Twice - Twitter Content Embeddings

Xianjing Liu[1], Behzad Golshan[1], Kenny Leung[1], Aman Saini[1], Vivek Kulkarni[1], Ali Mollahosseini[1] and Jeff Mo[1]

[1]*Twitter Cortex*

Abstract

In this short paper, we describe a new model for learning content-based tweet embeddings that serve to be generically useful as signals for a variety of down-stream predictive tasks. In contrast to prior approaches that only leverage cues from the raw text, we take a holistic approach and propose Twice, a model for learning tweet embeddings that (a) leverages cues beyond the raw text (including media) and (b) attempts to yield representations optimized for overall similarity, a combination of topical, semantic, and engagement similarity. Offline evaluations suggest that our model yields richer and superior embeddings compared to the benchmark models in the tasks evaluating on both academia dataset and twitter products.

Keywords

Deep Learning, Recommendation, Embedding, NLP

1. Introduction

A rich representation of a tweet that captures nuances in meaning is critical for most predictive models at Twitter (including Topics, Health, Recommendations etc.). Consequently, there is an urgent need for models that can encode or summarize a tweet's content into a dense representation – a representation that can then be used in various downstream models. Some key surface areas which will be using tweet embeddings include Home Timeline, Notifications, Topics, and potentially Health models. An important requirement is that the tweet embedding be generically useful on a variety of predictive tasks and not necessarily be useful for only a specific apriori task. Finally, we expect downstream models to only consume the embeddings, without having to worry about the inner workings of the underlying model.

Taking a bird's eye view, the need is simply for a tweet representation/embedding that captures similarity between tweets by embedding them in a vector space. Specifically, tweets that are "similar" must be close in the vector space and tweets that are not "similar" should ideally be far in this vector space with respect to a suitable metric. Zooming in, for practical modeling it is useful to attempt to operationalize this vague notion of "similar" and attempt to be more specific here. We can attempt to capture the following notions of similarity: (a) *Semantic Similarity* – tweets which are similar in meaning should be close in the embedding space. An example pair of semantically similar sentences is "The quick brown fox jumped over the lazy dog" and "The brown fox leaped over the lazy dog". (b) *Topic Similarity* – tweets that are about the same topic should be close in the embedding space. An example pair here would be "Don Bradman is the greatest cricket player of all time" and "India won the Cricket World Cup" since both tweets are about "Cricket". (c) *Engagement Similarity* – tweets that share engagement audiences are deemed similar.

In this paper, we present Twice – a model that attempts to capture the above notions of tweet similarity. In contrast to most prior work which only seeks to embed raw tweet text using standard pre-trained language models, Twice models tweets holistically leveraging not only raw tweet text, but also incorporating cues from the associated media, and hyperlinks. We evaluate Twice on a suite of offline benchmark tasks and demonstrate that our proposed model significantly outperforms several baseline approaches.

2. Related Work

Our work is very closely related to work in the area of learning sentence embeddings which seeks to learn dense representations of sentences and capture sentence similarity. One of the earliest works on learning dense embeddings of sentences is the work of Le and Mikolov [1] which generalized the Skipgram word-embedding models [2] to learn sentence embeddings and paragraph embeddings. With the rise of convolutional and recurrent neural network models, several approaches to learn sentence embeddings were proposed [3, 4, 5, 6, 7]. Additionally, a couple of these works sought to learn representations of tweets by applying these networks to Twitter

DL4SR'22: Workshop on Deep Learning for Search and Recommendation, co-located with the 31st ACM International Conference on Information and Knowledge Management (CIKM), October 17-21, 2022, Atlanta, USA

xianjingabbyl@twitter.com (X. Liu); bgolshan@twitter.com (B. Golshan); kennyleung@twitter.com (K. Leung); amansaini@twitter.com (A. Saini); vkulkarni@twitter.com (V. Kulkarni); amollahosseini@twitter.com (A. Mollahosseini); jeffm@twitter.com (J. Mo)

text [5, 4]. Finally, with the introduction of Transformers and pretrained language models, the current state of art approaches now use pre-trained language models coupled with contrastive loss functions to learn sentence embeddings [8, 9, 10, 11, 12, 13, 14, 15]. All of these approaches only look at embedding generic sentences and are not attuned to embedding tweets where deeper semantic cues, and multi-modal content can be used to obtain rich representations. Our model TWICE, in-turn builds on these works, but also incorporates cues specific to tweets (like media and hyperlink) to yield rich embeddings of tweets.

