test partitions

For each randomized split, we use the filtered instances (see Section 4.2) to create the Train-Dev-Test partitions, comprising approximately 80%, 10%, and 10% of the instances, respectively. In the creation of the Train set of a split, we exclude the $\langle t, c_1, c_2 \rangle$ instances associated to four targets t (i.e., 10% of the benchmark's targets). We include these instances in Dev and Test to enforce the Out-of-Vocabulary (OOV) evaluation. Specifically, we include in Dev the instances associated with two targets, and in Test the instances of the remaining excluded targets.

Notably, we ensure that each partition has a distinct set of OOV targets, such that the intersection of the OOV sets for each split is empty.

B Model evaluation

We evaluate almost all the pre-trained models available at `https://www.sbert.net/index.html`. Specifically, we considered only pre-trained models trained on tasks based on textual similarity and excluded those trained on other tasks (e.g., models for Image Search). Table 6 reports results for all the evaluated models.

For the sake of transparency and completeness, we have included the computation of Precision (PR) and Recall (RE) for each considered class. Specifically, for label 1, PR and RE are calculated as $\frac{TP}{(TP+FP)}$ and $\frac{TP}{(TP+FN)}$ respectively. Similarly, for label 0, PR and RE are computed as $\frac{TN}{(TN+FN)}$ and $\frac{TN}{(TN+FP)}$. In scientific literature, these latter metrics are also known as Negative Predictive Value and Sensitivity. For the sake of clarity, we preferred using PR and RE for *label 0* and *label 1* instead of distinguishing between Precision (PR), Recall (RE), Negative Predictive Value (NPV), and Specificity (SP).

C Fine-tuning

For each randomized split, we fine-tuned each considered model on the Train set and subsequently validated its performance on the Dev set. To do this, we employed the AdamW optimizer, coupled with a linear learning rate warm-up applied to the first 10% of the Train set. We used grid search to optimize hyper-parameters, with a particular focus on fine-tuning the learning rate by testing values from the set {1e-6, 2e-6, 5e-6, 1e-5, 2e-5}. We do not use weight decay, since our initial experiments did not yield any additional benefits. During the training, we leveraged an early stopping strategy. In particular, we fine-tuned each pre-trained model on TRiC instances using the *contrastive loss* (Hadsell et al., 2006). This loss minimizes the distance between embeddings of similar sentences and maximizes the distance for dissimilar sentences. We finally ceased training when there was no further improvement observed on the Dev set. Details on the setup of hyper-parameters are shown in Table 4.

D Hyper-parameters

Models	Learning Rate
`all-distilroberta-v1` (ADR)	1e-05
`distiluse-base-multilingual-cased-v1` (DBM)	1e-05
`paraphrase-multilingual-MiniLM-L12-v2` (PAM)	2e-05
`paraphrase-multilingual-mpnet-base-v2` (PAR)	5e-06
`multi-qa-mpnet-base-cos-v1` (MQA)	1e-05

Table 4: Models with learning rates.

E Tweets' length

In Figure 2, we illustrate the distribution of tweet lengths within the `TRoTR` benchmark.

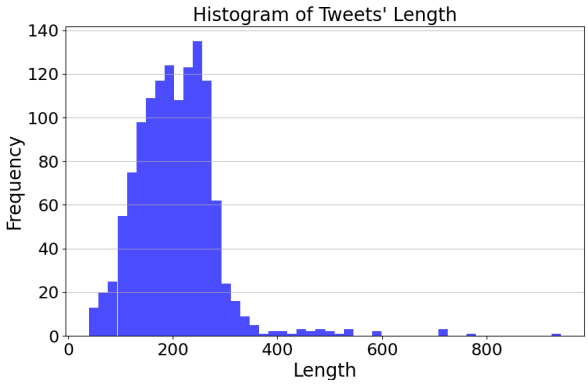

Figure 2: The histogram displays the distribution of tweet lengths in the TROTR benchmark, with the x-axis representing the length of tweets (in characters) and the y-axis indicating the frequency of tweets within specified length ranges. A total of 50 bins are used to provide a detailed view of the length distribution.

F Annotation

Annotating topic relatedness, instead of relying on explicit topic labels, closely resembles recent work exemplified in the Word-in-Context task (Pilehvar and Camacho-Collados, 2019), which relies on annotating word meaning relatedness rather than explicit sense labels. The methodology underlying this approach is thoroughly elucidated in our guidelines, submitted as supplementary material along with our paper. The topic relatedness is evaluated by using the four-point DURel relatedness scale (see Figure 1). Annotator were trained in a 30-minute online session and tested on a small set of 25 instances (tutorial). In particular, we ensured that each annotator achieved a minimum agreement (measured by Spearman correlation) of at least .550 with the tutorial judgments. We interpreted these results as reliable, and consequently, we proceeded with the annotation of our benchmark. Then, we derive TRiC and TRaC labels after conducting an empirical analysis of the agreement of each level of our topic relatedness scale (see Section 4.2).

