A Natural Language Processing Based Framework for Early Detection of Anorexia via Sequential Text Processing

Notebook for the BioNLP-IISERB Lab at CLEF 2024

Prateek Sarangi[1], Sumit Kumar[1], Shraddha Agarwal[1] and Tanmay Basu[1]

[1]*Department of Data Science and Engineering, Indian Institute of Science Education and Research, Bhopal, India*

Abstract

Task 2 of eRisk shared tasks in CLEF 2024 aims to develop text mining solutions for early prediction of anorexia using sequentially posted texts over social media. Anorexia is an eating disorder, a kind of mental illness, where people have distorted perceptions of their body weights and accordingly manipulate their food habits, which often results in deficient body weight. The aim here is to identify anorexia by processing the interactions over social media of an individual through text mining models. The organisers provided training data with ground truths for training the models and test data for evaluating their performance. The BioNLP research group at the Indian Institute of Science Education and Research Bhopal (IISERB) participated in Task 2 and submitted five runs for five text mining frameworks. Five different classifiers and individual feature engineering techniques were used to develop frameworks. The bag-of-words model and transformer-based embedding were used as features for individual classifiers. The performance of multiple classifiers was evaluated using the training corpus. Then Random Forest, Adaptive Boosting, Logistic Regression, Support Vector Machine (SVM), and Longformer classifiers were chosen to run on the test set. Experimental results show that SVM and AdaBoost classifiers using the TF-IDF-based weighting schemes achieved the highest precision score among all submitted runs in Task 2 of eRisk 2024. However, the performance of our models in terms of the other metrics, like recall, f-score, ERDE, etc., is not reasonably good compared to the other runs. Hence, we plan to develop transformer-based embeddings from scratch using data collected from multiple social media platforms.

Keywords

BioNLP, Information Extraction, Text Classification, Mental Health, Anorexia Detection

1. Introduction

The rise of social media has revolutionised communication and opened up new opportunities for sociological and psychological research, particularly in mental health [1]. Anorexia nervosa, a severe eating disorder characterised by an intense fear of weight gain and a distorted body image leading to self-starvation and significant health issues, is a crucial area of focus. Social media's widespread use provides a unique, real-time view into behaviours and expressions that may be useful to identify such mental disorders[2, 3, 4]. The vast amount of multiple generated content on platforms like Facebook, Twitter, and Reddit offers an extensive data pool for detecting early signs of mental health conditions like anorexia [5]. Such contents, including written posts, comments, shared photos, and likes, can be analysed for patterns indicating the onset of anorexia, which may be helpful for early interventions [6]. Recognising social media's potential as a public health tool, the Conference and Labs of the Evaluation Forum (CLEF) launched the eRisk initiative in 2018 to explore this potential. Initially focused on depression, the initiative later included anorexia, showcasing a broader commitment to using big data for mental health monitoring and intervention.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

✉ sarangiprateek80@gmail.com (P. Sarangi); sumitkumar9297@gmail.com (S. Kumar); shraddhaagarwal2001@gmail.com (S. Agarwal); tanmay@iiserb.ac.in (T. Basu)

🌐 https://www.linkedin.com/in/prateeksarangi/ (P. Sarangi); https://www.linkedin.com/in/sumit-kumar-787203178/ (S. Kumar); https://www.linkedin.com/in/shraddha-agarwal-98743320a/ (S. Agarwal); https://sites.google.com/view/tanmaybasu/ (T. Basu)

iD 0009-0009-3729-7833 (P. Sarangi); 0009-0008-2069-3082 (S. Kumar); 0009-0001-1395-4570 (S. Agarwal); 0000-0001-9536-8075 (T. Basu)

In 2024, the CLEF eRisk second shared task highlights the importance of sequential text processing to identify anorexia markers as they appear in social media posts [7]; in the way the mental health professionals analyse patient behaviours over a while for treating a mental illness [8]. This method aids not only in early detection but also in understanding the progression of anorexia through digital footprints. The Bio-NLP group at the Indian Institute of Science Education and Research Bhopal (IISERB) has developed comprehensive text mining frameworks for this task. These frameworks explore text feature extraction methods like transformer-based embeddings and classical bag-of-words models for semantic interpretation of social media text to identify the indicators of anorexia [9]. Subsequently, various text classifiers viz. Random Forest, Adaptive Boosting, Logistic Regression, SVM, and Longformer were trained on multiple features derived from the text data to categorise the posts to either anorexia or control group. The proposed frameworks are presented in section 2 of the paper. The organisers provided a training corpus with ground truths for developing the models and a separate test corpus for evaluating the performance of the submitted runs. The empirical analysis of the proposed frameworks on both the training and test corpora are provided in section 3.3 in terms of the evaluation metrics provided by eRisk organisers, like ERDE, Latency, F-score, etc. [10, 11].