3. TWICE

At its core, TWICE is a BERT model [16] trained on a multi-task loss function. Figure 1 shows the model architecture of TWICE model. In particular, we consider the following three tasks, each attempting to capture notions of similarity as noted in Section 1. More specifically, we optimize a standard BERT model on the following tasks:

- TOPIC PREDICTION: The task is to predict the concept topics associated with the tweet. This enables the representation to capture topical similarity. In particular, we optimize binary cross entropy loss since this is a multi-label prediction setting where a tweet may be associated with multiple topics. For instance, the tweet "America is heading back to the Moon, folks. No astronauts, but likely to glean loads of data." belongs to "Space" and "Science" topics. The total number of concept topics in this task is 419.
- ENGAGEMENT PREDICTION: Given a representation of the user (obtained by encoding a user biography) and a tweet, the task is to predict if the user engages with the tweet. This task is essentially identical to the task in the well-known CLIP (Contrastive Language-Image Pre-Training) model [17] except that instead of embedding images using an image encoder, we embed the user biographies using a standard BERT encoder. The loss function used is identical to the one described in [17]. Training on this task enables us to capture tweet similarity based on user engagement patterns and may be particularly useful to model especially when down-stream products may want to maximize user engagements.
- LANGUAGE PREDICTION: Because we desire multilingual support, we would like tweet representations to also encode language cues, so that tweets of the same language tend to be closer than ones from different languages. Therefore, we explicitly train on the task of predicting the language of the tweet and use the standard cross-entropy loss function for this task.

The full loss function is simply the average of the above three loss functions. Finally, to obtain a dense representation of the tweet, we simply use the representation of the [CLS] token. TWICE leverages cues from the entire tweet and not just the raw tweet text. In particular, in addition to the raw text, we also leverage media cues by obtaining media annotations for any associated media as described in [18]. These media annotations are simply concatenated to the raw text via a separator token before being input to the model. Similarly, when a tweet has hyperlinks, we extract the first 100 tokens of the webpage title and description as encoded in the linked HTML page. These features are also appended to the input.

Training procedure. TWICE is trained on a dataset of 200 million tweets sampled over a 90 day interval. We also associate with these tweets the users who engaged with them (which the CLIP task component requires). The model was optimized using standard ADAM with weight decay as the optimization procedure and trained for 5 epochs until convergence.

4. Experiments

We evaluate TWICE both quantitatively and qualitatively each of which we discuss below.

4.1. Quantitative Evaluation

Setup. We quantify the effectiveness of tweet embeddings in capturing content similarity via measuring their performance on three benchmark tasks – tasks that reflect how well embeddings capture notions of similarity noted in Section 1:

- **SemEvalPIT** [19]. This benchmark is an academic benchmark and consists of about 1000 pairs of tweets with similarity scores obtained by human judgments. We measure the performance of embeddings on this task by computing the Spearman correlation of similarity scores obtained for these tweet pairs in embedding space with human judgements. Higher correlations suggest better embeddings reflecting better alignment with human-derived similarity judgments.
- **Recalling Favorites (Favs).** In order to measure the effectiveness of these embeddings in down-stream predictive models of engagement, we consider the task of recalling (based on just a top k nearest neighbor lookup) which tweets a given user favorites from more than 5k candidate of tweets given their past engagement history. Higher scores reflect better embeddings.