G Annotation guidelines

Annotation guidelines

Your task is to rate the degree of topic relatedness between two texts in which a *text sequence* is used. For instance, presented with a pair as in the below table, you are asked to rate the topic relatedness of the texts in which **Love your neighbor as yourself** is used.

Text 1	Text 2
Love your neighbor as yourself. There is no commandment greater than these. You're a hypocritical Christian who ignores the greatest commandment because you're a bigot.	Jesus didn't tell you to be a bigot! Jesus had nothing to say about LGBTQIA+ people, but he did say to **love your neighbor as yourself**. #loveislove ❤️ 🧡 💛 💚 💙 💜

What is a topic?

The topic of a text answers the question "*What is this text about?*"

An example of a topic for **Text 1** and **Text 2** above is *bigotry*. However, you may identify a different topic for **Text 1** or **Text 2**, as the perception of a text is subjective. For example, in **Text 1** you may identify *hypocrisy* as the topic, while in **Text 2** you may identify *LGBTQIA+* as the topic.

It is also often the case that multiple topics can be identified in one text. For example, in **Text 1**, possible topics may include: *bigotry*, *hypocrisy*, *commandment*. In **Text 2**, possible topics may include: *bigotry*, *LGBTQIA+*, *love*.

Do not worry about finding the exact words to describe the topic. Just make sure you have a clear idea to compare how different topics relate to each other. Indeed, your task is to rate how closely related topics are, not to label them with specific names.

Task structure

You will be shown two texts displayed next to each other. In both texts, a common subsequence is marked in **bold**. Your task is to evaluate, for each of these pairs, how strong the topic relatedness is between the two texts.

Note that the topic information is not available, so you are asked to identify the latent topics in the texts before rating. While a common subsequence is marked in bold, please focus on the entire text during your evaluation. It is essential that you first read each text in a pair individually and determine the most plausible topic(s) for that text **before** comparing the two texts in the pair.

The judgment scale

The scale that you will be using for your judgments ranges from 1 (the two texts in a pair have completely unrelated topics) to 4 (the two texts in a pair have precisely identical topics). This four-point scale of topic relatedness is shown below.

- 4 – Identical
- 3 – Closely Related
- 2 – Distantly related
- 1 – Unrelated
- – Can't decide

Annotation examples

We now consider some evaluation examples to illustrate the different degrees of topic relatedness you might encounter in annotation. Please note that these are only examples, and you should always give your subjective opinion.

Example A: Judgment 4-Identical

The two texts in **Example A** are judged to be addressing the same topic (rating: 4) since both texts refer to the *reliance on a higher power for strength and support during challenging times*.

Text 1	Text 2
In the midst of life's storms, when fear and uncertainty surround us, let us remember to trust in God. Remember his word: **fear not, for I am with you**; do not be dismayed, for I am your God. I will strengthen you and help you.	Sometimes, I just feel like giving up... But the Lord gives me strength to keep going. **So do not fear, for I am with you**; Be not dismayed, for I am your God. I will strengthen you, Yes, I will help you, I will uphold you with My righteous right hand.

Example B: Judgment 3-Closely Related

In contrast to the previous example, the two texts in **Example B** are judged to be closely related in topic (rating: 3) as they both refer to *bigotry*. However, there is some difference in the topic(s) expressed in Text 1 (*accusing someone of hypocrisy and bigotry*) compared to Text 2 (*promoting love and acceptance*).

Text 1	Text 2
Love your neighbor as yourself. There is no commandment greater than these. You're a hypocritical Christian who ignores the greatest commandment because you're a bigot.	Jesus didn't tell you to be a bigot! Jesus had nothing to say about LGBTQIA+ people, but he did say to **love your neighbor as yourself**. #loveislove ❤️🧡💛💚💙💜

Example C: Judgment 2-Distantly Related

In **Example C**, the two texts are judged to be distantly related in topic (rating: 2) because, while they share a common aspect (i.e., time), they emphasize the *balance between work and rest* and the *constant checking of the time throughout the day*, respectively.

Text 1	Text 2
For everything there is a season, a time for every activity under heaven. As we embrace the weekend, let's remember to strike a balance between work and rest ⚖️, allowing ourselves time to rejuvenate and find inspiration in the world around us. 🌎🌞	**For everything there is a season, a time for every activity under heaven**. I don't know about you, but I constantly look at my watch throughout the day 🕰️🕰️. What time is it? What time are we supposed to be there? How much time will it take?

Example D: Judgment 1-Unrelated

A rating of 1 is assigned to two texts of a target sequence that are entirely unrelated in the topics they express, as seen in **Example D**. Note that this pair of texts is more different than the two texts in **Example C**.

Text 1	Text 2
At a large Crimean event today Putin quoted the Bible to defend the special military operation in Ukraine which has killed thousands and displaced millions. His words **Greater love has no one than this: to lay down one's life for one's friends**. And people were cheering him. Madness!!!	It's the wonderful pride month!! ❤️🧡💛💚💙💜 Honestly pride is everyday! Love is love don't forget I love you. Remember this!: My command is this: Love each other as I have loved you. **Greater love has no one than this: to lay down one's life for one's friends**.

