Exploiting Twitter as Source of Large Corpora of Weakly Similar Pairs for Semantic Sentence Embeddings

Marco Di Giovanni
Politecnico di Milano
Università di Bologna
marco.digiovanni@polimi.it

Marco Brambilla
Politecnico di Milano
marco.brambilla@polimi.it

Abstract

Semantic sentence embeddings are usually supervisedly built minimizing distances between pairs of embeddings of sentences labelled as semantically similar by annotators. Since big labelled datasets are rare, in particular for non-English languages, and expensive, recent studies focus on unsupervised approaches that require not-paired input sentences. We instead propose a language-independent approach to build large datasets of pairs of informal texts weakly similar, without manual human effort, exploiting Twitter's intrinsic powerful signals of relatedness: replies and quotes of tweets. We use the collected pairs to train a Transformer model with triplet-like structures, and we test the generated embeddings on Twitter NLP similarity tasks (PIT and TURL) and STSb. We also introduce four new sentence ranking evaluation benchmarks of informal texts, carefully extracted from the initial collections of tweets, proving not only that our best model learns classical Semantic Textual Similarity, but also excels on tasks where pairs of sentences are not exact paraphrases. Ablation studies reveal how increasing the corpus size influences positively the results, even at 2M samples, suggesting that bigger collections of Tweets still do not contain redundant information about semantic similarities.[1]

1 Introduction and Related Work

Word-level embeddings techniques compute fixed-size vectors encoding semantics of words (Mikolov et al., 2013; Pennington et al., 2014), usually unsupervisedly trained from large textual corpora. It has always been more challenging to build high-quality sentences-level embeddings.

Currently, best sentence-embeddings approaches are supervisedly trained using large labeled datasets (Conneau et al., 2017; Cer et al., 2018; Reimers and Gurevych, 2019; Chen et al., 2019; Du et al., 2021; Wieting et al., 2020; Huang et al., 2021), such as NLI datasets (Bowman et al., 2015; Williams et al., 2018) or paraphrase corpora (Dolan and Brockett, 2005). Round-trip translation has been also exploited, where semantically similar pairs of sentences are generated translating the non-English side of NMT pairs, as in ParaNMT (Wieting and Gimpel, 2018) and Opusparcus (Creutz, 2018). However, large labeled datasets are rare and hard to collect, especially for non-English languages, due to the cost of manual labels, and there exist no convincing argument for why datasets from these tasks are preferred over other datasets (Carlsson et al., 2021), even if their effectiveness on STS tasks is largely empirically tested.

Therefore, recent works focus on unsupervised approaches (Li et al., 2020; Carlsson et al., 2021; Wang et al., 2021; Giorgi et al., 2020; Logeswaran and Lee, 2018), where unlabeled datasets are exploited to increase the performance of models. These works use classical formal corpora such as OpenWebText (Gokaslan and Cohen, 2019), English Wikipedia, obtained through Wikiextractor (Attardi, 2015), or target datasets without labels, such as the previously mentioned NLI corpora.

Instead, we propose a Twitter-based approach to collect large amounts of *weak* parallel data: the obtained couples are not exact paraphrases like previously listed datasets, yet they encode an intrinsic powerful signal of relatedness. We test pairs of quote and quoted tweets, pairs of tweet and reply, pairs of co-quotes and pairs of co-replies. We hypothesize that quote and reply relationships are weak but useful links that can be exploited to supervisedly train a model generating high-quality sentence embeddings. This approach does not require manual annotation of texts and it can be expanded to other languages spoken on Twitter.

We train models using triplet-like structures on the collected datasets and we evaluate the results on the standard STS benchmark (Cer et al., 2017), two

[1] Code available at https://github.com/marco-digio/Twitter4SSE

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 9902–9910
November 7–11, 2021.

Twitter NLP datasets (Xu et al., 2015; Lan et al., 2017) and four novel benchmarks.

Our contributions are four-fold: we design an language-independent approach to collect big corpora of weak parallel data from Twitter; we fine-tune Transformer based models with triplet-like structures; we test the models on semantic similarity tasks, including four novel benchmarks; we perform ablation on training dataset, loss function, pre-trained initialization, corpus size and batch size.