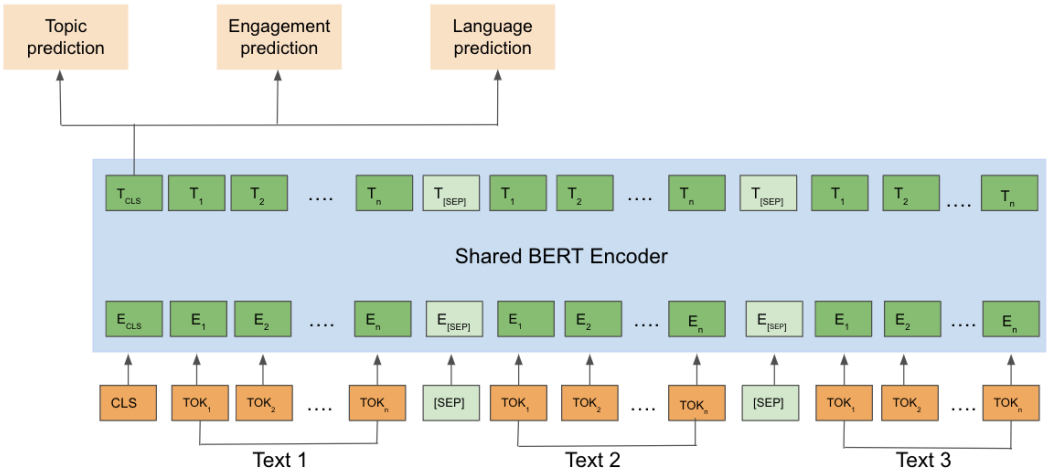

Figure 1: TWICE model is a BERT based model trained on a multi-task loss function.

- **Topic Assignment Precision (Topics).** We compute the precision of topic assignments on a test set of topical tweets. Higher precision suggest better encoding of topical similarity. Once again, here we only base our decisions using a k-NN classifier.

Our rationale for restricting ourselves to using very simple nearest neighbor approach based on cosine similarity of tweet embeddings is based on the intuition that higher quality representations would inherently demonstrate a higher-degree of "ease of extraction" of the predictive information. It is for this reason precisely that one needs to use simple models as opposed to very complex deep predictive models. We made a design choice to use a NN-based approach which supported quick implementation but simple shallow models are another alternative. Finally, we summarize model performance by reporting the harmonic mean over tasks.

Baselines. We consider the following baseline models:

- BERT: This is just the standard pre-trained BERT model [16] and serves as the simplest but strong baseline that one could use to embed tweets.
- SIMCSE: SIMCSE [13] is a state-of-the-art sentence embedding approach that learns sentence embedding using an unsupervised approach. The main idea of SIMCSE is to pass a tweet X twice through BERT (with dropout enabled). This yields two different (noisy) representations of X. The idea is to these as positive examples. X and all other examples in the batch are treated as negative examples. The objective to maximize cosine similarity in the representations of the positive pair and minimize this between X and the negative examples. It is to be noted here that we train SIMCSE on Twitter data.
- HASHSPACE: HASHSPACE is a BERT based model trained on the task of hashtag prediction where the model is optimized to predict the correct hashtag associated with a tweet from a set of 100K hashtags. Hashspace currently as deployed at Twitter only uses the raw tweet text as cues to make predictions, and does not model tweets holistically.
- TOPICSPACE: TOPICSPACE addresses two main limitations of HASHSPACE. First, in contrast to HASHSPACE, we model tweets holistically and leverage cues from media, and hyperlinks as well. Second, we simplify the predictive task. Instead of learning to predict a label from a universe of 100K labels (hashtags), we only learn to predict one or more topics from a space of 419 concept topics. The intuition is that to capture similarity, it is sufficient to capture fairly broad topics than seek to capture extremely fine-grained hashtags. By making these two changes, we note that we can learn a model with a better fit to the data yielding richer representations.
- CLIP: The original CLIP in [17] is a neural network trained on 400 millions of (image, text) pairs. Our CLIP model is identical to the original CLIP model in the model architecture. Our CLIP model is trained using a multi-modal method in which