The experimental evaluation shows that two of our runs containing Adaptive Boosting and SVM classifiers using the TF-TDF model achieve the first and second ranks in precision among all the runs submitted for this task. Moreover, our Longformer model achieves the best latency and speed, and another run having a Logistic Regression and Entropy-based bag of words model achieves the best speed among all 44 runs submitted for task 2 of eRisk 2024. However, our models could not achieve reasonable recall, F1 and ERDE scores compared to many other runs submitted for this task. Our models show poor performance for ranking-based evaluations. Hence, we need to investigate this direction to improve the performance. We plan to train a transformer-based model from scratch using the data collected from various social media sites like Reddit and Twitter to identify the subtle nuances in the sequential texts processed over time. Furthermore, in future, we need to consider the timestamp of the posts as a significant indicator in the model for semantic interpretation of the sequential texts.

2. Proposed Frameworks

In the effort to identify early signs of anorexia through sequential text processing of social media conversations, particularly from Reddit, we have developed various text classification frameworks. Our methodologies are designed to harness the vast amount of unstructured text in these digital interactions, formatted in XML and provided by the organisers of the eRisk 2024 challenge.

2.1. Feature Engineering

The feature engineering approaches we have explored are crafted to capture the intricate linguistic and semantic patterns from sequential social media texts to identify the signs of anorexia. Both classical Bag of Words (BoW)—-based features and recent transformer-based embeddings are used to train the classification models. The following three feature selection techniques generated features from the given training corpus.

2.1.1. TF-IDF Weighting Scheme of BoW

Each document, representing a user's aggregated posts, is converted into a vector with each unique term as a feature. This model provides a fundamental basis for more advanced analyses. The BoW approach[12, 13] considers each unique term as a feature, known as unigram, to constitute the set of features of a text collection known as vocabulary. Sometimes, it considers two, three, or multiple terms as one feature based on the significance of the text sequence, known as bigram, trigram, or n-grams, respectively. In the experimental analysis, we explored the classifiers' performance using unigrams, bigrams and trigrams. After generating the dictionary of a corpus, the Term Frequency-Inverse

Document Frequency (TF-IDF) weighting is used to develop the vector of a given text[14, 15, 16, 17]. TF counts the number of terms in a given text, whereas IDF of a term, say, t, is defined as

$$\text{IDF}(t) = \log\left(\frac{N}{DF(t)}\right)$$

Where N is the total number of texts in the text collections, and $DF(t)$ is the number of texts in a corpus containing the term t.

2.1.2. Entropy Based Weighting Scheme of BoW

This scheme assigns weights to the terms in a document based on their entropy, which measures the amount of information or uncertainty associated with each term. Many researchers use the entropy-based term weighting technique to form a term-document matrix from a text collection [13, 15, 16, 17, 18]. This method is developed in the spirit that the more important term is the more frequent one that occurs in fewer documents, taking the distribution of the term over the corpus into account [15]. The weight of a term in a document is determined by the entropy of the term frequency of the term in that document [15]. The weight (W_{ij}) of the i^{th} term in the j^{th} document is defined by the Entropy[1] [15, 16] model as follows:

$$W_{ij} = \log\left(tf_{ij} + 1\right) \times \left(1 + \frac{\sum\limits_{j=1}^{N} P_{ij}\log P_{ij}}{\log(N+1)}\right), \quad \text{where,} \quad P_{ij} = \frac{tf_{ij}}{\sum\limits_{j=1}^{N} tf_{ij}} \tag{1}$$

Here, N is the total number of documents in the corpus, and tf_{ij} is the frequency of i^{th} term in the j^{th} document of the corpus. Generally, the BoW model generates many terms, making the term-document matrix sparse and high dimensional, which can badly affect the performance of the text classifiers [17]. Hence, χ^2-statistic-based term selection technique was used for both TF-IDF and Entropy-based term weighting schemes in the experiments to identify essential terms from the term-document matrix, which is a widely used technique for term selection [13, 16, 19].