2 Datasets

We download the general Twitter Stream collected by the Archive Team Twitter[2]. We select English[3] tweets posted in November and December 2020, the two most recent complete months up to now. They amount to about $27G$ of compressed data ($\sim 75M$ tweets).[4] This temporal selection could introduce biases in the trained models since conversations on Twitter are highly related to daily events. We leave as future work the quantification and investigation of possible biases connected to the width of the temporal window, but we expect that a bigger window corresponds to a lower bias, thus a better overall performance.

We collect four training datasets: the Quote Dataset, the Reply Dataset, the Co-quote Dataset and the Co-reply Dataset.

The **Quote Dataset (Qt)** is the collection of all pairs of quotes and quoted tweets. A user can *quote* a tweet by sharing it with a new comment (without the new comment, it is called *retweet*). A user can also retweet a quote, but it cannot quote a retweet, thus a quote refers to an original tweet, a quote, or a reply. We generate *positive* pairs of texts coupling the quoted texts with their quotes.

The **Reply Dataset (Rp)** is the collection of all couples of replies and replied tweets. A user can reply to a tweet by posting a public comment under the tweet. A user can reply to tweets, quotes and other replies. It can retweet a reply, but it cannot reply to a retweet, as this will be automatically considered a reply to the original retweeted tweet. We generate *positive* pairs of texts coupling tweets with their replies.

The **Co-quote Dataset (CoQt)** and **Co-reply Dataset (CoRp)** are generated respectively from the Qt Dataset and the Rp Dataset, selecting as *positive* pairs two quotes/replies of the same tweet.

To avoid *popularity-bias* we collect only one positive pair for each quoted/replied tweet in every dataset, otherwise viral tweets would have been over-represented in the corpora.

We clean tweets by lowercasing the text, removing URLs and mentions, standardizing spaces and removing tweets shorter than 20 characters to minimize generic texts (e.g., variations of "Congrats" are common replies, thus they can be usually associated to multiple original tweets). We randomly sample $250k$ positive pairs to train the models for each experiment, unless specified differently, to fairly compare the performances (in § 5 we investigate how the corpus size influences the results). We also train a model on the combination of all datasets (**all**), thus $1M$ text pairs.

We show examples of pairs of texts from the four datasets in the Appendix.

3 Approach

We select triplet-like approaches to train a Tranformer model on our datasets. We extensively implement our models and experiments using sentence-transformers python library[5] and Huggingface (Wolf et al., 2020). Although the approach is model-independent, we select four Transfomer models (Vaswani et al., 2017) as pre-trained initializations, currently being the most promising technique ($\sim 110M$ parameters):

RoBERTa base (Liu et al., 2019) is an improved pre-training of BERT-base architecture (Devlin et al., 2019), to which we add a pooling operation: MEAN of tokens of last layer. Preliminary experiments of pooling operations, such as MAX and [CLS] token, obtained worse results;

BERTweet base (Nguyen et al., 2020) is a BERT-base model pre-trained using the same approach as RoBERTa on $850M$ English Tweets, outperforming previous SOTA on Tweet NLP tasks, to which we add a pooling operation: MEAN of tokens of last layer;

Sentence BERT (Reimers and Gurevych, 2019) are BERT-base models trained with siamese or triplet approaches on NLI and STS data. We select two suggested base models from the full list of

[2] https://archive.org/details/twitterstream

[3] *English* tweets have been filtered accordingly to the "lang" field provided by Twitter.

[4] We do not use the official Twitter API because it does not not guarantee a reproducible collections (Tweets and accounts are continuously removed or hidden due to Twitter policy or users' privacy settings).

[5] https://github.com/UKPLab/sentence-transformers

trained models: bert-base-nli-stsb-mean-tokens (S-BERT) and stsb-roberta-base (S-RoBERTa).

We test the two following loss functions:
Triplet Loss (TLoss): given three texts (an anchor a_i, a positive text p_i and a negative text n_i), we compute the text embeddings (s_a, s_p, s_n) with the same model and we minimize the following loss function:

$$max(||s_a - s_p|| - ||s_a - s_n|| + \epsilon, 0)$$

For each pair of anchor and positive, we select a *negative* text randomly picking a positive text of a different anchor (e.g., about the Quote dataset, anchors are quoted tweets, positive texts are quotes and the negative texts are quotes of different quoted tweets);
Multiple Negative Loss (MNLoss) (Henderson et al., 2017): given a batch of positive pairs $(a_1, p_1), ..., (a_n, p_n)$, we assume that (a_i, p_j) is a negative pair for $i \neq j$ (e.g., Quote Dataset: we assume that quotes cannot refer to any different quoted tweet). We minimize the negative log-likelihood for softmax normalized scores. We expect the performance to increase with increasing batch sizes, thus we set $n = 50$, being the highest that fits in memory (see § 5 for more details).

