# Transformer-based Topic Modeling to Measure the Severity of Eating Disorder Symptoms

Notebook for the eRisk Lab at CLEF 2023

Diana-Nicoleta Grigore[1], Ioana Pintilie[1]

[1]*Faculty of Mathematics and Computer Science, University of Bucharest, Romania*

Abstract

In this paper, we describe a topic-driven approach for detecting the severity of eating disorder symptoms. We extract more task relevant embeddings with the help of a MentalBERT model pretrained on ED data. We then employ the use of BERTopic to extract probability scores associated with identified discussion themes. These become features used to predict the answers given by users in the Eating Disorder Examination Questionnaire, based on their social media post history. The task is introduced in the CLEF eRisk 2023 competition, in which we participate as team RiskBusters. We obtain the best results in the Shape Concern Subscale and are competitive on all the other metrics.

**Keywords**

topic-based classification, social media, eating disorder detection, mental health transformers

## 1. Introduction

As a highly accessible way of communication, social media proves to be the perfect medium for self-expression. Under the benefit of anonymity, users share personal experiences and insights, come forward as advocates for mental health support communities or seek information. Eating disorders are a growing concern impacting people worldwide and early detection is crucial to ensuring positive outcomes for those affected. As symptoms are hidden in day-to-day behaviours, using data coming from online sources offers a better chance of capturing them.

Mental Health professionals use the Eating Disorder Examination Questionnaire (EDE-Q), a self-reporting tool, to understand the range, frequency and severity of symptoms and how those affect a person [1]. CLEF's eRisk 2023 competition proposes a task for measuring the impact of ED symptoms using a Reddit posts dataset and the answers the users give to the aforementioned questionnaire. The work presented in this paper describes the RiskBusters team's approach to predicting an individual's answers to the questions, based on their social media presence.

We extract common patterns in the user's discourse, using a framework for topic modeling that is based on transformers [2] and return the probabilities with which a set of topics appear in the users' messages. The resulting scores serve as features for several TabPFN [3] classifiers, each predicting the answer to one question.

*CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece*
diana.grigore13@s.unibuc.ro (D. Grigore); ioana.pintilie@s.unibuc.ro (I. Pintilie)
0009-0003-6371-8394 (D. Grigore); 0009-0005-9844-8069 (I. Pintilie)

We domain-adapt MentalBERT on Anorexia data with MLM pretraining. We obtain promising results on the proposed task by employing these embeddings to examine a user's discussion theme distribution. We observe that using these topic probabilities as input for classification enables us to get good performance with little information.

## 2. Related Work

Mental Health Disorder detection based on social media posts has become a popular research area in recent years. In particular, work on eating disorders focused on identifying individuals at risk using traditional topic modeling [4] or transformers [5].

Assessing the severity of eating disorder symptoms by predicting responses to the EDE-Q has been previously addressed in the 2022 edition of eRisk [6]. Due to the limited number of available samples and the complexity of the questionnaire, the task presents a unique challenge. Participating systems mostly make use of techniques that capture the semantic similarity between the questions and the posts, achieving good performance, considering that no training data was available for this previous iteration of the task. To this end, the systems use either transformer embeddings to encode the available data [7], or pretrained word vectors with additional feature engineering to extract relevant keywords in the questions [8]. The eRisk 2018 [9] anorexia dataset is used for additional fine-tuning or evaluation. The best performing system [10] uses a fine-tuned BERT [11] and cosine similarity to assign symptom severity.

Using the topical dimension of social media posts for understanding sentiment is a commonplace technique in NLP. Traditional modeling approaches (LDA [12], Top2Vec [13]) rely on individual words for discovering topics [14], but fail to capture complex contextual relationships in sentences. By employing the use of transformers, topic modeling can be framed as an embedding clustering task, where each created topic has a descriptive, contextual latent representation that can be purposed downstream [2]. This method has been mostly applied on datasets sourced from discussion trees of short posts (coming from Twitter or Reddit) for the analysis of trend evolution. For example, numerous works have observed how the COVID-19 pandemic shaped opinions [15] [16] or increased the number of mental health issues [17], [18], [19], [20]. Closer to our work, one study [21] describes how aspects from the sphere of eating disorders (such as dieting, substance abuse, and increased physical activity) came up in a number of topics when BERT embeddings were used. Some approaches go further and focus on user-level, by adding classifiers to flag posts as suggesting depressive, anxious or autistic behaviour [22] or trying to recommend therapeutic techniques fit for specific situations [23].

