Measuring Similarity of Opinion-bearing Sentences

Anonymous EMNLP submission

Abstract

For many NLP applications of online reviews, comparison of two opinion-bearing sentences is key. We argue that, while general purpose text similarity metrics have been applied for this purpose, there has been limited exploration of their applicability to opinion texts. We address this gap in the literature, studying: (1) how humans judge the similarity of pairs of opinion-bearing sentences; and, (2) the degree to which existing text similarity metrics, particularly embedding-based ones, correspond to human judgments. We crowdsourced annotations for opinion sentence pairs and our main findings are: (1) annotators tend to agree on whether or not opinion sentences are *similar* or *different*; and (2) embedding-based metrics capture human judgments of "opinion similarity" but not "opinion difference". Based on our analysis, we identify areas where the current metrics should be improved. We further propose to learn a similarity metric for opinion similarity via fine-tuning the Sentence-BERT sentence-embedding network based on review text and weak supervision by review ratings. Experiments show that our learned metric outperforms existing text similarity metrics and especially show significantly higher correlations with human annotations for differing opinions.

1 Introduction

Online reviews are an integral part of e-commerce platforms. Consumers utilize these reviews to make purchasing decisions, and businesses use this feedback to improve products or services. With the ever-growing number of reviews, NLP research has focused on methods to make sense of this vast data resource, including applications for opinion summarization (Suhara et al., 2020; Bražinskas et al., 2020b; Mukherjee et al., 2020; Chu and Liu, 2019; Angelidis and Lapata, 2018) and opinion search (Poddar et al., 2017).

A key characteristic of text in this domain is that it contains opinion-bearing sentences (hereafter, "opinion sentences"). As in preceding work (Pontiki et al., 2016), we view an opinion as having an aspect (e.g., the feature of a product or dimension of a service) and an appraisal (e.g., a positive or negative sentiment). In many applications, one needs to determine if two related opinion sentences are comparable in meaning. From an applied viewpoint, one might think of two opinions being comparable if they support the same recommendation, with respect to the relevant aspect. To compare two opinions, prior work has employed text similarity metrics, where cosine similarity based on TF-IDF (Angelidis and Lapata, 2018) or embedding representations (Suhara et al., 2020) is used to measure opinion sentence similarity.

We group existing text similarity metrics broadly into two types: lexical- and embedding-based approaches. The lexical-based approaches, such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), evaluate text by capturing the overlap in surface forms, such as n-grams of tokens. However, they are often ineffective when texts employ paraphrases or synonyms. Embedding-based metrics, such as Word Mover's Distance (WMD) (Kusner et al., 2015), MoverScore (Zhao et al., 2019), and Sentence-BERT (SBERT) (Reimers and Gurevych, 2019), typically relax the restriction of strict string matching by comparing continuous representations for words and sentences. Such approaches have shown to work well for various NLP applications, including areas involving comparison of sentence meaning; for example, paraphrase detection, question answering, or summarization (Wang et al., 2018; Lan and Xu, 2018; Suhara et al., 2020). However, a detailed study investigating their relationship to corresponding human judgments of opinion sentences is lacking.

In other text similarity settings, such as summary evaluation (Zhao et al., 2019), caption eval-

uation (Zhang et al., 2020) and machine translation evaluation (Mathur et al., 2019), embedding-based metrics out-perform lexical-based metrics, as demonstrated by its increased correlation with human judgment scores. However, embedding-based metrics have not yet been evaluated on opinion texts. The success of these embedding-based metrics in other types of text (such as news) cannot be guaranteed for opinion texts. This is because opinion text can be associated with a sentiment polarity. Opinion bearing words that are opposite in sentiment polarity are semantically related. Yet, many of these embedding-based metrics are often trained on semantic relatedness but not specifically sentiment polarity.

We address the gap of lacking research on similarity for opinion-bearing texts in the literature with the following research questions: (1) how do humans evaluate similarity of two opinion sentences? (2) how well do existing metrics capture similarity in a way similar to humans? and if not well, (3) how do we develop metrics to more effectively measure similarity for opinion sentences? We address the first question by conducting a crowdsourcing task that collects human annotations for the degree of similarity of two opinion sentences.[1] For the second research question, we examine the correlation of the text similarity metrics against our crowdsourced annotations. For the third question, we explore approaches to fine-tune embedding-based metrics for similarity of opinion texts.

