DIVERSITY AND BIAS IN AUDIO CAPTIONING DATASETS

Irene Martín-Morató, Annamaria Mesaros

Computing Sciences
Tampere University, Finland
{irene.martinmorato, annamaria.mesaros}@tuni.fi

ABSTRACT

Describing soundscapes in sentences allows better understanding of the acoustic scene than a single label indicating the acoustic scene class or a set of audio tags indicating the sound events active in the audio clip. In addition, the richness of natural language allows a range of possible descriptions for the same acoustic scene. In this work, we address the diversity obtained when collecting descriptions of soundscapes using crowdsourcing. We study how much the collection of audio captions can be guided by the instructions given in the annotation task, by analysing the possible bias introduced by auxiliary information provided in the annotation process. Our study shows that even when hints are given with the audio content, different annotators describe the same soundscape using different vocabulary. In automatic captioning, hints provided as audio tags represent grounding textual information that facilitates guiding the captioning output towards specific concepts. We also release a new dataset of audio captions and audio tags produced by multiple annotators for a subset of the TAU Urban Acoustic Scenes 2019 dataset, suitable for studying guided captioning.

Index Terms— audio captioning, bias, lexical diversity.

1. INTRODUCTION

Audio captioning is defined as the general audio content description using free-text [1]. As a free-text description of the content in terms of sound events in a soundscape, it is an important step in understanding the dynamics of a sound scene. Most environmental sound datasets (e.g. AudioSet [2], FSD50K [3], TAU Urban Acoustic Scenes [4]) are annotated with one or multiple labels or tags, providing only basic information on the content, and lack information on more intricate relationships e.g., how sounds overlap or follow each other, and other specific attributes. On the other hand, audio captioning (manual or automatic) has the potential to provide rich descriptions of audio content for various needs.

Image captioning has been an active research area for long, and has established certain practices for data collection and for evaluation of automatic methods, that are currently adopted as such in audio captioning. Often, Amazon Mechanical Turk (MTurk) was used to collect large amount of annotated data. Image captioning datasets like PASCAL [5], or Flickr8k [6] also highlight the main problems of using MTurk to collect annotations, such as grammar and spelling mistakes or empty annotations. Nevertheless, MTurk remains the method of choice for efficient and fast data collection.

The amount of audio captioning datasets and related work in audio captioning is very small in comparison to the vast amounts of data and related scientific literature available for image and video captioning. The few existing datasets for audio captioning include AudioCaps [7], a large-scale dataset containing 50K audio files, most files having one human-written description, and Clotho [8], a dataset of 5K clips, each having five human-written descriptions.

We argue that data collection is always prone to bias, being affected by how the annotation task is presented and what kind of instructions, examples, and auxiliary information is provided to the annotator. Moreover, perception of sounds is affected by other co-occurring and overlapping sounds [9]. On one hand, this can lead to a diverse set of free-form descriptions, if the clips to be captioned contain many sounds, because different annotators may choose to describe different sounds. On the other hand, an observation from automatic image captioning is that models do not have the capability of taking into account user interest: when the image to be described is complex, the models produce global descriptions that try to balance the information from the perspective of readability and informativeness [10]. This has lead to studies of diversity of automatic image descriptions [11], and novel methods for guiding the captioning by using a guiding text that refers to either groundable or ungroundable concepts in the image [10].

In this work, we study how human-produced audio captions are affected by bias introduced through auxiliary information during the annotation process. We investigate the lexical diversity of three audio captioning datasets, to determine how the possible bias affects the vocabulary and similarity of the free-text descriptions provided to the same clip by different annotators. The main contributions of this paper are twofold. Firstly, we observe that human annotators can be guided towards describing target content in audio clips without explicit instructions, and without affecting the richness of the language used in the descriptions. Secondly, we release a new crowdsourced dataset of captioned acoustic scene clips and corresponding audio tags, together with the annotator competence estimated based on the tags [12]. The captions provide an extension to the TAU Urban Acoustic Scenes 2018 dataset, and allow using it for automated guided captioning based on the tags as grounding text, while the estimated annotator competence offers a measure of trust in the individual annotations.