Our approach is unique, as we propose the use of a domain-adapted transformer (pretrained on eating disorder-related content) to estimate the user-oriented topic distribution as an input for simple classifiers and further predict a degree for the symptoms captured in the EDE-Q.

## 3. Method

We measure the severity of Eating Disorder signs by leveraging our user-level topic distribution method, presented in Figure 1. The training dataset of this task [24] consists of social media post and comment history from 28 Reddit users. For each user, the ground truth is a set of integers

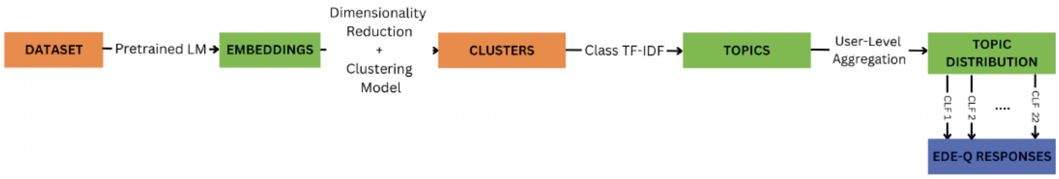


**Figure 1:** The end-to-end EDE-Q answering pipeline consists of topic distribution extraction and classification performed for each user.

ranging from 0 to 6, representing their answers to the questions in the EDE-Q. We approach the problem by employing a classifier for each question, where the answer corresponds to the degree of experiencing the target symptom. There is a total of 22 questions, grouped in 4 main categories: Restraint, Eating, Shape and Weight Concern. Each class refers to a major symptom set experienced with an eating disorder diagnosis and is analyzed from different perspectives. We train our base classifiers on the topic data from all users and observe that the label distribution is skewed towards either 0 or 6.

## 3.1. Transformer-Based Topic Modeling

We start by hypothesizing that the discussion subjects present in users' posts on Reddit will be informative enough to be used as features for downstream classification. The first step in our solution consists of discovering these topics using transformers. To this end, we employ the use of the BERTopic [2] framework, due to its highly customizable nature. We start by generating embeddings, usually from pretrained language models. We try both publically available transformers and our own domain-adapted versions, as well as other embedding generation techniques, such as the Universal Sentence Encoder [25].

The next pipeline step is reducing the dimensionality of these sentence embeddings and clustering them. We use UMAP [26] as our default dimensionality reduction technique. The clustered messages can now form the discussion topics and can be obtained with BERTopic's class TF-IDF. Due to limitations of the classification model used in the following step, we force HDBSCAN [27], our clustering model, to generate at most 100 clusters. At inference time, we use the topics obtained and generate scores for each post in the test dataset.

## 3.2. Domain-adaptive pretraining

We expect that extracting embeddings from transformers trained on mental health content will lead to identification of topics that are more relevant to our task. We therefore choose to experiment with MentalBERT [28], a transformer trained on posts collected from social media, covering topics such as depression, suicide and suicidal ideation, anxiety, posttraumatic stress disorder, and bipolar disorder.

Given that the datasets used in the training process do not explicitly include any ED content, we continue MentalBERT's pretraining with a masked language modeling objective [29] on the eRisk 2018 anorexia dataset [9]. Since the data was released for the task of detecting users

**Table 1**
Results for all variations of the method submitted by our team

| Run | $MAE$ | $MZOE$ | $MAE_{macro}$ | $GED$ | $RS$ | $ECS$ | $SCS$ | $WCS$ |
|---|---|---|---|---|---|---|---|---|
| baseline-all0s | 2.419 | **0.674** | 2.803 | 3.207 | 2.138 | 3.221 | 3.028 | 2.682 |
| baseline-all6s | 3.581 | 0.834 | 3.995 | 3.839 | 4.814 | 3.650 | 3.950 | 3.318 |
| baseline-average | **2.091** | 0.859 | 1.957 | 2.391 | **1.592** | 2.398 | 2.162 | **2.002** |
| distilroberta8 | 2.338 | <u>0.691</u> | 1.922 | 2.294 | 1.866 | 2.492 | 1.999 | 2.425 |
| mentalbert32 | 2.352 | 0.699 | 1.858 | **2.127** | 2.025 | **2.365** | 2.034 | 2.466 |
| mentalbert1epoch32 | 2.396 | 0.704 | 1.861 | 2.178 | <u>1.859</u> | 2.484 | 1.957 | 2.468 |
| mentalbert3epochs32 | 2.419 | 0.709 | 1.898 | 2.251 | 1.935 | 2.440 | 2.037 | 2.445 |
| mentalbert10epochs8 | 2.346 | 0.705 | 1.859 | 2.217 | 1.862 | 2.398 | **1.898** | 2.395 |
| mentalbert10epochs32 | <u>2.334</u> | 0.702 | **1.854** | 2.230 | 1.898 | 2.381 | 1.947 | <u>2.378</u> |
| mpnet32 | 2.408 | 0.696 | 1.936 | 2.365 | 2.048 | 2.536 | 1.985 | 2.414 |
| use32 | 2.347 | 0.696 | 1.975 | 2.534 | 1.911 | 2.443 | 2.215 | 2.494 |

at risk, we only keep the posts with a positive label for pretraining. We use the unsupervised MLM training implementation[1] provided in the Sentence-Transformers framework [30].