2.1.3. Transformer-Based Embeddings

Utilising cutting-edge transformer architectures like Longformer[2], we create dense vector representations of text that capture deep contextual nuances far beyond traditional models. Bidirectional Encoder Representations from Transformers (BERT) is a contextualised word representation model based on a masked language model and pre-trained using bidirectional transformers on general domain corpora, i.e., English Wikipedia and books [20]. The Longformer model[21] was chosen because it performs better than BERT at understanding long-term relationships in texts[17]. It creates feature embeddings to help detect early signs of anorexia by recognising language patterns, from explicit mentions of body image issues to under-expressed signs of distress.

2.2. Text Classification Methods

Our approach to text classification combines engineered features with several machine learning algorithms chosen for their ability to handle complex relationships within high-dimensional data. We use Support Vector Machines (SVM) with linear and RBF kernels for their efficiency in high-dimensional spaces and binary classification tasks, finding the optimal hyperplane to separate classes and maximising the margin between them. This method is particularly effective for its robustness against overfitting [22]. Logistic Regression (LR), a probabilistic model, is also employed with L1 and L2 regularisation due

[1] https://radimrehurek.com/gensim/models/logentropy_model.html

[2] https://huggingface.co/allenai/longformer-base-4096

to its interpretability and effectiveness for binary outcomes [23]. LR uses the probability of a particular class by fitting a logistic function to the data, providing precise, interpretable coefficients for each feature, which helps understand the influence of different features on the likelihood of anorexia.

In addition, we have used Adaptive Boosting (AB), which improves classification accuracy by combining multiple weak classifiers, such as decision trees, in our model. AB sequentially applies a weak classifier to the data, adjusting the weights of misclassified instances so subsequent classifiers focus more on difficult cases, enhancing performance on complex textual data [24]. Furthermore, we incorporate Longformer, a Transformer-based model, which is pre-trained on large datasets and fine-tuned on our specific dataset [21]. This model is used to understand more extended context and relationships between words in a text to capture language patterns in user interactions.

By training these classifiers on the features derived from the textual data, we aim to identify individual posts indicating the risk of anorexia. We systematically analyse the performance of these models on training data and validate them on unseen test data to develop a scalable and effective tool for the early detection of anorexia. The SVM, Logistic Regression, and AdaBoost models were implemented using Scikit-learn[3], while the transformer-based models were fine-tuned using the Hugging Face Transformers library [25]. In Section 3.3, we present the experimental results, showcasing the efficacy of our frameworks for identifying signs of anorexia through social media interactions.

3. Experimental Analysis

The dataset provided by the eRisk 2024 challenge organisers includes 2,335 files, each containing a collection of Reddit posts from an individual user. In XML format, these files include metadata such as post dates and titles. Among these, 2,062 files are labelled as class "0" (no signs of anorexia, i.e., control group), while 273 are labelled as class "1" (potential signs of anorexia).

3.1. Data Preprocessing

For our analysis, we extracted and merged all conversations of each subject from the XML files. It was observed that the text field was empty in some files, but the title field had vital text, ensuring that either had the data to be considered in such cases instead of just concentrating solely on the text field. This preprocessed data was input for our feature engineering and classification pipelines.

3.2. Experimental Settings

To assess the performance of our frameworks on the raining corpus, we utilised several evaluation metrics commonly employed in binary classification tasks, including precision, recall and F1-score. Stratified 5-fold cross-validation has been conducted on the training dataset to evaluate our models' performance and tune the hyperparameters. Grid search[4] and random search techniques were used to identify the optimal hyperparameter settings for each model following the cross-validation scores. The best-performing models from the cross-validation stage were then tested on the unseen test dataset. The results of this evaluation are presented and discussed in Section 3.3.

3.3. Results and Discussion

The BioNLP-IISERB team's participation in the eRisk 2024 challenge aimed to detect anorexia early via sequential text processing. The approach involved several experimental runs tailored to explore the optimal parameters and methodologies.