The paper is organized as follows: Section 2 describes how the collection of free-form descriptions for acoustic scene audio clips was set up and post-processed. Section 3 explains how we measure the vocabulary bias, the lexical diversity and the similarity of the captions. Section 4 shows the results of the analysis; finally, Section 5 presents conclusions and future work.

2. DATASETS FOR AUDIO CAPTIONING

We collected captions for a subset of TAU Urban Acoustic Scenes 2019 [4], through a process designed such that human annotators

This paper has received funding from Academy of Finland grant 332063 "Teaching machines to listen".

were given hints on the audio content. The comparative analysis of three datasets allows understanding how the diversity and the vocabulary of the captions is influenced by the annotation setup.

2.1. MACS: Multi-Annotator Captioned Soundscapes

The data to be annotated consists of recordings from three acoustic scenes (airport, public square and park) of the TAU Urban Acoustic Scenes 2019 development dataset. A number of 3930 files were annotated, each file being 10-seconds long. The 133 annotators, students taking an audio signal processing course, were randomly assigned a maximum of 131 files each. Annotators were assigned into 30 groups, aiming that each group will provide annotations to the same set of files.

The annotation procedure used a web-based interface, and annotators were given examples of correct annotations before they started. The annotation process consisted in two tasks. The annotator was provided with a list of ten classes and an audio clip that could be played back multiple times, and was required to first select the sounds present in the audio clip from the given list. Afterwards, the annotator was required to write a free-form one sentence description of the clip, using a minimum of 5 words. The sound labels provided were: *birds singing, dog barking, adults talking, children voices, traffic noise, music, footsteps, siren, announcement speech* and *announcement jingle*. The instructions neutrally mentioned that using these sounds in the free-form description is fine. We hypothesize that by giving annotators a tagging task and a preselected list of sounds, we bring to their attention certain content, and therefore influence the produced caption without explicitly mentioning on what content to focus on. The produced captions were then processed by removing punctuation (!., :; ?()−), replacing symbols and numbers by their non-numerical form (e.g. "1″ to "one″, "+″ to "and″) and correcting minor grammar mistakes (using Ginger Software through *gingerit*).

We publish the complete dataset, which we call MACS[1], consisting of the captions and tags assigned by each annotator to each of the files, and the estimated competence for each annotator. Annotator competence is calculated using multi-annotator competence estimation (MACE) [13] as described in [12] as a measure of trustworthiness of the individual annotations.

2.2. Other audio captioning datasets

AudioCaps [7], is a collection of sentence-long descriptions for a subset of AudioSet [2], focused on the audio input. The video was provided to be played if necessary, and the AudioSet tags were presented to the annotator as hints. The dataset contains over 46k files of 10 seconds each, and one caption per file, collected using MTurk. We consider that the tags given as hints and the video, if played, introduced some bias to the content described by the captions. Clotho [1] was also collected using MTurk using a three-step framework composed of captioning, grammar correction, and rating of the captions [8]. It contains five captions per clip, for audio clips 15 to 30 seconds long that were collected from Freesound [14]. We consider this dataset as having no bias, since the captions are based solely on the audio clip provided, and no additional information regarding the possible active sounds or clip content was available to annotators.

MACS contains audio recorded in the wild, which compared to Clotho may have more complex acoustic content. Freesound samples are typically highly representative of the tagged sound and often contain only the indicated sound without much background [15, p.51]. On the other hand, the clips in MACS and AudioCaps may contain uncontrolled sequences or co-occurrence of multiple sounds, as they happened naturally in the recorded environment.

[1]MACS dataset: https://zenodo.org/record/5114771

whistling footsteps and adults talking	5 words
adults talking and someone whistling	5 words
adults talk and whistle outside	5 words
people talking followed by footsteps and whistling	7 words
adults chattering and whistling nearby	5 words
total number of words: 27; unique words: 14	**TTR = 0.51**

Table 1: TTR, which represents the ratio of unique words with respect to the total number of words

3. DIVERSITY, BIAS AND SIMILARITY

This section presents an overview of the metrics we use for assessing diversity and evaluating similarity of the captions. There is no clear consensus on metrics regarding similarity of text; however, we employ a few metrics inspired from machine translation, automatic captioning, and natural language processing, which are most often used to benchmark certain vocabulary characteristics.