We train the models for 1 epoch[6] with AdamW optimizer, learning rate 2×10^{-5}, linear scheduler with 10% warmup steps on a single NVIDIA Tesla P100. Training on $250k$ pairs of texts requires about 1 hour, on $1M$ about 5 hours.

4 Evaluation

We evaluate the trained models on seven heterogeneous semantic textual similarity (STS) tasks: four novel benchmarks from Twitter, two well-known Twitter benchmarks and one classical STS task. We planned to test the models also on Twitter-based classification tasks, e.g., Tweeteval (Barbieri et al., 2020). However, the embeddings obtained from our approach are *not* designed to transfer learning to other tasks, but they should mainly succeed on similarity tasks. A complete and detailed evaluation of our models on classification tasks is also not straightforward, since a classifier must be selected and trained on the top of our models, introducing further complexity to the study. We leave this analysis for future works.

4.1 Novel Twitter benchmarks

We propose four novel benchmarks from the previously collected data. Tweets in these datasets are discarded from *every* training set to avoid unfair comparisons. We frame these as ranking tasks and we pick normalized Discounted Cumulative Gain (nDCG) as metric (Järvelin and Kekäläinen, 2002)[7]. We propose these datasets to highlight that benchmark approaches are not able to detect similarities between related tweets, while they can easily detect similarities between formal and accurately selected texts. Thus the necessity for our new models.

Direct Quotes/Replies (DQ/DR): Collections of $5k$ query tweets, each one paired with 5 positive candidates (quotes/replies of the query tweets) and 25 negative candidates (quotes/replies of other tweets). We rank candidates by cosine distance between their embeddings and the embedding of the query tweet.

Co-Quote/Reply (CQ/CR): Similar to the previous tasks, we focus on co-quotes/co-replies, i.e., pairs of quotes/replies of the same tweet. These datasets are collections of $5k$ query quotes/replies, each one paired with 5 positive candidates (quotes/replies of the same tweet) and 25 negative candidates (quotes/replies of other tweets). We rank candidates by cosine distance between their embeddings and the embedding of the query tweet.

4.2 Established benchmarks

We select two benchmarks from Twitter, PIT dataset and Twitter URL dataset (TURL), and the STS benchmark of formal texts. We pick Pearson correlation coefficient (Pearson's r) as metric.

PIT-2015 dataset (Xu et al., 2015) is a Paraphrase Identification (PI) and Semantic Textual Similarity (SS) task for the Twitter data. It consists in 18762 sentence pairs annotated with a graded score between 0 (no relation) and 5 (semantic equivalence). We test the models on SS task.

Twitter URL dataset (Lan et al., 2017) is the largest human-labeled paraphrase corpus of 51524

[6] We briefly tested the training for two epochs in preliminary experiments, but we noticed no evident benefits. Moreover, increasing the number of epochs enhances the risk of overfitting the noise included in tweets since these texts are noisy and we do not perform validation.

[7] nDCG is a common ranking-quality metric obtained normalizing Discounted Cumulative Gain (DCG). The scores range from 0 to 1, the higher the better. Thus, 1 represents a perfect ranking: the first ranked document is the most relevant one, the second ranked document is the second most relevant one, and so on.

sentence pairs and the first cross-domain benchmarking for automatic paraphrase identification. The data are collected by linking tweets through shared URLs, that are further labeled by human annotators, from 0 to 6.

STS benchmark datasets (Cer et al., 2017) is a classical dataset where pairs of formal texts are scored with labels from 0 to 5 as semantically similar. It has been widely used to train previous SOTA models, so we do not expect our models trained on informal weak pairs of texts to outperform them. However, it is a good indicator of the quality of embeddings and we do expect our models to not deteriorate on accuracy with respect to their initialized versions.

4.3 Baselines

We compare our models with the pre-trained initializations previously described: RoBERTa-base and BERTweet (MEAN pooling of tokens) and S-BERT and S-RoBERTa, pre-trained also on STSb.