Social media data

The texts provided for the annotation task were gathered from Twitter and may contain offensive language, discriminatory content, and other sensitive material.

Some texts may occur more than once during annotation. They may vary in length, ranging from very short to very long, and some may appear ungrammatical. Additionally, you may encounter words spelled differently than you are used to (e.g., *veeeeery*), and some abbreviations may be used (e.g., *lol*, i.e. *lots of laugh*).

Try to disregard these issues during the annotation.

H Inter-annotator agreement

Id *Bible reference*	**Target sequence** *Reused text*	**Agreement** *Krippendorf's* α	**Agreement** *Avg. pairwise Spearman's* ρ	**# Instances** *Number of pairs*
Overall	***Overall***	***.420 / .709***	***.506 / .811***	***6,300 / 3,821***
(Matthew 18:22)	Seventy times seven	.118 / .764	.619 / .857	150 / 95
(John 17:21)	That all may be one	.494 / .677	.183 / .782	150 / 79
(Matthew 5:39)	Turn the other cheek	-.036 / .193	.510 / .405	120 / 132
(Matthew 7:7)	Seek and you will find	.210 / .097	.558 / .475	150 / 125
(Psalm 23:1)	The Lord is my shepherd	.213 / .138	.476 / .431	150 / 117
(John 8:32)	The truth will set you free	.250 / .217	.347 / .368	150 / 10
(1 Corinthians 13:4)	Love is patient, love is kind	.282 / .798	.382 / .834	150 / 61
(Matthew 7:1)	Judge not, that ye be not judged	.472 / .450	.555 / .469	150 / 96
(Ecclesiastes 3:1)	For everything there is a season	.263 / .369	.443 / .557	150 / 104
(Romans 8:28)	All things work together for good	.110 / -.030	.543 / .430	150 / 128
(2 Corinthians 5:7)	For we walk by faith, not by sight	.383 / .823	.437 / .866	150 / 79
(Psalm 121:7)	The Lord will keep you from all harm	.169 / .178	.213 / .166	150 / 103
(Mark 12:17)	Give to Caesar what belongs to Caesar	.196 / .117	.460 / .522	150 / 106
(Proverbs 27:5)	Better is open rebuke than hidden love	.431 / .828	.481 / .911	150 / 95
(Exodus 20:3)	You shall have no other gods before me	.259 / .506	.331 / .646	150 / 80
(Genesis 1:1)	In the beginning God created the heaven	.151 / .177	.219 / .277	150 / 66
(Romans 12:10)	Love one another with brotherly affection	.410 / .566	.571 / .657	150 / 101
(Leviticus 20:13)	If a man lies with a male as with a woman	.315 / .492	.339 / .517	150 / 86
(Joshua 1:9)	Be strong and courageous. Do not be afraid	.223 / .822	.281 / .833	150 / 69
(Mark 9:23)	Everything is possible for one who believes	.081 / .509	.118 / .557	150 / 81
(Philippians 4:13)	I can do all things through Christ who strengthens me	.128 / .624	.349 / .787	150 / 73
(Ephesians 5:25)	Husbands, love your wives, as Christ loved the church	.487 / .802	.507 / .802	150 / 95
(Matthew 5:44)	Love your enemies and pray for those who persecute you	.073 / -.004	.389 / .532	150 / 104
(John 15:12)	My command is this: Love each other as I have loved you	.288 / .589	.426 / .790	150 / 97
(Isaiah 43:4)	You are precious in my eyes and honored, and I love you	.421 / .439	.625 / .730	150 / 110
(Matthew 7:25)	The rain came down, the streams rose, and the winds blew	.166 / .625	.406 / .802	150 / 71
(1 Timothy 2:12)	But I suffer not a woman to teach, nor to usurp authority	.302 / .232	.315 / .217	150 / 71
(Proverbs 10:12)	Hatred stirs up conflict, but love covers over all wrongs	.172 / .148	.377 / .517	150 / 100
(Hosea 8:7)	They have sown the wind, and they shall reap the whirlwind	.261 / .621	.423 / .668	150 / 55
(Proverbs 12:25)	Anxiety weighs down the heart, but a kind word cheers it up	.093 / .518	.253 / .777	150 / 81
(1 John 4:8)	Whoever does not love does not know God, because God is love	.329 / .355	.373 / .393	150 / 121
(Solomon 4:7)	You are altogether beautiful, my darling; there is no flaw in you	.385 / .782	.537 / .878	150 / 92
(Leviticus 18:22)	You shall not lie with a male as with a woman; it is an abomination	.423 / .648	.458 / .753	150 / 101
(Psalm 118:24)	This is the day that the Lord has made; let us rejoice and be glad in it	.294 / .847	.492 / .884	150 / 77
(Proverbs 31:10)	A wife of noble character who can find? She is worth far more than rubies	.108 / .775	.261 / .797	150 / 74
(John 15:13)	Greater love has no one than this: to lay down one's life for one's friends	.347 / .355	.640 / .737	150 / 125
(Matthew 11:28)	Come to me, all you who labour and are overburdened, and I shall give you rest	-.007 / -.024	.268 / .355	150 / 101
(Jeremiah 17:9)	The heart is deceitful above all things, and desperately wicked; who can know it?	.164 / .432	.311 / .591	150 / 75
(Hebrews 11:1)	Now faith is confidence in what we hope for and assurance about what we do not see	.391 / .853	.410 / .870	150 / 72
(Luke 17:3)	Take heed to yourselves. If your brother sins against you, rebuke him; and if he repents, forgive him	-.011 / .267	.124 / .485	150 / 91
(2 Corinthians 5:17)	Therefore, if anyone is in Christ, he is a new creation. The old has passed away; behold, the new has come	.307 / .655	.414 / .744	150 / 75
(1 Samuel 16:7)	The Lord does not look at the things people look at. People look at the outward appearance, but the Lord looks at the heart	.432 / .665	.508 / .733	150 / 80