### 3.3. Final Classification

After obtaining the topic scores for all posts, we aggregate the probabilities at user level. Each user is assigned a feature vector of size $N \leq 100$, where $N$ is the number of extracted topics. Due to the small number of users, we need a solution that works well on low-dimensional data. The best fit is TabPFN [3], a transformer trained for supervised classification of tabular data, by approximating Bayesian inference on synthetic datasets drawn from causal priors. It achieves state-of-the-art performance on small datasets.

In order to predict the answers to the EDE-Q, each question is treated as an individual classification problem. We fit a TabPFN model on the feature vectors corresponding to the users in the training set, and learn to output an answer ranging from 1 to 6, which should correlate to the severity of the symptom targeted by the question, as experienced by the user.

## 4. Results

We report the results for our submitted runs in Table 1, based on the eight metrics used to evaluate all systems on the unannotated test data made available to participants [24].

The Mean Zero-One Error ($MZOE$) metric reflects the fraction of incorrect predictions for a user's questionnaire response, while the Mean Absolute Error ($MAE$) represents the average deviation from predicted values to the ground truth. The $MAE_{macro}$ is appropriate for imbalanced ordinal classification problems, as it computes the $MAE$ for each class and weighs the results equally. Since the measures evaluate performance at the user level, the reported result is averaged across all users in the dataset.

[1] https://github.com/UKPLab/sentence-transformers/blob/master/examples/unsupervised_learning/MLM

The Restraint Subscale ($RS$), Eating Concern Subscale ($ECS$), Shape Concern Subscale ($SCS$), and Weight Concern Subscale ($WCS$) are concerned strictly with the set of questions that address each symptom class. These metrics compute the $RMSE$ between the mean value of the responses filled in by the user for the corresponding questions and those outputted by the system. Based on the mean performance across the four subscale measures, a global score is obtained, which is then used in the Global ED ($GED$) metric, computed as the $RMSE$ between the ground truth global score and the model's global score.

The run names in Table 1 reflect the attempted variations on our method. We mainly experiment with different embeddings as input to the topic modeling pipeline. For MentalBERT, we submitted results for the default model, as well as after 1, 3, and 10 MLM pretraining epochs. We also specify the $N\_ensemble\_configurations$ hyperparameter for the final TabPFN classifiers, as we found, during validation, that tuning this parameter can control the model's tendency to skew the predictions towards either 0 or 6.

To ensure a fair comparison, we include the baseline performances, as reported in the task overview [24], covering three scenarios: predicting only 0, predicting only 6, and predicting the average user response. The best results are highlighted in bold. For metrics where either of the baselines is not surpassed, we underline the value for the model that came closest. The $MZOE$ all 0s and the $WCS$ average baselines were not outperformed by any participating system.

Overall, our best results are achieved with the MentalBERT model further pretrained on social media anorexia data. In particular, our *mentalbert10epochs* runs outperform the others on 4 out of the 9 metrics, while also achieving the highest $SCS$ score amongst all participating systems. This suggests that the embeddings from a model with specialized domain knowledge help identify topics informative enough to capture more intricate aspects of an eating disorder diagnosis, such as shape concern symptoms.

When it comes to the more general perspective captured by the $GED$, the MentalBERT with no additional pretraining performs best amongst our runs. This model also leads to the highest $ECS$, showing that even non-task specific mental health knowledge aids performance.

The DistilRoBERTa [31] sentence transformer [30] is the most competitive with the MentalBERT models, as it comes closest to outperforming the $MZOE$ all 0s baseline and is the best scoring run on this metric compared to other participating systems as well.

## 4.1. Qualitative Analysis

We further analyze the topics extracted by the customized BERTopic pipeline, when the posts in the conversation tree are embedded by the MentalBERT model with additional pretraining for 10 epochs. These topics represent the distribution before user level aggregation, and are indicative of general conversation trends present in the dataset. As expected due to the diverse nature of online conversations, many of the discovered topics are unrelated to eating disorders.