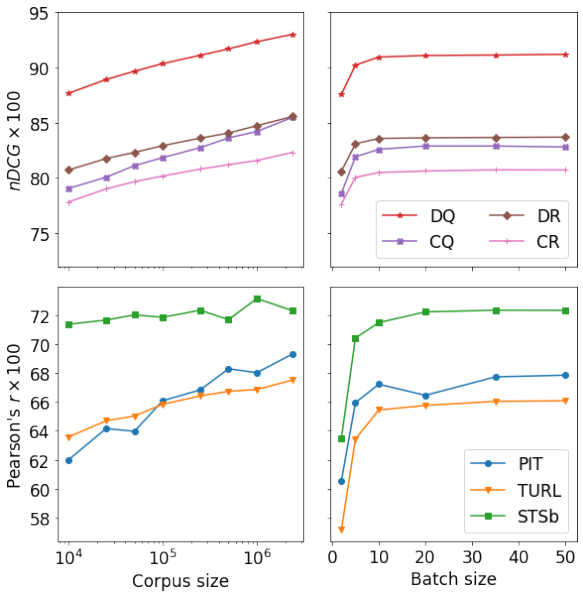

Figure 1: $nDCG \times 100$ and Pearson's $r \times 100$ varying Corpus size (left) and Batch size (right) on Our-BERTweet trained on Quote dataset with MNLoss. Results are averages of 5 runs.

5 Results and Ablation Study

In Table 1 we show the results of the experiments.

As expected, we conclude that baseline models perform poorly in the new benchmarks, being trained for different objectives on different data, while *Our-BERTweet (all)* obtains the best performances. On established datasets, our training procedure improves the corresponding pre-trained versions. The only exception is when our model is initialized from S-BERT and S-RoBERTa and tested on TURL, where we notice a small deterioration of performances (0.5 and 0.1 points respectively) and on STSb-test, since baselines where trained on STSb-train. This result proves that our corpora of weakly similar texts are valuable training sets and specific NLI corpora are not necessary to train accurate sentence embeddings. We remark that for many non-English languages, models such as S-BERT and S-RoBERTa cannot be trained since datasets such as STSb-train do not exist yet[8].

The best initialization for novel benchmarks and PIT is BERTweet, being previously unsupervisedly trained on big amaunts of similar data, while for TURL and STSb the best initializations are S-BERT and S-RoBERTa respectively. MNLoss always produces better results than a simple TripletLoss, since the former compares multiple negative samples for each positive pair, instead of just one as in the latter.

The training dataset does not largely influence the performance of the model on novel benchmarks, while, on enstablished benchmarks, Qt and Rp are usually better than CoQt and CoRp training datasets. However, the concatenation of all datasets (all) used as training set almost always produces better results than when a single dataset is used.

Figure 1 (left) shows that performances improve by increasing the corpus size of Qt dataset. Since they do not reach a plateau yet, we expect better performances when a wider magnitude of Tweets is collected.

Figure 1 (right) shows the performance of the same model when varying batch size in MNLoss, i.e., the number of negative samples for each query. The performance plateaus at about 10, setting a sufficient number of negative samples. However, we set it to a higher value because it implies a faster training step.

6 Conclusions

We propose a simple approach to exploit Twitter in building datasets of weak semantically similar texts. Our results prove that exact paraphrases, such as in NLI datasets, are not necessary to train accurate models generating high-quality sentence-embeddings, since models trained on our datasets

[8]Recently, multilingual approaches have been succesfully tested (Reimers and Gurevych, 2020).

Model	DQ	CQ	DR	CR	Avg	PIT	TURL	STSb
RoBERTa-base	42.9	39.1	55.0	41.0	44.5	39.5	49.7	52.5
BERTweet	46.9	42.5	56.7	44.1	47.5	38.5	48.2	48.2
S-BERT	53.7	43.9	60.5	45.4	50.9	43.8	**69.9**	84.2
S-RoBERTa	52.4	42.8	59.1	44.1	49.6	57.3	69.1	**84.4**
Our-RoBERTa-base (all)	80.8	68.5	83.0	66.1	74.6	58.8	67.5	74.2
Our-BERTweet (all)	83.7	72.1	84.2	68.3	**77.1**	66.1	67.1	72.4
Our-S-BERT (all)	79.0	66.6	81.5	64.6	72.9	57.7	69.4	76.1
Our-S-RoBERTa (all)	80.2	67.8	82.6	65.6	74.0	60.1	69.0	78.9
Our-RoBERTa (Qt)	75.9	63.6	79.3	61.2	70.0	60.7	66.8	74.9
Our-BERTweet (Qt)	80.8	68.9	81.7	65.0	74.1	**67.4**	66.0	72.4
Our-S-BERT (Qt)	73.6	61.5	77.7	59.8	68.1	57.6	69.1	79.3
Our-S-RoBERTa (Qt)	74.6	62.6	78.4	60.5	69.0	58.1	68.8	80.7
Our-BERTweet (Co-Qt)	80.7	70.6	80.8	65.9	74.5	63.6	64.3	70.9
Our-BERTweet (Rp)	81.5	68.4	82.2	65.8	74.5	63.8	67.3	72.3
Our-BERTweet (Co-Rp)	79.3	69.0	81.7	67.5	74.4	62.1	64.3	67.3
Our-BERTweet-TLoss (Qt)	67.7	60.8	71.5	56.9	64.2	53.1	43.4	44.7