Table 5: Biblical passages included in TRoTR and their inter-annotator agreement agreement. We report data using the x / y format, where x denotes the data on the entire set of instance pairs, and y denotes the data post-filtering process.

	Standard Test Set								Out-of-vocabulary (OOV) Test set							
	Label 0			*Label 1*			***All***		*Label 0*			*Label 1*			***All***	
Models	PR	RE	F1	PR	RE	F1	F1	SP	PR	RE	F1	PR	RE	F1	F1	SP
paraphrase-multilingual-MiniLM-L12-v2	**.96±.02**	.46±.09	.61±.08	.41±.09	.96±.02	.57±.08	.61±.07	**.58±.08**	.96±.04	.43±.17	.57±.16	.37±.15	.95±.05	.52±.15	.59±.12	**.49±.22**
+MASK	*.89±.05*	*.88±.06*	*.88±.03*	*.71±.10*	*.72±.10*	*.70±.05*	*.83±.03*	*.67±.04*	*.89±.09*	*.86±.09*	*.87±.06*	*.65±.19*	*.71±.18*	*.65±.12*	*.83±.05*	*.60±.13*
multi-qa-mpnet-base-cos-v1	.94±.03	.42±.11	.58±.11	.40±.10	.94±.03	.55±.09	.58±.09	.55±.09	.94±.09	.39±.19	.53±.20	.36±.19	.96±.03	.50±.18	.55±.16	**.49±.21**
+MASK	*.88±.05*	*.87±.07*	*.88±.04*	*.71±.10*	*.71±.12*	*.69±.06*	*.83±.04*	*.68±.05*	*.89±.07*	*.86±.10*	*.87±.06*	*.63±.18*	*.69±.16*	*.63±.13*	*.83±.05*	*.62±.13*
all-distilroberta-v1	.95±.03	.47±.13	.62±.11	.42±.11	.93±.04	.57±.10	.61±.10	.55±.09	.94±.07	.45±.20	.58±.20	.38±.19	.93±.06	.51±.18	.58±.16	.48±.20
+MASK	*.89±.05*	*.87±.07*	*.87±.03*	*.70±.14*	*.72±.12*	*.69±.07*	*.82±.03*	*.67±.06*	*.90±.07*	*.85±.10*	*.87±.05*	*.62±.21*	*.71±.18*	*.63±.14*	*.82±.05*	*.62±.15*
all-mpnet-base-v2	.93±.03	.48±.14	.62±.13	.42±.12	.91±.03	.57±.10	.61±.11	.53±.10	.93±.09	.44±.22	.56±.21	.38±.20	.94±.05	.51±.18	.57±.18	.48±.20
+MASK	*.88±.06*	*.84±.09*	*.85±.05*	*.66±.12*	*.71±.11*	*.67±.04*	*.81±.04*	*.66±.06*	*.89±.08*	*.82±.11*	*.85±.07*	*.59±.20*	*.73±.14*	*.62±.11*	*.81±.05*	.61±.15
paraphrase-multilingual-mpnet-base-v2	.95±.03	.40±.10	.56±.09	.39±.09	.95±.04	.55±.08	.56±.07	.56±.09	.93±.11	.35±.18	.49±.19	.34±.15	.95±.06	.49±.16	.52±.15	.47±.25
+MASK	*.89±.05*	*.85±.07*	*.87±.04*	*.69±.10*	*.75±.11*	*.70±.05*	*.83±.03*	*.68±.03*	*.90±.08*	*.83±.13*	*.86±.07*	*.63±.19*	*.75±.17*	*.65±.10*	*.82±.05*	.62±.11
all-MiniLM-L12-v2	.95±.03	.40±.12	.55±.11	.39±.11	.94±.04	.54±.10	.55±.10	.52±.08	.94±.03	.37±.18	.50±.18	.35±.18	.93±.06	.48±.18	.52±.16	.47±.17
+MASK	*.88±.05*	*.87±.08*	*.87±.03*	*.70±.13*	*.72±.11*	*.69±.06*	*.82±.03*	*.68±.04*	*.89±.08*	*.85±.10*	*.86±.05*	*.62±.21*	*.71±.17*	*.62±.14*	*.82±.04*	*.62±.13*
multi-qa-distilbert-cos-v1	**.96±.03**	.33±.11	.48±.11	.37±.10	**.97±.02**	.53±.09	.50±.10	.53±.09	**.97±.06**	.29±.18	.42±.19	.33±.16	**.97±.05**	.47±.17	.46±.16	.47±.21