We make several contributions: (1) we collect and release a dataset of 1635 sentence pairs with similarity scores; (2) we show that annotators broadly agree on whether an opinion sentence pair is "similar" or "different"; (3) we demonstrate that text similarity metrics have weak correlation to human judgments of opinion similarity, and that they perform poorly with differing opinions in particular; (4) we conduct an analysis of differing opinions to characterize the limitations of such approaches when dealing with opinion sentences; and, (5) we propose to learn a metric for similarity of opinion texts by fine-tuning SBERT via weak supervision by review ratings. Our experiments show that the fine-tuned SBERT model outperforms existing metrics for distinguishing different opinions and for measuring similarity of opinion sentences.

[1]We will release the data upon publication.

2 Related Work

Our research is related to text similarity metrics, which broadly include lexical-based metrics, embedding-based metrics and learned metrics.

Lexical-based Metrics ROUGE (Lin, 2004) is a commonly used metric for opinion summary evaluation. It measures similarity between texts by counting the overlaps of n-grams. BLEU (Papineni et al., 2002) is the default metric for machine translation evaluation. Similar to ROUGE, it also relies on counting overlaps in n-grams. Such lexical matching methods face the same limitation in evaluating texts that are similar in meaning but expressed with different words (Ng and Abrecht, 2015; Shimanaka et al., 2018). METEOR (Denkowski and Lavie, 2014) is proposed to relax the exact n-gram matching to allow matching words with its synonyms.

Embedding-based Metrics Embedding-based metrics are proposed to overcome the limitations of lexical-based metrics (Zhelezniak et al., 2019; Clark et al., 2019; Zhang et al., 2020). Word Mover's Distance (WMD) (Kusner et al., 2015) and MoverScore (Zhao et al., 2019) measure how similar two texts are by accumulating the distance between word embeddings and contextual embeddings, respectively. Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) is a sentence encoder that can be used with cosine similarity to capture similarity of meaning between sentences.

Metric Learning The objective of metric learning is to learn a task specific similarity measure. There are two broad approaches to metric learning. Supervised metric learning requires a training dataset for the task. For example, machine translation metrics learn to score machine translations against humans translations from previous machine translation datasets with human annotations (Shimanaka et al., 2018; Mathur et al., 2019). Sentence similarity can be learnt using Siamese network of sentence encoders with the Manhattan distance with a semantic relatedness dataset (Mueller and Thyagarajan, 2016). However, this approach to metric learning requires a labelled dataset which is not always available.

Alternatively, metric learning by weak supervision uses related data to guide the learning. During the training phase, the training dataset can be differ-

ent from the end task and even trained with a different training objective. To get the sentence similarity of a pair of sentences, a Siamese network is trained with Natural Language Inference datasets (SNLI and MNLI) using cross entropy loss (Reimers and Gurevych, 2019). To learn the thematic similarity of sentences, the metric is trained on a Triplet network using triplets of sentences from Wikipedia sections (Ein Dor et al., 2018).

To sum up our discussion, it is notable that the lexical-based metric ROUGE is still widely used in the literature for similarity of opinion texts in tasks like opinion summarisation evaluation (Amplayo and Lapata, 2020; Bražinskas et al., 2020a). Although ROUGE correlates well with human judgments at the system level but it performs poorly at the summary text level (Bhandari et al., 2020).

3 Human Comparisons of Opinion Sentences

3.1 Data and Annotations

Our dataset is based on that of the SemEval 2016 Task 5: "Aspect-Based Sentiment Analysis Subtask 1" (Pontiki et al., 2016). The SemEval datasets contain review sentences on laptops and restaurants in English. We selected two sentences from reviews on the same product or business to create a sentence pair, for which we collected human judgments of similarity. We constructed 1800 sentence pairs using sentences of at least 3 and at most 25 tokens.

To ensure that judgments were not trivially about different features of a product, we kept at least one aspect the same between sentences. In this way, annotations would depend on the expression of the appraisal.

Human judgments were collected using Amazon Mechanical Turk.[2] Only annotators with "Mechanical Turk Masters Qualification" were considered. Annotators were asked to rate the similarity in meaning of each pair of opinion sentences on a 5-level Likert scale, using methodology borrowed from the Semantic Textual Similarity task (STS) shared task (Cer et al., 2017). In our annotation task, the scale ranged from 0 ("completely different opinion") to 4 ("completely same opinion"), with the middle value, 2, indicating a partial match.