Table 2 shows the experimental comparison of various Bag of Words models across different classifiers on a training corpus. Support Vector Machine (SVM) with TF-IDF feature selection emerged as the top performer, achieving the highest precision (0.9639) and F1 score (0.8009). Logistic Regression (LR)

[3] http://scikit-learn.org/stable/supervised_learning.html

[4] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

Table 1
BioNLP-IISERB team results for eRisk 2024 Challenge Task 2

Team Details	
Number of runs	5
Number of user writings processed	10
Processing time	09:39 (from first to last response)

Table 2
Performance of Various Bag of Words Models on Different Classifiers on Training Corpus

Classifier	Feature Selection	# Terms	Recall	Precision	F1
LR	Entropy	2500	0.5128	0.7650	0.6140
	TF-IDF	10000	**0.6777**	0.8685	**0.7613**
RF	Entropy	500	0.1465	0.5714	0.2332
	TF-IDF	1000	0.5971	0.7689	0.6722
SVM	Entropy	1000	0.3626	**0.9802**	0.5294
	TF-IDF	10000	**0.6850**	**0.9639**	**0.8009**
Adaboost	Entropy	500	0.3993	0.7365	0.5178
	TF-IDF	1500	0.6410	0.8102	0.7157
LongFormer	-	-	0.0741	0.0606	0.0667

also performed strongly, particularly with TF-IDF, achieving the highest recall (0.6777) among all models. The Adaboost classifier showed moderate performance, while Random Forest models generally underperformed. Surprisingly, the LongFormer model showed remarkably poor performance across all metrics and needed further investigation.

A clear trend can be seen in the superiority of TF-IDF over Entropy-based feature selection across all traditional classifiers, often by a significant margin. This suggests that TF-IDF's ability to capture both term importance and distinctiveness is precious for this classification task. The optimal number of terms varied across classifiers, with SVM and LR performing best with 10,000 terms, while RF and Adaboost achieved their best performance with fewer terms (1000 and 1500, respectively). These results highlight the importance of classifier selection and feature engineering in text classification tasks, with SVM and LR coupled with TF-IDF emerging as strong baseline models for similar problems.

Table 3
Decision-Based Evaluations of Proposed Frameworks on Test Data

Runs	P	R	F1	ERDE$_5$	ERDE$_{50}$	latency$_{TP}$	speed	latency weighted F1
BioNLP-IISERB 0 (Entropy+BoW+LR)	0.53	0.23	0.32	0.10	0.09	2.00	1.00	0.32
BioNLP-IISERB 1 (TF-IDF+BoW+LR)	0.54	0.75	0.62	0.08	0.04	4.00	0.99	0.62
BioNLP-IISERB 2 (Longformer)	0.58	0.16	0.25	0.10	0.10	1.00	1.00	0.25
BioNLP-IISERB 3 (TF-IDF+BoW+SVM)	0.67	0.51	0.58	0.08	0.06	3.00	0.99	0.58
BioNLP-IISERB 4 (TF-IDF+BoW+AB)	0.73	0.62	0.67	0.08	0.05	4.00	0.99	0.66

The performance exhibited in Table 3 with considerable variation across multiple experimental runs. Run 1 (TF-IDF+BoW+LR) demonstrated superior performance with a high F1 score (0.62), indicating an optimal balance between precision and recall. Run 3 and run 4 showed consistent performance, both achieving an F1 score of 0.58 and 0.67, respectively. Notably, these runs also exhibited the lowest ERDE50 values (0.08), suggesting enhanced early recognition capabilities.

In contrast, run 0 (Entropy+BoW+LR) and run 2 (Longformer) obtained F1 scores of 0.32 and 0.25,

Table 4
Ranking Based Performance of Proposed Frameworks on Test Data

Writings	Metrics	BioNLP-IISERB0 (Entropy + SVM)	BioNLP-IISERB1 (TFIDF + SVM)	BioNLP-IISERB2 (Longformer)	BioNLP-IISERB3 (Entropy + AdaBoost)	BioNLP-IISERB4 (TFIDF + AdaBoost)
1	P@10	0.10	0.00	0.00	0.10	0.20
	NDCG@10	0.19	0.00	0.00	0.06	0.21
	NDCG@100	0.06	0.07	0.05	0.09	0.10
100	P@10	0.00	0.00	0.00	0.00	0.00
	NDCG@10	0.00	0.00	0.00	0.00	0.00
	NDCG@100	0.00	0.00	0.00	0.00	0.00
500	P@10	0.00	0.00	0.00	0.00	0.00
	NDCG@10	0.00	0.00	0.00	0.00	0.00
	NDCG@100	0.00	0.00	0.00	0.00	0.00
1000	P@10	0.00	0.00	0.00	0.00	0.00
	NDCG@10	0.00	0.00	0.00	0.00	0.00
	NDCG@100	0.00	0.00	0.00	0.00	0.00