+MASK	*.88±.06*	*.86±.06*	*.87±.03*	*.68±.11*	*.73±.10*	*.69±.05*	*.82±.03*	*.68±.05*	*.89±.10*	*.85±.08*	*.86±.05*	*.61±.19*	*.71±.14*	*.63±.11*	*.82±.05*	*.62±.14*
multi-qa-mpnet-base-dot-v1	.92±.05	.48±.14	.62±.11	.42±.11	.89±.06	.56±.09	.61±.09	.51±.10	.91±.15	.45±.19	.59±.17	.38±.18	.92±.07	.51±.17	.59±.14	.46±.22
+MASK	*.87±.07*	*.86±.08*	*.86±.03*	*.69±.12*	*.65±.19*	*.63±.08*	*.80±.03*	*.63±.05*	*.87±.09*	.85±.09	*.85±.06*	*.62±.23*	*.63±.20*	*.57±.13*	*.80±.05*	*.57±.12*
all-MiniLM-L6-v2	**.96±.02**	.40±.10	.55±.10	.39±.10	.95±.04	.55±.09	.56±.08	.53±.09	**.97±.03**	.37±.17	.51±.19	.35±.17	.95±.07	.49±.18	.54±.14	.44±.23
+MASK	*.88±.05*	*.88±.06*	*.88±.03*	*.72±.12*	*.70±.12*	*.69±.06*	*.83±.03*	*.67±.05*	*.89±.07*	*.88±.09*	*.88±.05*	*.67±.22*	*.66±.19*	*.62±.14*	*.83±.04*	*.61±.16*
distiluse-base-multilingual-cased-v1	**.96±.02**	.26±.12	.40±.14	.35±.09	**.97±.03**	.51±.09	.43±.12	.54±.09	.96±.08	.21±.19	.31±.23	.31±.14	**.97±.05**	.45±.16	.38±.18	.44±.23
+MASK	*.87±.07*	*.88±.07*	*.87±.03*	*.72±.14*	*.66±.16*	*.66±.09*	*.81±.03*	*.64±.04*	*.88±.09*	*.88±.09*	*.87±.05*	*.66±.23*	*.64±.25*	*.58±.19*	*.82±.04*	*.58±.12*
distiluse-base-multilingual-cased-v2	**.96±.03**	.26±.08	.40±.10	.34±.09	.97±.03	.50±.09	.43±.09	.54±.10	.96±.08	.21±.16	.32±.20	.30±.15	.96±.08	.44±.17	.38±.16	.44±.25
+MASK	*.87±.06*	*.89±.07*	*.87±.03*	*.72±.14*	*.66±.14*	*.66±.09*	*.82±.03*	*.65±.04*	*.88±.08*	*.88±.10*	*.87±.05*	*.66±.24*	*.64±.23*	*.60±.18*	*.82±.05*	*.59±.12*
multi-qa-distilbert-dot-v1	.93±.04	.40±.12	.55±.11	.39±.09	.92±.05	.54±.09	.56±.09	.51±.09	.92±.12	.36±.16	.50±.16	.34±.15	.92±.07	.48±.16	.53±.11	.43±.19
+MASK	*.85±.05*	*.87±.08*	*.85±.03*	*.69±.15*	*.60±.16*	*.61±.08*	*.79±.02*	*.62±.05*	*.86±.09*	*.87±.09*	*.86±.05*	*.66±.24*	*.58±.22*	*.55±.16*	*.80±.03*	*.57±.14*
paraphrase-albert-small-v2	**.96±.02**	.36±.09	.52±.09	.38±.09	.96±.02	.54±.09	.53±.07	.53±.09	.95±.10	.32±.16	.46±.18	.33±.14	**.97±.04**	.48±.16	.50±.12	.43±.25
+MASK	*.88±.06*	*.84±.07*	*.86±.03*	*.65±.11*	*.70±.14*	*.66±.07*	*.80±.02*	*.65±.05*	*.88±.08*	*.82±.12*	*.84±.07*	*.56±.19*	*.67±.20*	*.58±.14*	*.80±.05*	*.57±.14*
multi-qa-MiniLM-L6-cos-v1	.95±.03	.37±.09	.52±.09	.38±.10	.95±.04	.53±.10	.53±.08	.52±.10	.91±.14	.34±.18	.48±.19	.34±.17	.94±.08	.47±.17	.50±.16	.42±.25
+MASK	*.88±.05*	*.88±.04*	*.88±.02*	*.70±.09*	*.69±.09*	*.68±.06*	*.83±.02*	*.66±.04*	*.87±.09*	*.87±.07*	*.87±.05*	*.61±.19*	*.64±.19*	*.60±.14*	*.82±.04*	*.60±.15*
cross-encoder/stsb-roberta-large	.33±.08	**.99±.02**	.49±.09	**.97±.03**	.19±.10	.30±.13	.36±.11	.52±.07	.29±.14	**.99±.03**	.42±.16	**.94±.15**	.11±.15	.18±.19	.28±.15	.42±.20