We processed the annotations based on three quality-based criteria: (1) Filter out annotators with low accuracy on quality control sentence pairs; (2) Identify and filter out anomalous annotators; and, (3) Require a minimum of three annotations per sentence pair after filtering out annotators.

[2] https://www.mturk.com/

Domain	Pairs	Alpha	Avg. #Annot.	Avg. Var
Laptop	621	0.541	3.504	0.483
Restaurant	1014	0.624	3.673	0.414
Total Pairs	1635			

Table 1: Statistics on our annotated dataset. Krippendorff's alpha, average number of annotations and average variance of annotations, per pair for each domain.

#Levels	Grouping	Laptop	Restaurant
2	(0,1) (2,3,4)	0.524	0.665
3	(0,1) (2) (3,4)	0.536	0.624
3	(0) (1,2,3) (4)	0.250	0.312

Table 2: Agreement for different Likert scales.

3.2 Analysis

Statistics of our dataset with annotations are shown in Table 1. After processing for quality, the dataset includes 1635 sentence pairs from reviews on two domains. The inter-annotator agreement is measured using Krippendorff's alpha (reliability coefficient) for ordinal levels (Artstein and Poesio, 2008), with coefficients of 0.541 and 0.624 for Laptop and Restaurant, respectively, indicating a moderate level of agreement.

We investigate the appropriateness of the 5-point Likert scale by comparing the inter-annotator agreement at different levels of Likert scale. Interestingly, while the 5-point scale had the highest level of agreement, a 3-point scale (grouping 0 and 1 together, and 3 and 4 together) led to a similar level of agreement, as shown in Table 2. Grouping the center three levels together led to worse

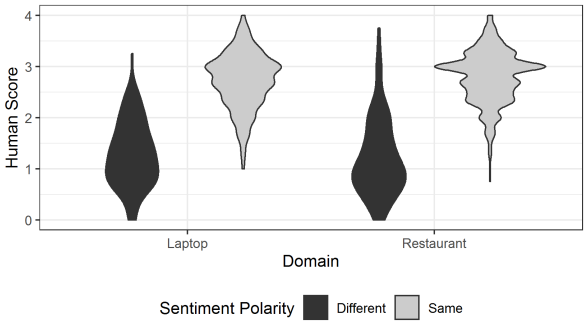

Figure 1: Violin plot of human score of sentence pair by domain and sentiment polarity match.

	Laptop		Restaurant	
Metric	**P**	**K**	**P**	**K**
ROUGE-1	0.078	0.035	0.086	0.068
ROUGE-2	0.102	0.079	0.042	0.027
ROUGE-L	0.070	0.025	0.085	0.070
SPICE	0.074	0.071	0.112	0.116
WMD	0.217	0.184	0.249	0.151
SBERT	**0.430**	**0.311**	**0.450**	**0.331**
MoverScore	0.229	0.172	0.206	0.156

Table 3: Correlation of existing embedding metrics and human scores. The highest correlation is in bold. P: Pearson and K: Kendall.

agreement. When grouping levels into 2 bins (0,1 vs 2,3,4), agreement increased for the Restaurant data ($\alpha = 0.665$) but decreased for the laptop data ($\alpha = 0.524$). We refer to the two-level groupings as *broadly different* and *broadly similar* opinions. We draw on this 2-level distinction, also with moderate annotator agreement, later in this paper. Given the moderate level of agreement achieved, we argue that the human judges generally agreed on these similarity judgments. This also suggests that the 5-point scale we collected our annotations on is an appropriate choice.

We explored this agreement further by examining the relationship of these judgments to sentiment polarity. Our selection of sentence pairs were sampled with constraints on aspect but were unconstrained by sentiment, which is already annotated in the SemEval dataset. We grouped sentences pairs with the same polarity and contrasting polarity and examined how humans judged similarity of opinions. We present these results in Figure 1. These violin plots of human scores of sentence pairs with "same" sentiment polarity is spread above level 2 and the violin plots of human scores of sentence pairs with "different" sentiment polarity is below level 2. This is consistent with what one would expect given an ordinal rating of similarity, and supports the use of a 5-point Likert scale.