	SemEvalPIT	Favs	Topics	Mean (HM)
Bert	0.025	0.043	0.063	0.038
Simcse	0.218	0.022	0.205	0.055
Hashspace	0.336	0.064	0.230	0.131
Topicspace	0.264	0.075	0.599	0.160
Clip	0.225	0.136	0.271	0.194
Twice	0.302	0.102	0.429	0.194

Table 1
Quantitative performance of various tweet embedding approaches on our benchmark suite. We report the Spearman correlation for SemEvalPIT, Top-K recall for Favs and Precision for Topics. Note that our model Twice outperforms most standard baselines including the currently deployed Hashspace.

the model attempts to predict whether a piece of media and text come from the same tweet or not. In this setting we replace the image encoder in the original Clip model with a user-biography encoder.

Results. Table 1 shows the results of our evaluation on our benchmark suite. Based on these results we can make the following conclusions: (a) First, observe that just using standard models like Bert for embedding tweets does not yield superior embeddings. It is imperative to learn embeddings from Twitter data. (b) Second, state-of-the-art unsupervised methods for sentence embedding perform worse than supervised methods which is inline with prior work on sentence embeddings as well. (c) Topicspace significantly outperforms Hashspace overall. This is because Topicspace leverages cues from beyond tweet text and also uses a simpler but more intuitive task. (d) Models like Topicspace and Clip which solely optimize for a specific notion of similarity tend to perform significantly better on the corresponding evaluation tasks than other models simply because the underlying representations are optimized to capture that specific similarity notion over others. (e) Finally, note that our proposed model Twice generally outperforms all of these baseline approaches overall. While the mean performance of Twice and Clip are identical, note that Twice outperforms Clip on both **SemEvalPIT** and the **Topics** tasks significantly with a slight drop on the **Favs** task.

To summarize, all in all Twice demonstrates superior performance over prior production models and yields improved embeddings of tweets by leveraging cues beyond the tweet text.

4.2. Quantitative Evaluation and Analysis of Usage in Health Products

Setup. We evaluate the performance of content embeddings on our health platform. We use Twice embeddings as features in a shallow model to predict Spam and Terms of Service (ToS) violations. The spam prediction is a binary classification task to predict whether a tweet is a spam or not. The ToS violation prediction is a multi-label classification task. There are 8 labels in total. Some examples of ToS violation labels include 'violence', 'threaten someone', 'suicide', etc.

Results. Table 2 shows the average precision score of using the Twice embeddings on the tasks of Spam detection and ToS violation classification. We compare the result with the benchmark models Bert and Hashspace. The results show that Twice outperforms both the standard Bert and Hashspace models in both the Spam detection and ToS violation tasks.

4.3. Qualitative Evaluation and Analysis of Usage in Content Recommenders

In order to evaluate our model qualitatively, we also built a web page where one can enter a tweet ID and see the nearest neighbors to the given tweet from a given predetermined universe of tweets – a scenario that reflects the usage of embeddings for candidate generation in content recommenders. Figure 2 shows the nearest neighbors for a couple of seed tweets as a demonstration. Note that the nearest neighbors reflect the broad topic of the seed tweet and are similar in content to the seed tweet suggesting that our model is able to capture content similarity between tweets.

While the process outlined in this section can be used for candidate generation in content recommenders (by finding tweets similar to a user's interests and past engagements), through a qualitative analysis conducted offline, we have identified a list of challenges that need to be addressed in-order to ensure good quality candidates are returned.

- **Seed Selection.** Using tweets with which users have positively engaged as seeds to fetch more interesting content is a natural choice. However, some seed tweets may have very little content which makes them unsuitable for retrieving candidate tweets. Examples include tweets that contain frequent phrases like "good morning", everyday greetings, and daily life updates.

	Spam	ToS
BERT	0.41	0.328
HASHSPACE	0.44	0.286
TWICE	0.47	0.347

Table 2
The average precision score of using TWICE embeddings for the Spam detection and TOS violation classification tasks.