Table 1: $nDCG \times 100$ (novel benchmarks) and Pearson's $r \times 100$ (established benchmarks). We indicate our models with the *Our-* prefix followed by the name of the initialization model, between parentheses the training dataset. If not specified, we use MNLoss. Results are averages of 5 runs.

of *weak* pairs perform well on both enstablished and novel benchmakrs of informal texts.

The intrinsic relatedness of quotes with quoted texts and replies with the replied texts is particularly useful when building large datasets without human manual effort. Thus, we plan to expand the study to other languages spoken in Twitter. Two months of English data are more than enough to build large datasets, but the time window can be easily extended for rarer languages, as today more than 9 years of data are available to download. Finally, we also hypothesize that this approach can be adapted to build high-quality embeddings for text classification tasks. We will extensively explore this on Twitter-related tasks.

7 Ethical Considerations

We generate the training datasets and novel benchmarks starting from the general Twitter Stream collected by the Archive Team Twitter, as described in § 2. They store data coming from the Twitter Stream and share it in compressed files each month without limits. This collection is useful since we can design and perform experiments on Twitter data that are completely reproducible. However, it does not honor users' post deletions, account suspensions made by Twitter, or users' changes from public to private. Using Twitter official API to generate a dataset is not a good option for reproducibility since parts of data could be missing due to Twitter Terms of Service. We believe that our usage of Twitter Stream Archive is not harmful since we do not collect any delicate information from tweets and users. We download textual data and connections between texts (quotes and replies), and we also remove screen names mentioned in the tweets during the cleaning step.

However, we agree that Twitter Stream Archive could help malicious and unethical behaviours through inappropriate usage of its data.

References

Giuseppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus

Sahlgren. 2021. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations*.

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.

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. A multi-task approach for disentangling syntax and semantics in sentence 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 2453–2464, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Mathias Creutz. 2018. Open subtitles paraphrase corpus for six languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Jingfei Du, Edouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Veselin Stoyanov, and Alexis Conneau. 2021. Self-training improves pre-training for natural language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5408–5418, Online. Association for Computational Linguistics.

John M. Giorgi, Osvald Nitski, Gary D. Bader, and Bo Wang. 2020. Declutr: Deep contrastive learning for unsupervised textual representations.

Aaron Gokaslan and Vanya Cohen. 2019. Openwebtext corpus. `http://Skylion007.github.io/OpenWebTextCorpus`.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *ArXiv e-prints*.

James Y. Huang, Kuan-Hao Huang, and Kai-Wei Chang. 2021. Disentangling semantics and syntax in sentence embeddings with pre-trained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1372–1379, Online. Association for Computational Linguistics.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. In *Proceedings of The 2017 Conference on Empirical Methods on Natural Language Processing (EMNLP)*, pages 1235–1245. Association for Computational Linguistics.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv e-prints*, page arXiv:1907.11692.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021. Tsdae: Using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

John Wieting, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A bilingual generative transformer for semantic sentence embedding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1581–1594, Online. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Bill Dolan. 2015. SemEval-2015 task 1: Paraphrase and semantic similarity in Twitter (PIT). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 1–11, Denver, Colorado. Association for Computational Linguistics.

Appendix

We show randomly selected examples of training data from Qt and CoQt datasets in Figure 2, and from Rp and CoRp datasets in Figure 3.