4 On Metrics and Human Judgments

The following baseline metrics are chosen for our investigation: (1) ROUGE variants, ROUGE-1, ROUGE-2 and ROUGE-L without stemming and without stopword removal, using ROUGE-2.0 (Ganesan, 2018); (2) SPICE (Anderson et al., 2016), an image captioning evaluation metric that compares the scene graph of one text against the reference text; and, (3) WMD, using the implementation of WMD from Gensim (Řehůřek and Sojka, 2010) with the normalized 300-dimension word2vec trained on Google News. We follow Clark et al. (2019) to transform WMD scores to a similarity score using $exp^{-(WMD)}$; (4) SBERT, using sentence-transformers library in python; and (5) MoverScore, using the authors' implementation. These metrics are representative of the different types of metrics. ROUGE is a lexical-based metric, SPICE is a metric that incorporates representations of sentence meaning, and WMD, SBERT and MoverScore are embedding-based metrics.

	Laptop		Restaurant	
Metric	**P**	**K**	**P**	**K**
Broadly Different				
ROUGE-1	0.244	**0.180**	0.022	0.034
ROUGE-2	0.082	0.051	0.123	0.102
ROUGE-L	**0.257**	0.166	0.006	0.028
SPICE	−0.006	−0.026	0.014	0.032
WMD	0.038	0.060	**0.233**	**0.110**
SBERT	0.156	0.108	0.119	0.070
MoverScore	−0.033	−0.019	−0.148	−0.073
Broadly Similar				
ROUGE-1	0.053	0.022	0.137	0.076
ROUGE-2	0.104	0.069	0.073	0.036
ROUGE-L	0.058	0.022	0.152	0.088
SPICE	0.132	0.092	0.130	0.104
WMD	0.255	0.177	0.146	0.114
SBERT	**0.391**	**0.272**	**0.399**	**0.276**
MoverScore	0.330	0.230	0.323	0.192

Table 4: Correlation of metric and human scores for *broadly different* or *broadly similar* groups. The highest correlation is in bold. P: Pearson and K: Kendall.

Pearson and Kendall correlations are reported in Table 3. Pearson correlation is often used in the literature for text similarity evaluation. However, Pearson correlation can be misleading because it is a measure of linear relationship, sensitive to outliers and requires the two variables to be approximately normally distributed (Reimers et al., 2016). We therefore include the Kendall correlation, a non parametric correlation that is not limited to linear relationship, less sensitive to outliers and does not make any assumption about the distribution of variable. Amongst the baseline metrics, SBERT has the highest correlation but the correlation is still weak.

4.1 Broadly Different and Broadly Similar

We grouped the data using the binary split (*broadly different* and *broadly similar*) presented in Sec-

tion 3.2 and calculate the correlations again, presenting these in Table 4. We observe that correlations for *broadly different* are generally lower for baseline metrics (e.g., Pearson correlation ranging from -0.033 to 0.257 for Laptop) than comparable values for the *broadly similar* group. This suggests that the metrics tested have difficulty in determining difference in meaning of opinion sentences.

5 SOS: Sentence-BERT for Opinion Similarity

Our earlier analysis shows that existing embedding-based metrics have low correlation with human judgments for sentence pairs of opposite sentiment polarity. We hypothesize that the performance of embedding-based metrics can be improved with sentiment polarity information.

A straightforward approach to improving embedding-based metrics for opinion similarity is to use sentimentally trained word embeddings. WMD is based on the Word2Vec word embeddings trained on Google news. For opinion similarity, the sentiment specific word embeddings trained on tweets and the sentiment information associated with emoticons Tang et al. (2014) can be used. We call this baseline approach WMD-SSWE.

Motivated by the observation that BERT-based sentence embeddings has shown superior performance for measuring sentence similarity, we propose to learn a metric for opinion similarity based on SBERT. Our metric is a Siamese network of SBERT that takes in two sentences as inputs and outputs a similarity score. To overcome the problem of costly human annotated similarity score for training, we propose to train the metric through weak supervision based on review ratings. We call our model SOS (SBERT for Opinion Similarity).

Online reviews usually contain a review text and review rating. The review text contains opinion texts and the review rating provides an overall sentiment polarity of review text. For popular review platforms, the review rating typically spans a score of 1 to 5. A review rating of 1 is negative and 5 is positive. We can draw the connection that a review text associated with higher rating is positive. Similarly, a review text associated with lower rating is negative. In our work, we consider positive reviews to have a star rating of 4 and 5, while negative reviews to have a star rating of 1 and 2. We omit reviews of star ratings of 3. For the same product, review texts with the same sentiment polarity (both positive or both negative) are deemed to be similar and review texts with different sentiment polarity (one positive and one negative) are different. This forms the basis of creating the training datasets for fine-tuning our model.