+MASK	*.70±.13*	*.68±.15*	*.66±.07*	*.87±.06*	*.87±.08*	*.87±.03*	*.81±.03*	*.66±.04*	*.62±.27*	*.64±.28*	*.57±.21*	*.87±.10*	*.86±.11*	*.86±.06*	*.80±.06*	*.62±.08*
paraphrase-MiniLM-L3-v2	.95±.03	.28±.09	.43±.10	.35±.08	.96±.03	.51±.09	.46±.08	.49±.11	.96±.05	.23±.19	.34±.21	.31±.14	**.97±.05**	.45±.16	.41±.17	.40±.27
+MASK	*.87±.05*	*.86±.09*	*.86±.04*	*.68±.12*	*.68±.11*	*.66±.05*	*.81±.03*	*.65±.04*	*.88±.07*	*.85±.12*	*.86±.06*	*.61±.21*	*.64±.22*	*.59±.15*	*.81±.05*	*.59±.14*
msmarco-distilbert-dot-v5	.93±.04	.36±.10	.51±.09	.37±.09	.93±.03	.52±.08	.52±.08	.47±.08	.92±.09	.31±.18	.43±.20	.32±.14	.92±.08	.46±.15	.48±.13	.38±.19
+MASK	*.87±.05*	*.91±.04*	*.89±.03*	*.75±.10*	*.66±.08*	*.69±.06*	*.84±.03*	*.64±.04*	*.87±.09*	*.90±.05*	*.88±.06*	*.67±.18*	*.60±.16*	*.61±.15*	*.83±.06*	*.58±.10*
msmarco-MiniLM-L12-cos-v5	.91±.04	.44±.09	.59±.08	.39±.09	.90±.05	.54±.08	.58±.07	.44±.08	.91±.09	.44±.17	.58±.16	.36±.16	.88±.10	.49±.16	.59±.12	.38±.19
+MASK	*.85±.05*	*.88±.06*	*.86±.03*	*.68±.11*	*.60±.10*	*.62±.06*	*.80±.03*	*.59±.04*	*.85±.10*	*.88±.08*	*.86±.06*	.62±.21	*.55±.21*	*.53±.16*	*.79±.05*	*.54±.12*
multi-qa-MiniLM-L6-dot-v1	.89±.07	.54±.07	**.67±.06**	.42±.09	.84±.08	.55±.08	**.64±.05**	.46±.10	.87±.16	.51±.15	**.63±.15**	.37±.16	.83±.11	.49±.15	.62±.12	.37±.26
+MASK	*.83±.07*	*.86±.06*	*.84±.03*	*.61±.13*	*.56±.12*	*.56±.07*	*.76±.04*	*.53±.07*	*.82±.12*	*.86±.08*	*.83±.07*	*.53±.23*	*.50±.20*	*.47±.17*	*.76±.08*	*.45±.18*
msmarco-MiniLM-L6-cos-v5	.93±.03	.41±.10	.56±.10	.39±.09	.92±.06	.54±.09	.56±.08	.44±.10	.93±.07	.38±.18	.52±.18	.34±.16	.91±.12	.48±.17	.54±.14	.37±.22
+MASK	*.85±.06*	*.87±.07*	*.86±.04*	*.67±.10*	*.62±.14*	*.62±.07*	*.79±.03*	*.59±.04*	*.85±.11*	*.86±.09*	*.85±.07*	*.60±.17*	*.58±.24*	*.55±.16*	*.79±.05*	*.54±.11*
msmarco-distilbert-base-tas-b	.93±.04	.36±.13	.51±.13	.38±.09	.93±.05	.53±.09	.52±.11	.45±.10	.92±.10	.32±.22	.44±.21	.33±.15	.92±.10	.47±.16	.48±.17	.36±.23
+MASK	*.86±.07*	*.86±.08*	*.86±.03*	*.67±.14*	*.64±.14*	*.63±.07*	*.80±.03*	*.62±.05*	*.86±.11*	*.87±.11*	*.85±.06*	*.61±.23*	*.59±.26*	*.53±.20*	*.79±.07*	*.56±.14*
cross-encoder/stsb-distilroberta-base	.33±.08	.96±.04	.49±.08	.94±.06	.23±.10	.35±.12	.40±.10	.43±.08	.29±.14	.96±.06	.43±.15	.89±.21	.17±.16	.27±.19	.34±.15	.36±.21
+MASK	*.66±.13*	*.61±.15*	*.61±.07*	*.85±.07*	*.86±.09*	*.85±.04*	*.78±.04*	*.59±.04*	*.58±.22*	*.56±.25*	*.51±.17*	*.85±.11*	*.84±.12*	*.84±.07*	*.77±.07*	*.55±.08*
msmarco-distilbert-cos-v5	.94±.03	.30±.09	.45±.11	.36±.09	.95±.03	.51±.09	.48±.09	.42±.09	.91±.12	.26±.14	.38±.17	.31±.14	.94±.06	.44±.15	.43±.13	.34±.17