3.1. Lexical diversity

One simple measure that represents the variety in vocabulary, or lexical diversity, is the type-token ratio (TTR). TTR is often used in measuring language acquisition in infants or learners of a second language, to assess if the learner uses the same words over and over, or uses a variety of different words to communicate [16].

TTR is defined as the number of distinct words (tokens), divided by the total number of words. Therefore, it ranges from a theoretical 0 (infinite repetition of a single word) and 1 (no repetition at all). In practice, the value is influenced by the length of the analyzed text: the longer the analyzed text, the lower the calculated TTR, because of using more of the same words. Moving-Average-TTR (MATTR) [17] was proposed to remove text length dependency; however it is dependent on the window length, being equivalent to calculating TTR for a smaller fixed window size. An example for calculating TTR is presented in Table 1, using five descriptions assigned to the same audio file. We use TTR to have a simple understanding of the use of different words in the datasets under study.

3.2. Vocabulary bias

We propose to measure the vocabulary bias as the proportion of hinted sounds with respect to the number of sounds mentioned in the caption. For identifying sounds in the caption, we use the AudioSet taxonomy, consisting of approximately 600 classes, considering that it provides a comprehensive list of the most common sounds encountered in our everyday environments.

We analyze only AudioCaps and MACS for bias, because they were provided with hints during the annotation process. For AudioCaps, the hints are the tags associated to the clip in AudioSet (possibly incorrect). For MACS, the hints are the ten tags among which the annotator was asked to mark the sounds present in the clip. For example, "whistling footsteps and adults talking" contains three sounds (whistling, footsteps and talking) of which two (foot-

Captions	Jaccard	BLEU-4	BERTscore	sBERT
a person whistling and singing *people are talking and singing*	0.38	0.00	0.92	0.91

Table 2: Example of similarity metrics calculated for two captions of the same audio clip.

steps, talking) were given as hint, therefore the bias is 0.66, while for "adults talking and someone whistling" the bias is 0.5.

3.3. Similarity

Automatic captioning methods are evaluated using metrics from machine translation, to compare the machine-generated captions with human produced free descriptions of the same items. In this study, we are interested to evaluate the similarity of descriptions produced by different annotators for the same audio example. One basic approach to calculate similarity is the Jaccard similarity coefficient, or intersection-over-union, for two sets that are compared. For two sentences a and b, Jaccard index is defined as

$$J(a,b) = \frac{|S_a \cap S_b|}{|S_a \cup S_b|}, \tag{1}$$

where $S_{a/b}$ is the tokenized version of the sentence. $J(a,b) = 0$ means that sentences a and b do not have any token in common, while $J(a,b) = 1$ means that they contain the exact same tokens. Jaccard index is a fast low-cost metric for measuring similarity [8].

BLEU [18] is a commonly-used metric for comparing machine translated text to human-translated references. It does so by calculating the overlap between n-grams from the reference and candidate sentences. BLEU is defined as the geometric mean of the n-gram precision up to a certain length of n:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \tag{2}$$

where p_n is the modified n-gram precision multiplied by positive weights w_n, and BP is a brevity penalty applied when the generated text is too short. Most commonly reported is BLEU-4 (or cumulative 4-gram BLEU score), that incorporates 1-, 2-, 3-, and 4-grams, with a weight of 0.25 each. Because it measures overlap of n-grams, BLEU cannot handle synonyms and paraphrasing. Despite this, it is the most widely used automatic evaluation score in machine translation, and commonly reported in automatic captioning. Recent methods for calculating similarity in natural language processing use BERT [19], a model pretrained on large amounts of unlabeled data that can be fine-tuned with smaller amounts of labeled data. Building on BERT, BERTScore [20] calculates contextual embeddings to represent the tokens and computes matching using cosine similarity, optionally weighted with inverse document frequency scores. The BERT contextual embeddings can handle paraphrasing and different ordering, capturing distant dependencies in sentences. A slightly different approach is given by sentence-BERT (sBERT) [21], a modification of the BERT model using siamese networks [22]; s-BERT encodes an entire sentence into an embedding, instead of going token by token, then uses the cosine measure between the embedding vectors of two sentences.