We explore both Siamese networks and Triplet networks for training the opinion similarity model. The Siamese network for fine-tuning SOS, SOS^S, formulates opinion similarity as a classification task with a learning objective to classify a pair of sentences as similar or otherwise. The supervision is a binary signal that a pair is either similar or different. This approach is used by an unsupervised metric for a sentence similarity, where a Siamese network of SBERT is fine-tuned on SNLI and MNLI dataset which is a classification task Reimers and Gurevych (2019). For this work, we create training, development and test datasets of review pairs. Each pair contain reviews from the same product. The pair is either similar (either both positive or both negative) or different (one positive and one negative). We also ensure that the dataset is balanced with similar and different pairs. The training objective is cross entropy loss.

The second variant of our metric is to fine-tune with a triplet network for SOS, SOS^T. Each training instance is a triplet of an anchor example, positive example and negative example. The learning objective is triplet loss, which is to score the distance between anchor example and positive example to be smaller than the distance between anchor example and negative example by a margin. Each triplet is constructed from reviews for the same product. We randomly select a review to form the anchor, and randomly selected another review review that have the same sentiment polarity as the anchor example as the positive example. We then select another review with opposite sentiment polarity to the anchor example as the negative example. For this work, we create training, development and test datasets of review triplets.

Both Siamese and Triplet networks are effective for metric learning, but SOS^T performs better than SOS^S as the training triplets provide context that helps modeling the similarity more effectively. This finding is consistent with the literature for Semantic Text Similarity (Hoffer and Ailon, 2015).

6 Experiments and Results

We examine the performance of our model variants SOS^S and SOS^T on measuring similarity of sen-

tences. Specifically, we compare which training network, the Siamese or Triplet, is more appropriate to fine-tune our model for our task. Apart from comparing the networks, we also included variations in constructing the training pairs or triplets. We constructed pairs and triplets with the entire review text, first sentence of review text or random sentence of review text. We choose to include sentence variations because our task is at a sentence level therefore training examples at sentence level is an appropriate consideration.

For our experiments, we train four variants of SOS: (1) SOS-Siamese-PC (SOS^S_{PC}) and SOS-Triplet-PC (SOS^T_{PC})- trained with reviews from Amazon PC dataset[3]; and, (2) SOS-Siamese-Yelp (SOS^S_{Yelp}) and SOS-Triplet-Yelp (SOS^T_{Yelp})- trained with reviews from the Yelp Academic dataset[4]. We select these two review datasets with the intention to train our models on domain related reviews. The models on Amazon PC reviews is intended to roughly parallel the Laptop dataset and models on Yelp reviews roughly parallel the Restaurant dataset. Our experiments are written in python using the codes from sentence-transformers library. We use SBERT (stsb-bert-base) model in our metric, and fine-tuned with 10% warm up steps, 1 epoch and a batch size of 8. We ran each model three times and report the average correlation on our opinion similarity evaluation dataset.

We select our models based on the accuracy on the development datasets. For SOS^S models, SOS^S_{PC} and SOS^S_{Yelp} are both trained with 8000 training examples of entire reviews. Our best SOS^T models are SOS^T_{PC} which is fine-tuned with 1000 training triplets of entire reviews and margin of 1, and SOS^T_{Yelp} which is fine-tuned with 3000 training triplets of entire reviews and a margin of 7.

6.1 Main Results

Out of the models we propose, the metric learning models consistently outperform the best embedding-based model (SBERT) (Table 5). Our best model for Restaurant is SOS^T_{Yelp}. Although SOS^T_{Yelp} have the highest Pearson correlation for Laptop, its Kendall correlation is comparable to SOS^T_{PC}. This suggests that training on Yelp reviews can be generalized to both Laptop and Restaurant opinions. Our models outperform SBERT even when not fine-tuned in a relevant domain. On the other hand, WMD-SSWE have poor correlation with human judgments.

Comparing SOS^S and SOS^T models, triplet trained models achieve higher correlation than the models trained with pairs. A possible explanation is that triplets capture context information that is beneficial to evaluate sentence pairs that are broadly similar. This result is consistent with the observation by Hoffer and Ailon (2015).

For broadly different sentence pairs, almost all our models have a higher correlation than SBERT. This result supports our hypothesis that sentiment polarity information is helpful for opinion similarity and our proposed weak supervision methods (training by pairs and triplets) are effective to learn sentiment polarity information.