We provide relevant examples in Table 2. For anonymity purposes, we only include general examples and cluster names. We can see that our method successfully captures topics containing keywords associated with ED symptoms. As suggested by topic 80, the dataset also contains conversation surrounding recovery and treatment. Other adjacent mental health topics are identified as well, with keywords such as *bpd*, *autism* and *antidepressants*. We also observe that some clusters are formed around dieting, recipes or general mentions of food.

**Table 2**
Topic examples extracted with MentalBERT embeddings

| Topic name | Keywords | Frequency |
|---|---|---|
| 84_eating_body_purging_your | eating, body, purging, weight, food, eat | 143 |
| 83_bpd_mental_disorders_people | bpd, mental, disorders, people, autism | 64 |
| 80_you_re_your_ed | ed, recovery, eating, therapist | 23 |
| 69_ritalin_bupropion_effects_it | ritalin, bupropion, effects, antidepressants, take | 34 |
| 82_people_women_men | people, women, men, gender, fat, shaming | 67 |
| 42_vegan_impact_diet_vegetarian | vegan, impact, diet, vegetarian, eat | 42 |

## 5. Conclusion

We present a transformer-based topic modeling method to measure the severity of eating disorder symptoms, as implemented in our submission to CLEF's eRisk 2023 Task 3. We customize the BERTopic framework and obtain user-level topic distributions to be used as input features for downstream classification. To obtain more descriptive embeddings, we adapt the MentalBERT transformer to the Eating Disorder domain. We offer insight into the topics discovered by our model. Our ensembles reach the best performance on the Shape Concern Subscale and Mean Zero-One Error, even though all systems are below the baseline on the latter. The final results reflect the difficulty of estimating a person's answers to the EDE-Q and how more resources should be invested in the research of effective means of eating disorder symptom detection.

## 6. Acknowledgments

We would like to express our deepest gratitude to our coordinator, Ana-Sabina Uban, for the guidance offered throughout all stages of the development of this project.

## References

[1] M. O. C. G. Fairburn, Z. Cooper, Eating disorder examination edition 17.0d (2014). URL: https://www.credo-oxford.com/pdfs/EDE_17.0D.pdf.
[2] M. Grootendorst, Bertopic: Neural topic modeling with a class-based tf-idf procedure, arXiv preprint arXiv:2203.05794 (2022).
[3] N. Hollmann, S. Müller, K. Eggensperger, F. Hutter, Tabpfn: A transformer that solves small tabular classification problems in a second, 2022. arXiv:2207.01848.
[4] S. Zhou, Y. Zhao, R. Rizvi, J. Bian, A. F. Haynos, R. Zhang, Analysis of twitter to identify topics related to eating disorder symptoms, in: 2019 IEEE international conference on healthcare informatics (ICHI), IEEE, 2019, pp. 1–4.
[5] H. Yan, E. E. Fitzsimmons-Craft, M. Goodman, M. Krauss, S. Das, P. Cavazos-Rehg, Automatic detection of eating disorder-related social media posts that could benefit from a mental health intervention, International Journal of Eating Disorders 52 (2019) 1150–1156.

[6] J. Parapar, P. Martin-Rodilla, D. E. Losada, F. Crestani, Overview @erisk 2022: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2022, p. 233–256. URL: https://doi.org/10.1007/978-3-031-13643-6_18. doi:10.1007/978-3-031-13643-6_18.
[7] A. M. Mármol-Romero, S. M. J. Zafra, F. M. P. del Arco, M. D. Molina-González, M. T. M. Valdivia, A. Montejo-Ráez, SINAI at erisk@clef 2022: Approaching early detection of gambling and eating disorders with natural language processing, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 961–971. URL: https://ceur-ws.org/Vol-3180/paper-76.pdf.
[8] S. H. H. Saravani, L. Normand, D. Maupomé, F. Rancourt, T. Soulas, S. Besharati, A. Normand, S. Mosser, M. Meurs, Measuring the severity of the signs of eating disorders using similarity-based models, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 936–946. URL: https://ceur-ws.org/Vol-3180/paper-74.pdf.
[9] D. E. Losada, F. Crestani, J. Parapar, Overview of erisk: Early risk prediction on the internet (extended lab overview), in: L. Cappellato, N. Ferro, J. Nie, L. Soulier (Eds.), Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018, volume 2125 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018. URL: https://ceur-ws.org/Vol-2125/invited_paper_1.pdf.
[10] H. Srivastava, L. N. S, S. S, T. Basu, Nlp-iiserb@erisk2022: Exploring the potential of bag of words, document embeddings and transformer based framework for early prediction of eating disorder, depression and pathological gambling over social media, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 972–986. URL: https://ceur-ws.org/Vol-3180/paper-77.pdf.
[11] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423. doi:10.18653/v1/n19-1423.
[12] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022. URL: http://jmlr.org/papers/v3/blei03a.html.
[13] D. Angelov, Top2vec: Distributed representations of topics, ArXiv abs/2008.09470 (2020).
[14] R. Egger, J. Yu, A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts, Frontiers in Sociology 7 (2022).
[15] Q. Ng, C. Yau, Y. Lim, L. Wong, T. Liew, Public sentiment on the global outbreak of