Quote	Quoted tweet
me either he always got an attitude	frrrrr i cant stand that mfer-
hahahahahahaha he deleted and posted a new one already 😭	someone tell jinyoung to get rid of the date please 😭😭😭 it shows he received the poster on nov 4th helppp [#got7 #갓세븐 #got7 #아가새 #got7_breathoflove_lastpiece #got7_breath #got7_lastpiece]
according to multiple sources, a meeting was held in logar (post us-tb deal) where the haqqani leadership instructed its cadres to focus instead (of suicide attacks) on identifying and killing individuals who support the post 2001 order. why is the world not dealing with this?	open killing season on anyone attempting to improve afghanistan or to take it to a better place. this is the ultimate definition of terrorism: terrorise them to the point of silence
oh wow, thank you so much for this incredible review. you've just made our day 😄. merry christmas!!!	are you looking for last minute christmas presents you don't need to go to the shops for? i have a recommendation for you! hubby got me the packs app six months ago and we're loving it. excuse my terrible food photography as i try to explain why ... 1/?
that says it all about tory blair and the witch splodge	margaret hodge became leader of islington council in 1982. during the time she was in charge, many vulnerable children in the borough's "care homes" were abused, forced into prostitution & raped by people in positions of trust. tony blair later made her the minister for children!
park jihoon #treasure #트레저 #mamavote #treasure	goal : 1000 retweets [#2020mama] voted for #treasure on #mamavote ▏2020 mama ▏2020.12.06 (sun)
ok i have made my brain calm so ayern thank you so much again!!! i didnt expect to win ofc hahahha pero nag donate na din to help <3 this is just an extra blessing huhuhuhhu tysm lord	i put in 1 raffle entry for every 5 pesos donated according to the order of entries on the form then generated a random number which corresponds to the winner andddddd..... lucky #122 is !!! 🎉🎉🎉 congratulations on winning mingyu's signed tone up sun cream 🐶❤️
keep drinking the kool aid i believe in god not man. have a wonderful day.	it's not a lie. obama didn't replenish the ppe.
sven!!! the only cat i love with my whole heart.	our fearless leader, sven 📸:
lool this was posted before his 50 yd td catch and run smh	anyone playing against dalvin cook in fantasy

Quote 1	Quote 2
aint nobody looking at that damn zebra 😭	uno i was prepared to mute wz's name bc i thought this gif was gonna end up like the juyeon one with those captions 💀 😭
picture perfect indeed😌	nigeria map in the mud 💔
excruciating national heartache. healthcare workers we see you too. 💔	🥺😩😩😩 pull it together people! #covid19
and she persists, fierce women we believe in!!	love this! #electoralcollege #womenworthwatching #womengettingitdone
imagine calling someone toxic because they tryna defend their fave from psychopaths and bullies	elites they're calling you peoples names here o lmaooo 😈😈😈😈!!
cancel culture 🤝 being selective	there's so many tweets i don't know what "that" even means in this context ☠️
an update on esl pro tour & iem katowice 2021 dreamhack warcraft iii championship esl pro tour championship dates: march 4-7 (new dates!) $130,000 prize purse 16 player tournament (format unchanged)	the #iem katowice csgo, sc2 and wc3 tournaments will all be played as no audience, studio events. it's a great shame to go without an audience two years running, but it is what it is. we will see you in spodek when it's safe to do so.
why not pressure your party now to reverse the pause in the current legislation, before its -15? it wasn't struck out of the books in the 90's. the section on rent control was just paused it could be reinstated tomorrow, if wanted it w/o recalling the mlas	in what is by far the biggest break so far from existing govt policy, leadership candidate is pitching rent control to help address housing problems.
absolutely spot-on from - which means an even more fundamental rethink for small l liberals on left and right...	exactly. and the republican party has changed fundamentally. centre ground politics in the us is still in a v difficult long-term position
his punishment, living in indianapolis, will haunt him forever.	jackie we sincerely apologize for this totally unacceptable behavior, and will have a statement this morning about actions being taken harassment of this kind has no place or justification this is not ok

Figure 2: Examples of pairs of texts from Qt (top) and CoQt (bottom) datasets.