+MASK	*.88±.05*	*.84±.06*	*.85±.03*	*.64±.10*	*.71±.11*	*.66±.07*	*.80±.02*	*.62±.03*	*.88±.08*	*.82±.08*	.84±.05	*.56±.19*	*.67±.16*	*.59±.15*	*.80±.04*	*.56±.09*
cross-encoder/stsb-TinyBERT-L-4	.32±.09	.98±.03	.48±.10	.96±.03	.16±.13	.26±.16	.33±.14	.41±.07	.29±.14	.97±.05	.43±.17	.77±.39	.13±.18	.19±.23	.28±.20	.34±.19
+MASK	*.67±.15*	*.66±.16*	*.63±.07*	*.86±.07*	*.85±.09*	*.85±.04*	*.79±.04*	*.62±.04*	*.61±.23*	*.62±.26*	*.54±.17*	*.87±.10*	*.85±.11*	*.85±.05*	*.79±.05*	*.56±.11*
cross-encoder/stsb-roberta-base	.31±.08	.98±.02	.47±.09	.95±.05	.13±.07	.22±.10	.30±.08	.42±.07	.28±.14	.97±.05	.41±.16	.91±.15	.10±.10	.16±.15	.26±.13	.33±.20
+MASK	*.68±.10*	*.64±.15*	*.64±.08*	*.86±.06*	*.87±.07*	*.86±.03*	*.80±.04*	*.63±.06*	*.57±.21*	*.57±.26*	*.52±.20*	*.86±.11*	*.86±.10*	*.85±.06*	*.78±.08*	*.57±.11*
msmarco-bert-base-dot-v5	.93±.03	.32±.10	.47±.11	.36±.08	.94±.03	.51±.09	.49±.09	.45±.09	.91±.07	.26±.19	.38±.21	.31±.14	.92±.08	.45±.16	.43±.15	.33±.24
+MASK	*.87±.05*	*.90±.05*	*.88±.03*	*.74±.11*	*.66±.09*	*.69±.06*	*.83±.03*	*.65±.03*	*.86±.09*	*.90±.06*	*.88±.05*	*.66±.18*	*.58±.20*	*.58±.17*	*.82±.05*	*.58±.11*
cross-encoder/ms-marco-TinyBERT-L-2-v2	.32±.08	.97±.02	.48±.09	.93±.06	.17±.11	.28±.14	.34±.12	.34±.10	.29±.14	.97±.03	.43±.16	.78±.30	.13±.19	.20±.23	.29±.19	.26±.20
+MASK	*.67±.15*	*.64±.14*	*.63±.07*	*.86±.06*	*.86±.09*	*.85±.04*	*.79±.03*	*.60±.06*	*.60±.23*	*.61±.24*	*.55±.17*	*.87±.10*	*.86±.12*	*.85±.05*	*.79±.06*	*.55±.15*
cross-encoder/ms-marco-MiniLM-L-2-v2	.32±.08	.97±.02	.48±.09	.94±.05	.16±.12	.26±.15	.33±.13	.36±.10	.29±.14	.97±.05	.43±.16	.91±.15	.13±.20	.19±.24	.29±.20	.26±.23
+MASK	*.67±.14*	*.61±.13*	*.62±.07*	*.85±.06*	*.87±.07*	*.85±.03*	*.79±.03*	*.57±.08*	*.58±.22*	*.55±.25*	*.50±.18*	*.85±.11*	*.85±.10*	*.84±.06*	*.77±.07*	*.51±.16*
cross-encoder/ms-marco-MiniLM-L-4-v2	.32±.08	.95±.03	.47±.09	.89±.04	.18±.10	.29±.13	.35±.11	.31±.10	.29±.14	.93±.09	.42±.17	.91±.10	.16±.16	.24±.20	.32±.17	.24±.22
+MASK	*.63±.13*	*.64±.13*	*.62±.07*	*.86±.06*	*.83±.08*	*.84±.04*	*.78±.04*	*.56±.07*	*.56±.21*	*.62±.25*	*.53±.16*	*.87±.11*	*.81±.14*	*.83±.07*	*.77±.06*	*.52±.15*
cross-encoder/quora-roberta-base	.31±.08	**.99±.02**	.46±.09	.96±.04	.10±.05	.18±.07	.27±.07	.32±.08	.28±.14	.98±.05	.41±.17	.78±.39	.09±.10	.15±.15	.25±.13	.23±.17
+MASK	*.63±.12*	*.55±.07*	*.58±.08*	*.83±.04*	*.87±.03*	*.85±.03*	*.78±.03*	*.47±.09*	*.58±.30*	*.47±.18*	*.49±.20*	*.84±.08*	*.88±.08*	*.85±.05*	*.79±.04*	*.41±.16*
cross-encoder/quora-roberta-large	.31±.08	.97±.05	.46±.09	.26±.40	.09±.15	.13±.21	.23±.17	.31±.10	.28±.14	.97±.06	.41±.17	.25±.39	.08±.19	.11±.23	.22±.21	.22±.19