The three selected measures represent similarity at different granularity: Jaccard index treats the tokens as a set, disregarding the order of words; BLEU looks at n-gram overlaps, therefore very specifically focuses on the ordering of words, while BERTscore and sBERT are state-of-the-art similarity measures that give a holistic view of the semantic content. Table 2 presents an example of metrics values for two captions corresponding to the same audio file in MACS. The scores produced by the BERT-based models (0.91 and 0.92) reflect the fact that there is a high similarity in the content of the two sentences; the Jaccard score of 0.38 shows the proportion of identical words within the vocabulary, while BLEU-4 has difficulties in matching n-grams, returning a score of 0.0.

Dataset	Audio clips	Vocab. size	Unique sentences	Sentence length (std)
AudioCaps	57188	5218	52198	9.17 (4.27)
Clotho	5929	4373	29611	11.34 (2.78)
MACS	3930	2775	16262	9.46 (3.89)

Table 3: Statistics of the studied datasets.

4. EXPERIMENTAL RESULTS AND ANALYSIS

The statistics of the three studied datasets are presented in Table 3, with the vocabulary calculated without lemmatization. Note that, these numbers correspond to the current version of the downloaded datasets, and may differ from the ones reported in the original paper. The most used words in the MACS dataset (after lemmatization and stop word removal) are *talk*, *people* and *adult*, *noise*, and *bird*, of which the first three are parts of the provided tags. *Speak* is used in all its forms, but in a considerable less amount, since the given tag was *adults talking*. In AudioCaps the most used five words are: *man*, *speak*, *follow*, *talk*, and *engine*. In contrast, in Clotho the clips were selected specifically to not include speech [1], and the most used 5 words are: *bird*, *water*, *background*, *chirp*, and *someone*.

4.1. Lexical diversity

Lexical diversity is calculated in three different versions: (1) without any processing of the text; (2) with lemmatization; and (3) with lemmatization and removal of stopwords. Overall lexical diversity, calculated as TTR using all captions in each dataset, is presented in Table 4. TTR is lower when lemmatization is performed than without any processing because lemmatization merges some forms into the same unique word, decreasing the number of types. When stopwords removal is added, TTR is slightly higher because a significant amount of repetitive words is removed from the overall text. The overall lexical diversity is very low for all datasets, implying that, for all of them, a small set of words is used repeatedly to describe the audio. While the vocabulary of AudioCaps is larger than the other datasets, the total amount of text in it is also larger, resulting in a small TTR. If MATTR is calculated instead, overall diversity values increase when using a small window. AudioCaps has the highest diversity when MATTR is calculated using a relatively small window (10-1000 tokens), while for larger windows (5000-10000), Clotho is more diverse. In all cases, MACS has a smaller MATTR diversity, showing a high repetition of the vocabulary.

We also calculate local lexical diversity, i.e., TTR for the set of

S	L	AudioCaps	Clotho		MACS	
		overall	overall	local	overall	local
-	-	1.09%	1.30%	56.52%	1.80%	69.37%
-	✓	0.79%	0.91%	52.08%	1.38%	66.06%
✓	✓	1.27%	1.66%	60.43%	2.17%	71.02%

Table 4: Global and local lexical diversity of captions. S: removal of stopwords; L: lemmatization. AudioCaps has only a single caption per clip, thus we do not calculate local lexical diversity for it.