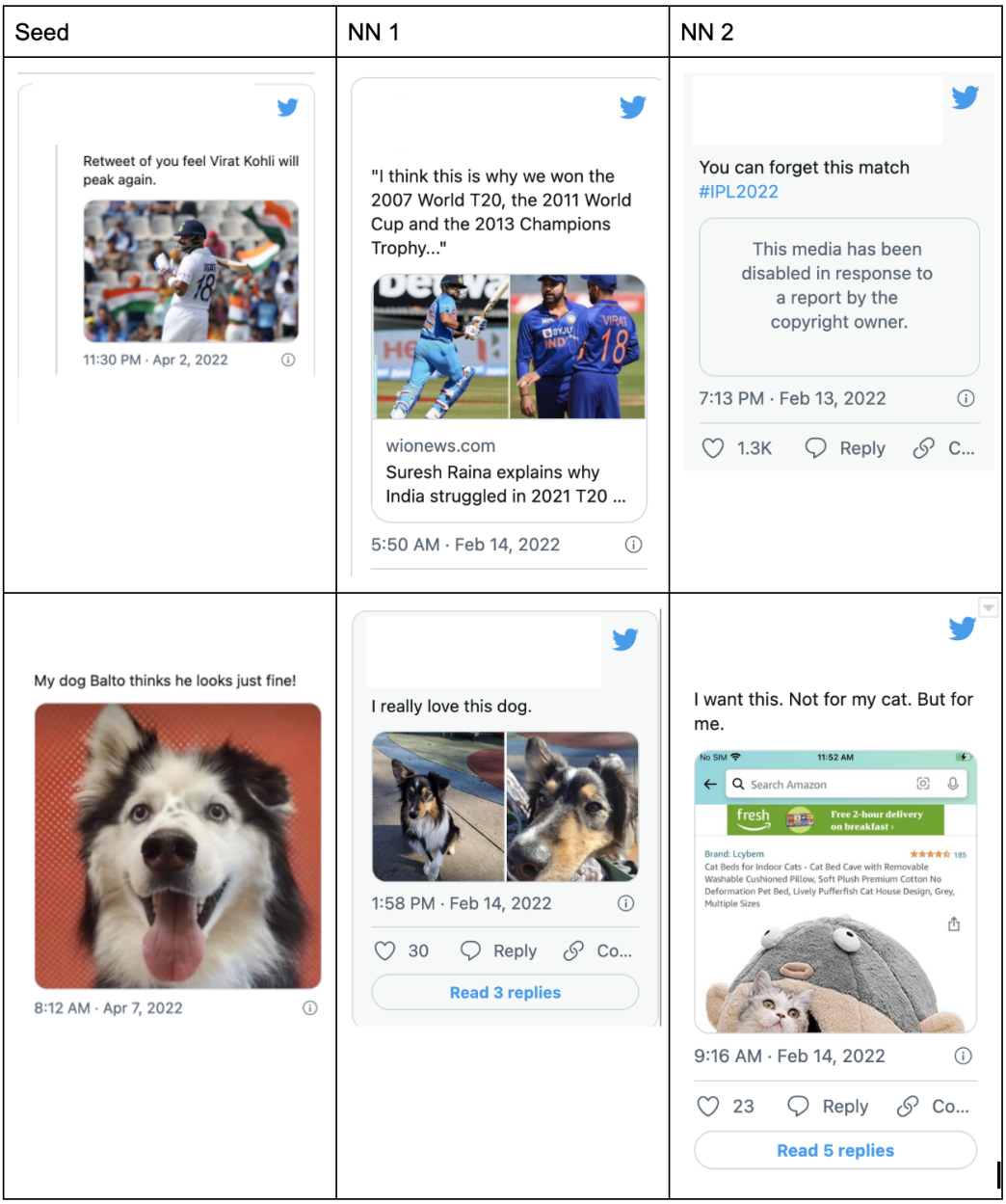

Figure 2: Nearest neighbors to a couple of seed tweets in the TWICE embedding space.