+MASK	*.40±.20*	*.76±.37*	*.40±.10*	*.22±.34*	*.29±.44*	*.25±.38*	*.30±.26*	*.48±.08*	*.35±.25*	*.73±.41*	*.32±.20*	*.23±.36*	*.29±.44*	*.25±.39*	*.30±.28*	*.42±.14*
cross-encoder/ms-marco-MiniLM-L-6-v2	.33±.09	.91±.05	.48±.09	.87±.06	.25±.11	.37±.12	.41±.11	.30±.09	.29±.15	.88±.12	.42±.17	.89±.10	.23±.15	.34±.16	.39±.14	.21±.18
+MASK	*.63±.14*	*.62±.13*	*.60±.08*	*.85±.06*	*.84±.07*	*.84±.03*	*.78±.03*	*.55±.07*	*.55±.24*	*.57±.26*	*.49±.20*	*.86±.11*	*.83±.12*	*.83±.05*	*.77±.05*	*.50±.15*
cross-encoder/ms-marco-MiniLM-L-12-v2	.33±.09	.79±.07	.46±.09	.81±.07	.35±.09	.49±.09	.49±.07	.24±.09	.30±.17	.78±.17	.41±.18	.82±.16	.34±.15	.47±.16	.48±.12	.20±.16
+MASK	*.58±.11*	*.58±.13*	*.56±.05*	*.83±.06*	*.81±.09*	*.82±.05*	*.75±.04*	*.48±.05*	*.47±.19*	*.53±.25*	*.45±.18*	*.83±.11*	*.80±.12*	*.81±.08*	*.74±.06*	*.44±.09*
cross-encoder/quora-distilroberta-base	.31±.08	.97±.06	.46±.09	.18±.36	.08±.18	.11±.23	.22±.18	.25±.10	.28±.14	.98±.05	.42±.17	.19±.38	.07±.20	.09±.23	.21±.21	.16±.20
+MASK	*.39±.20*	*.84±.31*	*.43±.11*	*.16±.32*	*.20±.39*	*.18±.35*	*.27±.25*	*.34±.10*	*.34±.19*	*.81±.37*	*.36±.21*	*.17±.34*	*.20±.40*	*.18±.36*	*.28±.28*	*.30±.18*
cross-encoder/qnli-electra-base	.33±.10	.45±.12	.36±.08	.74±.08	.63±.12	**.67±.08**	.58±.07	.04±.11	.31±.18	.49±.18	.34±.16	.78±.14	.64±.14	**.68±.09**	**.60±.11**	.07±.18
+MASK	*.41±.12*	*.36±.12*	*.35±.07*	*.74±.08*	*.77±.14*	*.74±.08*	*.63±.08*	*.07±.08*	*.40±.23*	*.38±.19*	*.32±.11*	*.77±.14*	*.78±.16*	*.75±.10*	*.64±.12*	*.11±.14*
cross-encoder/qnli-distilroberta-base	.31±.09	.50±.18	.35±.10	.73±.07	.53±.18	.60±.11	.53±.08	.05±.06	.30±.18	.48±.19	.32±.14	.75±.13	.53±.19	.59±.14	.54±.12	.02±.10
+MASK	*.46±.24*	*.31±.15*	*.30±.11*	*.73±.07*	*.77±.16*	*.74±.08*	*.61±.06*	*.13±.09*	*.32±.27*	*.26±.15*	*.24±.13*	*.75±.12*	*.78±.19*	*.74±.11*	*.62±.10*	*.15±.13*

Table 6: **TRiC evaluation** using various SBERT models on Subtask 1 and Subtask 2. Results are presented for each model using pre-trained models and the +MASK setting (*italic*). For Subtask 1, precision (PR), recall (RE), and Weighted -F1 scores (F1) are reported for both label 0 (i.e., different topics) and label 1 (i.e., roughly identical topics). For Subtask 2, Spearman correlation (SP) is reported on the overall set of instances. The reported metrics include standard deviations (±) across the 10 Test splits for comparative analysis. The superior performance for each metric between pre-trained models is highlighted in **bold**. Results for both Test and OOV Test sets are provided for completeness.