monkeypox: An unsupervised machine learning analysis of 352,182 twitter posts, Public Health 213 (2022) 1–4.
[16] M. Falkenberg, A. Galeazzi, M. Torricelli, N. Di Marco, F. Larosa, M. Sas, A. Mekacher, W. Pearce, F. Zollo, W. Quattrociocchi, et al., Growing polarization around climate change on social media, Nature Climate Change (2022) 1–8.
[17] R. Ebeling, C. A. C. Sáenz, J. C. Nobre, K. Becker, Analysis of the influence of political polarization in the vaccination stance: the brazilian covid-19 scenario, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 16, 2022, pp. 159–170.
[18] Y. Hua, H. Jiang, S. Lin, J. Yang, J. M. Plasek, D. W. Bates, L. Zhou, Using twitter data to understand public perceptions of approved versus off-label use for covid-19-related medications, Journal of the American Medical Informatics Association 29 (2022) 1668–1678.
[19] Q. X. Ng, S. R. Lim, C. E. Yau, T. M. Liew, Examining the prevailing negative sentiments related to covid-19 vaccination: Unsupervised deep learning of twitter posts over a 16 month period, Vaccines 10 (2022) 1457.
[20] A. Baird, Y. Xia, Y. Cheng, Consumer perceptions of telehealth for mental health or substance abuse: a twitter-based topic modeling analysis, JAMIA open 5 (2022) ooac028.
[21] K. Wanchoo, M. Abrams, R. M. Merchant, L. Ungar, S. C. Guntuku, Reddit language indicates changes associated with diet, physical activity, substance use, and smoking during covid-19, Plos one 18 (2023) e0280337.
[22] S. Sarkar, A. Alhamadani, L. Alkulaib, C.-T. Lu, Predicting depression and anxiety on reddit: a multi-task learning approach, in: 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2022, pp. 427–435. doi:10.1109/ASONAM55673.2022.10068655.
[23] E. Jeon, N. Yoon, S. Y. Sohn, Exploring new digital therapeutics technologies for psychiatric disorders using bertopic and patentsberta, Technological Forecasting and Social Change 186 (2023) 122130.
[24] J. Parapar, P. Martin-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2023: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 14th International Conference of the CLEF Association, CLEF 2023, Springer International Publishing, 2023, p. 585–592.
[25] D. Cer, Y. Yang, S. yi Kong, N. Hua, N. L. U. Limtiaco, R. S. John, N. Constant, M. Guajardo-Céspedes, S. Yuan, C. Tar, Y. hsuan Sung, B. Strope, R. Kurzweil, Universal sentence encoder, in: In submission to: EMNLP demonstration, Brussels, Belgium, 2018. URL: https://arxiv.org/abs/1803.11175, in submission.
[26] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426 (2018).
[27] C. Malzer, M. Baum, A hybrid approach to hierarchical density-based cluster selection, in: IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, MFI 2020, Karlsruhe, Germany, September 14-16, 2020, IEEE, 2020, pp. 223–228. URL: https://doi.org/10.1109/MFI49285.2020.9235263. doi:10.1109/MFI49285.2020.9235263.
[28] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, E. Cambria, MentalBERT: Publicly available pretrained language models for mental healthcare, in: Proceedings of the Thirteenth Lan-

guage Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 7184–7190. URL: https://aclanthology.org/2022.lrec-1.778.

[29] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N. A. Smith, Don't stop pretraining: Adapt language models to domains and tasks, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8342–8360. URL: https://aclanthology.org/2020.acl-main.740. doi:10.18653/v1/2020.acl-main.740.

[30] 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, Association for Computational Linguistics, 2019. URL: https://arxiv.org/abs/1908.10084.

[31] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter, CoRR abs/1910.01108 (2019). URL: http://arxiv.org/abs/1910.01108. arXiv:1910.01108.