Tweet	Reply
first time in 4 years a republican has mentioned the deficit.	we pay the to work for the president of russia & we pay republicans to work for putin who pays for dead americans. corruption is the currency of republicans. 🇺🇸
hussain haqqani's saath forum is denying links with efsas which posted its own participation at the second saath forum conference held in london uk on 16 october 2017 on efsas own website. link here: /1	efsas sent yoana barakova to attend the saath forum conference held in uk on 16 of october 2017. yoana barakova mentioned by name in the eudisinfolab report as an indian sponsored propagandists is seen with hussain haqqani posted by efsas website: /2
9.) sent documents w/ inflated numbers and hidden debts to make himself seem like a better business partner. these docs are now at the center of a newyorkstateag investigation -- a key part of trump's legal headaches post-election.	8.) defied real-estate industry wisdom by sinking $400m+ of his own cash into big real-estate projects. many of these look like bad bets, on properties that consistently lose $. (as nytimes confirmed in its great trump-taxes stories).
guys, we're the purple line, really super close to 6th and 5th place on ichart 😭 we need to get last piece chart higher on the respective korea streaming platforms and we'll definitely go up 💥💟💥💟 got7official #got7 #갓세븐	our solid #1 on genie daily chart and also #3 on genie real-time chart is hard carrying us on ichart💚 got7official #got7 #갓세븐
can't make it up— is now campaign with beto "let's go door to door and seize guns by force" o'rourke. ossoff previously was caught taking a hard position on guns in metro atlanta while running ads about protecting the second amendment in rural georgia.	john cornyn. what a loser.. you must have some pakistani in you. 😥😥
" ""the pm has said he loathes bullying and yet today he has comprehensively failed a test of his leadership, when he's had a report on his desk, precisely on this issue"" shadow home secretary nick thomas-symonds is ""shocked"" priti patel remains in post "	and that ladies and gents is called ministerial corruption.. enjoy!
i have been studying this old map for a while now. the map here is actually showing us that down or south of the sahara desert we have the ancient world meaning we have been existing before the nations above the sahara desert. meaning they all migrated from the ancient world.	and we also have a new jerusalem (jebu) above meaning there is definitely an old jerusalem (jebu).
this is why the democrats fought so hard to keep amy coney barrett out of the supreme court!	they knew it would come to this. glad the seat got filled.
. just when the complicity of the mainstream media had succeeded in making the transition to the new world order almost painless and unnoticed, all sorts of deceptions, scandals and crimes are coming to light. until a few months ago, it was easy to smear...	... as "conspiracy theorists" those who denounced these terrible plans, which we now see being carried out down to the smallest detail.
any questions? anyone? any trump supporters have any questions???	people need to understand this

Reply 1	Reply 2
why are we leaving? any one got a benefit to share yet with the majority of us who don't want brexit ?	2/ i'm told that the uk has offered 3 year status quo on access in the 12m to 200m zone of the u.k. eez but after that uk would have a free hand.
hello everyone including viewers. they should have cancelled long time ago, what are they waiting for. we don't want to bury innocent souls tshepho godfrey mollo boksburg gae zebediela makgophong #fullview #sabcnews	they must close these events, we have seen maskandi events people were over the set amount, people were not even wearing masks. so it's wise to suspend these events and those breaking rules must be punished... prisoned
#happinessindecember [#2020mama] voted for #redvelvet on #mamavote | 2020 mama | 2020.12.06 (sun)	1 red velvet best idol group alive luvies got your back #happinessroadto100m [#2020mama] voted for #redvelvet on #mamavote | 2020 mama | 2020.12.06 (sun) mnetmama
if ohanaeze said what ipob is doing did not have head, we will cut off their heads and put it there and it will have head	when you start the campaign for biafra restoration,we will begin to believe not trust you, for now, you people are anti igbo, that your own do not trust one bit. remember the clock is ticking. make hay while sun is shining. a word is enough for a fool
ihh this guy was a real baller🎯⚽	seen this video for 55th time in the past 2 weeks
happy birthday annaa❤️❤️	happy birthday jagananna
if you didn't totally punk out you would have been pardoned by now.	he had one of the best questions to sarah huckabee sanders in 2018. we still don't know what the answer is.
for years now your career is not yet stable and you can't work on that , all you could do is to publish bad news about others, crazy reporter #abt davido	how does his relationship with chioma affects the present nigeria economy?
in front of a live audience, which is allowed in nyc but not restaurants.	yup, just two "maskless" guys, sitting "2 feet apart" working at their "jobs" in front of a "live audience" making fun of people not willing to "social distance" "stay at home" & "lose their jobs".
you say that like it matters...like it could be true.	i imagine lie are infinite right? you can fabricate as much evidence as you want.

Figure 3: Examples of pairs of texts from Rp (top) and CoRp (bottom) datasets.