	Tag bias (std)	Word bias (std)
AudioCaps	0.33 (0.35)	0.35 (0.35)
MACS	0.38 (0.36)	0.49 (0.38)

Table 5: Calculated vocabulary bias.

descriptions assigned to the same clip. Here the types and tokens are counted based on the 5 captions of each clip, and the resulting clip-wise TTR values are averaged over the dataset. The results are presented in Table 4 for the datasets with multiple captions per clip. The comparison of local lexical diversity between Clotho and MACS shows that while Clotho has a larger vocabulary and slightly larger overall lexical diversity, MACS has a higher proportion of different words used to describe individual clips. The reason for this could be the source of audio clips: even though diverse in terms of sound categories, many Freesound clips often contain only the indicated sound, while the clips in MACS, being recorded in the wild, allow description of different details and sounds.

4.2. Vocabulary bias

We identify sound events present in the captions using the AudioSet vocabulary. We have merged our tags into the AudioSet vocabulary to deal with synonyms, e.g we added "talk" and "adult talk" to the vocabulary, because Audioset contains only the synonym "speech". We use a total of 722 labels to identify sounds in the captions. Table 5 shows the calculated bias for the two datasets with given hints. We also calculate word bias to account for the hints that do not match exact categories in AudioSet. About one third of the sounds mentioned in the captions are found in the given hints for both AudioCaps and MACS. On the other hand, for individual words, MACS has a much higher bias. Considering that we added our tags to the vocabulary, and that "adults talking" was the most frequently annotated tag in MACS, this confirms that the choice of words in the free-text description is influenced by the given hints. In addition, for MACS, the ratio of sounds selected by the annotators as tags but not mentioned in the caption is 29%. This means that, on average, over one fourth of the tags indicating sounds being active in a clip were not included in the free-form description. This can be explained by the complexity of the scenes, for which the caption is only a partial description of a complex acoustic content. The calculated bias for the guided annotation tasks is not considerably high, and it is interesting to note the tags missing from the caption. We hypothesize that the complexity of the acoustic scene can affect the diversity of the vocabulary more than it affects the observed bias. Indeed, a closer look at scene-wise lexical diversity shows that *airport* class has a local lexical diversity approximately 3 percent points higher than the park and street scenes, indicating that airport clips have a higher scene complexity than the other scenes in terms of events happening.

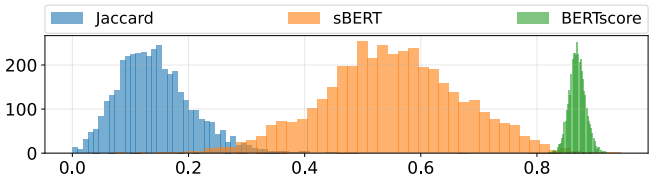

Figure 1: Similarity metrics for MACS dataset.

Dataset	BLEU-4	Jaccard	sBERT	BERTscore
Clotho	0.06 (0.04)	0.22 (0.09)	0.61 (0.13)	0.88 (0.01)
MACS	0.01 (0.02)	0.16 (0.08)	0.55 (0.12)	0.87 (0.01)

Table 6: Average similarity of the captions, using multiple metrics.

4.3. Similarity

We calculate similarity of the captions produced by different annotators for the same audio clip. The metrics are calculated for every pair of captions (10 pairs for a clip with 5 captions), and then averaged. Even though BLEU is generally meant to be used at corpus level, we use it at sentence level for comparison with the other metrics. The calculated values are presented in Table 6, and histogram plots of the clip-wise values for MACS are presented in Fig. 1.

We observe that BLEU provides very low similarity values, which implies diversity at least through ordering or paraphrasing. BLEU is higher for Clotho, and so is the Jaccard similarity index, indicating that descriptions of the same clip have more words in common for the captions in Clotho. This is in agreement with previously calculated local diversity that indicates MACS has more distinct words per clip. On the other hand, metrics based on BERT embeddings indicate high similarity for the descriptions of the same content. While sBERT is higher for Clotho, BERTscore is equally high. This effect may be due to the highly variable caption lengths in MACS, as sBERT groups sentences of same length for reducing computational load, and pads them to the longest one in each batch; according to the sentence length variance, this padding takes place more often in MACS than in Clotho. Both Clotho and MACS exhibit a high degree of caption similarity at clip level, irrespective of the difference in the characteristics of their audio content. On the other hand, the datasets do not have the same degree of diversity in terms of language used, showing its dependence on the nature of the complexity of the acoustic content.