For broadly similar sentence pairs, SOS^T mod-

	Laptop		Restaurant	
Metric	**P**	**K**	**P**	**K**
SBERT	0.430	0.311	0.450	0.331
WMD-SSWE	0.128	0.099	0.079	0.188
SOS^S_{PC}	0.507	0.354	0.668	0.492
SOS^S_{Yelp}	0.515	0.367	0.747	0.535
SOS^T_{PC}	0.584	**0.427**	0.634	0.466
SOS^T_{Yelp}	**0.606**	0.425	**0.794**	**0.569**

Table 5: Correlation of our metrics and human scores. The highest correlation is in bold. SBERT is the best baseline metric. P: Pearson and K: Kendall.

	Laptop		Restaurant	
Metric	**P**	**K**	**P**	**K**
Broadly Different				
SBERT	0.156	0.108	0.119	0.070
WMD-SSWE	0.244	0.298	0.063	0.171
SOS^S_{PC}	0.262	0.176	0.252	**0.196**
SOS^S_{Yelp}	0.389	0.273	0.232	0.121
SOS^T_{PC}	0.249	0.181	0.268	0.184
SOS^T_{Yelp}	**0.396**	**0.299**	**0.275**	0.173
Broadly Similar				
SBERT	0.391	0.272	0.399	0.276
WMD-SSWE	0.141	0.059	0.105	0.104
SOS^S_{PC}	0.266	0.215	0.366	0.323
SOS^S_{Yelp}	0.284	0.236	0.454	0.365
SOS^T_{PC}	**0.478**	**0.346**	0.462	0.333
SOS^T_{Yelp}	0.399	0.305	**0.529**	**0.401**

Table 6: Correlation of our metrics and human scores for *broadly different* or *broadly similar* groups. The highest correlation is in bold. SBERT is the best baseline metric. P: Pearson and K: Kendall.

[3] https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt

[4] https://www.yelp.com/dataset

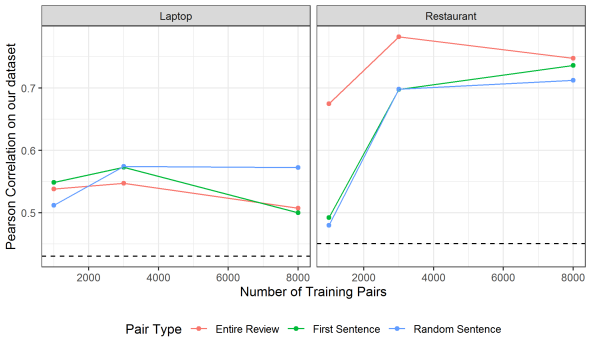

Figure 2: Pearson correlation of SOS^S models on our dataset. The dashed line is SBERT.

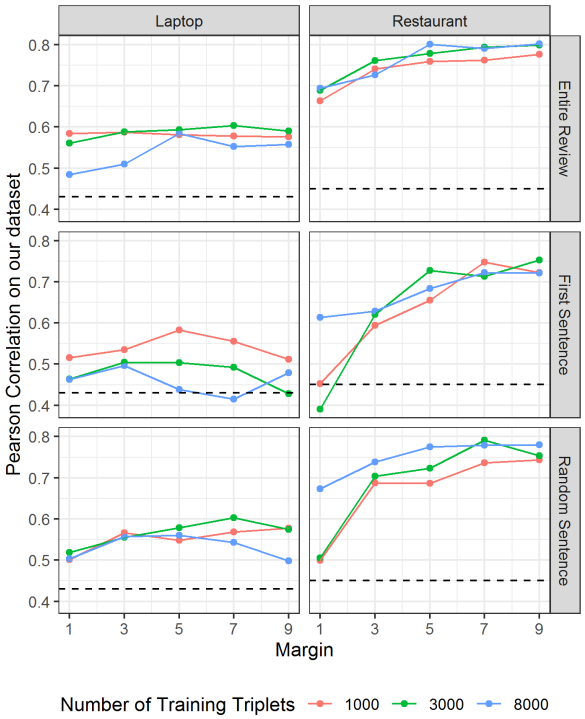

Figure 3: Pearson correlation of SOS^T models on our dataset. The dashed line is SBERT.

Model	Laptop	Restaurant
SBERT	0.430	0.450
Best	0.603	0.801
SOS^T	0.584	0.794

Table 7: Pearson correlation of our selected models and highest correlation amongst all models.

els consistently improve both Pearson and Kendall correlation over SBERT. On the other hand, our SOS^S models are not always better. This suggests that the triplet training is more appropriate for "Broadly Similar" pairs. Although the correlation on "Broadly Different" pairs have increased, it is still not as high as correlation on "Broadly Similar" pairs. We discuss this further in Section 7.