- **Content Candidates Quality.** Some tweets may have very little to no content and are non-topical. This includes tweets that may only have single emojis, may be very short in length, or only consist of a shortened URL. Such candidates should not be part of the universe. This is by far the most pervasive and immediate challenge to be addressed.
- **Health Considerations.** Some tweets can be spam, violence, and etc. Recommending such content is problematic and needs to be addressed before tweet representations can be used for content recommendations.
- **Recency.** Some tweets returned may be irrelevant (or at least non-engaging) simply because they are old and outdated. Candidate generators need to adapt their responses to the recency requirements of the product.

Note that these challenges are independent of the underlying tweet representation itself and may significantly hinder the quality of candidates even if the tweet representation model is of a superior quality. To address these challenges, we build various filters and apply to the candidate pool, we also carefully design and pick desirable seed tweets to generate high quality tweets for the user.

5. Conclusion

In this paper, we proposed a model for embedding tweets that goes beyond just modeling tweet text. Our goal has been to develop generically useful rich representations of tweets that can be used in a variety of downstream predictive models at Twitter. To that end, we have demonstrated through offline evaluation that our proposed model outperforms the benchmark models on various tweet products. As next steps, we seek to validate our model using online A/B tests in various product surfaces which serve as the ultimate litmus test.

References

[1] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: International conference on machine learning, PMLR, 2014, pp. 1188–1196.

[2] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems 26 (2013).

[3] J. Wieting, M. Bansal, K. Gimpel, K. Livescu, Towards universal paraphrastic sentence embeddings, arXiv preprint arXiv:1511.08198 (2015).

[4] S. Vosoughi, P. Vijayaraghavan, D. Roy, Tweet2vec: Learning tweet embeddings using character-level cnn-lstm encoder-decoder, in: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, 2016, pp. 1041–1044.

[5] B. Dhingra, Z. Zhou, D. Fitzpatrick, M. Muehl, W. Cohen, Tweet2vec: Character-based distributed representations for social media, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2016, pp. 269–274.

[6] F. Hill, K. Cho, A. Korhonen, Learning distributed representations of sentences from unlabelled data, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 1367–1377.

[7] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, A. Bordes, Supervised learning of universal sentence representations from natural language inference data, in: EMNLP, 2017.

[8] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, et al., Universal sentence encoder, arXiv preprint arXiv:1803.11175 (2018).

[9] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3982–3992.

[10] L. Wang, C. Gao, J. Wei, W. Ma, R. Liu, S. Vosoughi, An empirical survey of unsupervised text representation methods on twitter data, in: Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020), 2020, pp. 209–214.

[11] H. Yin, X. Song, S. Yang, G. Huang, J. Li, Representation learning for short text clustering, in: International Conference on Web Information Systems Engineering, Springer, 2021, pp. 321–335.

[12] S. Kayal, Unsupervised sentence-embeddings by manifold approximation and projection, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2021, pp. 1–11.

[13] T. Gao, X. Yao, D. Chen, Simcse: Simple contrastive learning of sentence embeddings, in: EMNLP (1), 2021.

[14] J. Huang, D. Tang, W. Zhong, S. Lu, L. Shou, M. Gong, D. Jiang, N. Duan, Whiteningbert: An easy unsupervised sentence embedding approach, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 238–244.

[15] D. Liao, Sentence embeddings using supervised con-

trastive learning, arXiv preprint arXiv:2106.04791 (2021).

[16] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[18] V. Kulkarni, K. Leung, A. Haghighi, CTM - a model for large-scale multi-view tweet topic classification, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track, Association for Computational Linguistics, Hybrid: Seattle, Washington + Online, 2022, pp. 247–258. URL: https://aclanthology.org/2022.naacl-industry.28. doi:10.18653/v1/2022.naacl-industry.28.

[19] W. Xu, C. Callison-Burch, B. Dolan, SemEval-2015 Task 1: Paraphrase and semantic similarity in Twitter (PIT), in: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, 2015, pp. 1–11. URL: https://www.aclweb.org/anthology/S15-2001. doi:10.18653/v1/S15-2001.