5. CONCLUSIONS

This paper presented a study of the lexical diversity, bias, and similarity of captions from three audio captioning datasets. A new set of captions was collected for everyday soundscapes, with provided sound event hints. However, these hints turned out to not be a significant source of bias; instead, the free-text descriptions are more affected by the complexity of the soundscape. Despite the hints, the captions in the studied datasets have a high lexical diversity, and while token and n-gram based similarities are relatively low, the semantic similarity between captions assigned to the same clips by different annotators was found to be high. The new captions are freely available, along with the tags provided by the same annotators. This dataset brings novel elements to audio captioning; for example the tag-caption pairs allow guided captioning, and the estimated annotator reliability provides a measure of trustworthiness for each caption, which can be used in the learning process.

6. REFERENCES

[1] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: an audio captioning dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 736–740.

[2] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[3] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: an open dataset of human-labeled sound events," *ArXiv*, vol. abs/2010.00475, 2020.

[4] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 9–13.

[5] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using Amazon's Mechanical Turk," in *Proceedings of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Los Angeles: Association for Computational Linguistics, Jun. 2010, pp. 139–147.

[6] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *J. Artif. Int. Res.*, vol. 47, no. 1, p. 853–899, May 2013.

[7] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating captions for audios in the wild," in *Proceedings of the 2019 Conference of the NAACL HLT, Vol. 1*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 119–132.

[8] S. Lipping, K. Drossos, and T. Virtanen, "Crowdsourcing a dataset of audio captions," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, Oct. 2019.

[9] N. J. Vanderveeer, "Ecological acoustics: Human perception of environmental sounds," 1979.

[10] E. G. Ng, B. Pang, P. Sharma, and R. Soricut, "Understanding guided image captioning performance across domains," 2020, arXiv:2012.02339.

[11] E. van Miltenburg, D. Elliott, and P. Vossen, "Measuring the diversity of automatic image descriptions," in *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1730–1741.

[12] I. Martín-Morató and A. Mesaros, "What is the ground truth? reliability of multi-annotator data for audio tagging," in *29th European Signal Processing Conference 2021 (EUSIPCO 2021)*, Dublin, Ireland, Aug 2021.

[13] D. Hovy, T. Berg-Kirkpatrick, A. Vaswani, and E. Hovy, "Learning whom to trust with MACE," in *Proceedings of the 2013 Conference of the NAACL HLT*. Atlanta, Georgia: Association for Computational Linguistics, Jun. 2013, pp. 1120–1130.

[14] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *ACM International Conference on Multimedia (MM'13)*, ACM. Barcelona, Spain: ACM, Oct. 2013, pp. 411–412.

[15] N. Turpault, "Analysis of the scientific challenges in ambient sound recognition in real environment," Ph.D. dissertation, Universite de Lorraine, Villers-lès-Nancy, France, May 2021.

[16] G. Youmans, "Measuring lexical style and competence: The type-token vocabulary curve," *Style*, vol. 24, no. 4, pp. 584–599, 1990.

[17] M. A. Covington and J. D. McFall, "Cutting the gordian knot: The moving-average type–token ratio (mattr)," *Journal of Quantitative Linguistics*, vol. 17, no. 2, pp. 94–100, 2010.

[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. USA: Association for Computational Linguistics, 2002, p. 311–318.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the NAACL HLT, Vol. 1*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[20] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in *International Conference on Learning Representations*, 2020.

[21] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese bert-networks," in *EMNLP/IJCNLP (1)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Association for Computational Linguistics, 2019, pp. 3980–3990.

[22] T. Ranasinghe, C. Orasan, and R. Mitkov, "Semantic textual similarity with Siamese neural networks," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. Varna, Bulgaria: INCOMA Ltd., Sep. 2019, pp. 1004–1011.