6.2 Granularity of Text

Granularity of the text in the training examples can potentially affect the performance of the metric (Ein Dor et al., 2018). We compare the performance of different models trained on training examples constructed by entire reviews, first sentence or random sentence.

We plot the Pearson correlation on the opinion similarity task of the SOS^S models in Figure 2 and SOS^T models in Figure 3. We present the results for Pearson correlation as Kendall correlation exhibits a similar trend.

Overall, the best models on our opinion similarity dataset are trained on examples of entire reviews except for SOS$^S_{PC}$ which is best with random sentence selection. We initially thought that the sentence examples will be more appropriate as our task is at a sentence level. However, our results show otherwise. One possible reason is that review text contains a mix of positive and negative opinions. Selecting the first sentence or a random sentence might not correspond the overall review rating resulting in a noisy training dataset which eventually reduced the effectiveness of training.

6.3 Optimizing on Development Dataset

We investigate the question of "How much do we fine-tune our model with weak supervision?". We select our best models based on the accuracy on the development set. However, is optimizing on the development dataset a good strategy to obtain the model that performs best on the end task?

We observe that optimizing the models on the development set does not give us the highest correlation on our opinion similarity task (Table 7). However, the performance of these models are not significantly differently (using two-sided Permutation Test for paired data at 5%) from the model with highest Pearson correlation amongst all models. Besides, all our selected models outperform the SBERT model. Hence, model selection based on the development dataset is a reasonable way to select our model for our task.

7 Error Analysis

We examined possible reasons why metrics have difficulty assessing differences in opinion. Table 9 shows how many pairs were deemed similar (in the

Sentence Pair	Human	SBERT	SOS^{T}_{PC}	SOS^{T}_{Yelp}	Explanation
Restaurant					
S1: Rice is too dry, tuna wasn't so fresh either. S2: Hands down, the best tuna I have ever had.	Q4	Q1	Q2	Q4	Opposite sentiment
S1: It was absolutely amazing. S2: This place is unbelievably over-rated.	Q4	Q1	Q3	Q4	Opposite sentiment and Implicit aspect (S1)
S1: Worst Service I Ever Had. S2: We waited over 30 minutes for our drinks and over 1 1/2 hours for our food.	Q1	Q4	Q4	Q3	Implied opinion (S2)
Laptop					
S1: You will not regret buying this computer! S2: I can't believe people like these computers.	Q4	Q1	Q2	Q4	Opposite sentiment
S1: This is very fast, high performance computer. S2: It wakes in less than a second when I open the lid.	Q1	Q4	Q3	Q3	Implicit aspect (S2) and Implied opinion (S2)

Table 8: Examples of sentence pairs where SBERT scores are inconsistent with human score. We expect the metric scores to be in similar quartiles of human scores. SOS^{T}_{PC} and SOS^{S}_{Yelp} are able to score sentence pairs of opposite sentiment more correctly but not sentence pairs of implicit aspect and implied opinion.

Metric	Laptop	Restaurant
Broadly Different Sentence Pairs	116	227
SPICE	0.069	0.062
WMD	0.181	0.172
SBERT	0.052	0.097
MoverScore	0.250	0.198
SOS^{T}_{PC}	0.000	0.018
SOS^{T}_{Yelp}	0.017	0.000
Broadly Similar Sentence Pairs	505	787
SPICE	0.000	0.000
WMD	0.028	0.044
SBERT	0.008	0.024
MoverScore	0.022	0.038
SOS^{T}_{PC}	0.004	0.013
SOS^{T}_{Yelp}	0.012	0.005

Table 9: Proportion of sentence pairs that are broadly different but scored in Q1 (Top 25%) and broadly similar but scored in Q4 (Bottom 25%) by metric scores.

top quartile (Q1)) when in fact the average human rating indicated they were *different*. For SBERT, the metric correlating best with human judgments (from Table 3), 5-10% of differing opinion pairs show human judgments diametrically oppose to metric scores. SOS variants reduce the percentage of wrongly scored pairs to almost 0%.

Table 8 presents examples for when automatic metrics are confused. Our analysis suggests three possible reasons for the weak correlation: sentence pairs that are opposite in sentiment polarity, implicit aspects, and implied opinions. To better understand the frequency of errors for our dataset, we sampled 100 sentence pairs from each domain and classified the challenges. Our annotations show that on average across both domains, 41% of the sentence pairs contain sentences that are opposite in polarity, 12% contain sentence pairs that contains implicit aspects and 10% contain implied opinions.

Although our SOS models have higher correlation than SBERT for "Broadly Different" pairs, correlations are still not at the same level for "Broadly Similar" pairs. However, SOS models are still not able to do well for opinions that contain implicit aspects or implied opinions. This is a possible reason for the lower correlation in "Broadly Different" pairs. We leave addressing these two challenges to future research.

8 Conclusions

We investigate how humans make similarity judgments over opinion sentences. We contribute a dataset of crowdsourced similarity judgments for opinion sentences. The agreement amongst annotators for judgments is moderate. We study the limitations of current text similarity methods when they are adopted for this task and our analysis show that this is likely due to the inability of current metrics to model *differing* opinions. By fine-tuning Siamese Sentence-BERT using weak supervision, we increase the Pearson correlation with human judgments to 0.606 and 0.794 on Laptop and Restaurant respectively of our opinion similarity dataset.

References

Reinald Kim Amplayo and Mirella Lapata. 2020. Unsupervised opinion summarization with noising and denoising. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1934–1945, Online.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision – ECCV 2016*, pages 382–398, Cham. Springer International Publishing.

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing Opinions: Aspect Extraction Meets Sentiment Prediction and They Are Both Weakly Supervised. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, Brussels, Belgium.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 9347–9359, Online.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020a. Few-shot learning for opinion summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 4119–4135, Online.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020b. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169, Online.

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 International Workshop on Semantic Evaluation*, pages 1–14, Vancouver, Canada.

Eric Chu and Peter Liu. 2019. MeanSum: A neural model for unsupervised multi-document abstractive summarization. In *Proceedings of the International Conference on Machine Learning*, pages 1223–1232, Long Beach, CA.

Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760, Florence, Italy.

Michael Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, MD.

Liat Ein Dor, Yosi Mass, Alon Halfon, Elad Venezian, Ilya Shnayderman, Ranit Aharonov, and Noam Slonim. 2018. Learning thematic similarity metric from article sections using triplet networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 49–54, Melbourne, Australia. Association for Computational Linguistics.

Kavita Ganesan. 2018. ROUGE 2.0: Updated and improved measures for evaluation of summarization tasks. *CoRR*, abs/1803.01937.

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

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the International Conference on International Conference on Machine Learning - Volume 37*, pages 957–966, Lille, France.

Wuwei Lan and Wei Xu. 2018. Neural network models for paraphrase identification, semantic textual similarity, natural language inference, and question answering. In *Proceedings of the International Conference on Computational Linguistics*, pages 3890–3902, Santa Fe, NM.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2019. Putting evaluation in context: Contextual embeddings improve machine translation evaluation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 2799–2808, Florence, Italy.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, page 2786–2792, Phoenix, AZ.

Rajdeep Mukherjee, Hari Chandana Peruri, Uppada Vishnu, Pawan Goyal, Sourangshu Bhattacharya, and Niloy Ganguly. 2020. Read what you need: Controllable aspect-based opinion summarization of tourist reviews. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1825–1828, Virtual Event, China.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for ROUGE. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930, Lisbon, Portugal.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA.

Lahari Poddar, Wynne Hsu, and Mong Li Lee. 2017. Author-aware aspect topic sentiment model to retrieve supporting opinions from reviews. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 472–481, Copenhagen, Denmark.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the International Workshop on Semantic Evaluation*, pages 19–30, San Diego, CA.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta.

Nils Reimers, Philip Beyer, and Iryna Gurevych. 2016. Task-oriented intrinsic evaluation of semantic textual similarity. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 87–96, Osaka, Japan.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 3982–3992, Hong Kong, China.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Conference on Machine Translation: Shared Task Papers*, pages 751–758, Belgium, Brussels.

Yoshihiko Suhara, Xiaolan Wang, Stefanos Angelidis, and Wang-Chiew Tan. 2020. OpinionDigest: A simple framework for opinion summarization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 5789–5798, Online.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1555–1565, Baltimore, MD.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, Online.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 563–578, Hong Kong, China.

Vitalii Zhelezniak, Aleksandar Savkov, April Shen, Francesco Moramarco, Jack Flann, and Nils Y. Hammerla. 2019. Don't settle for average, go for the max: Fuzzy sets and max-pooled word vectors. In *International Conference on Learning Representations*, New Orleans, LA.