respectively. These runs also had elevated ERDE50 values (0.10 and 1.00), indicating reduced effectiveness in early correct decision recognition. Table 3 further demonstrates several strengths of the proposed approaches, such as runs 1, run 3 and run 4 exhibited high precision balanced with satisfactory recall, resulting in robust F1 scores. Moreover, run 3 and run 4 showed lower latencies (0.99 and 0.99), indicating efficient decision-making processes. The consistent performance in Run 3 (TF-IDF+BoW+SVM) and 4 (TF-IDF+BoW+AB) suggests a reliable and stable approach.

However, certain limitations were observed where run 0 and run 2 displayed notably lower recall values (0.23 and 0.16), adversely affecting their overall F1 scores. Additionally, run 0 and run 2 exhibited higher latency, potentially due to inefficiencies in the early stages of model deployment.

The ranking-based evaluations in Table 4 depicted a decline in performance metrics as the number of processed writings escalated. Initially, promising detection capabilities at smaller datasets experienced a stark reduction in precision and NDCG scores, diminishing to zero as the data volume increased. This suggests that while the initial models perform well under control, smaller datasets, the effectiveness wanes significantly under larger scales, pointing to potential overfitting or the need for more robust generalisation capabilities in the models.

4. Conclusion

The BioNLP@IISERB team took part in the eRisk 2024 task 2 to develop robust text mining frameworks for early prediction of anorexia by analysing social media texts. The classical bag-of-words models and recent transformer-based methods were explored to generate potential features for identifying nuances in the given texts. Subsequently, different text classifiers were trained using these features to identify anorexia from the given social media texts. The experimental results suggests that these frameworks are capable of identifying textual patterns indicative of anorexia, however, there are rooms for further improvements. Some frameworks showed promising precision and recall scores on more minor texts, but we encountered substantial challenges in maintaining consistent performance as the volume of data increased. This suggests the need for continued research and refinement of our methods. By processing and analysing large-scale social media data, we can extract valuable insights into the onset and progression of conditions like anorexia, thus enabling timely and targeted support. In the future, more robust and dynamic models will be developed that can efficiently handle the growing volume of user-generated content while maintaining high accuracy in detection. Additionally, exploring multi-modal approaches that combine textual data with other types of information, such as images and social network data, could provide more comprehensive and nuanced insights into mental health conditions. In conclusion, by successfully applying these techniques, we would have a significant positive impact, providing early resilience and support for individuals dealing with anorexia and other mental health challenges. Future work should focus on stabilising recall performance and reducing latency to consistently enhance decision-making speed and accuracy.

Acknowledgements

Prateek Sarangi and Tanmay Basu acknowledge the support of the seed funding (INST/DSE/2023-2024/18) provided by the Indian Institute of Science Education and Research Bhopal, India.

References

[1] W. Ragheb, J. Azé, S. Bringay, M. Servajean, Attentive multi-stage learning for early risk detection of signs of anorexia and self-harm on social media., in: CLEF (Working Notes), 2019.

[2] M. De Choudhury, M. Gamon, S. Counts, E. Horvitz, Predicting depression via social media., ICWSM 13 (2013) 1–10.

[3] M. De Choudhury, S. Counts, E. Horvitz, Social media as a measurement tool of depression in populations, in: Proceedings of the 5th Annual ACM Web Science Conference, ACM, 2013, pp. 47–56.

[4] L. G. et al., Machine learning and natural language processing in mental health: systematic review, Journal of Medical Internet Research, 23 (2021) e15708.

[5] S. Paul, S. K. Jandhyala, T. Basu, Early detection of signs of anorexia and depression over social media using effective machine learning frameworks., in: CLEF (Working notes), 2018.

[6] B. W. Eidem, F. Cetta, J. L. Webb, L. C. Graham, M. S. Jay, Early detection of cardiac dysfunction: use of the myocardial performance index in patients with anorexia nervosa, Journal of adolescent health 29 (2001) 267–270.

[7] J. Parapar, P. M. Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2024: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, 15th International Conference of the CLEF Association, CLEF 2024, Springer International, Grenoble, France, 2024.

[8] A. Ranganathan, A. Haritha, D. Thenmozhi, C. Aravindan, Early detection of anorexia using rnn-lstm and svm classifiers., in: CLEF (Working Notes), 2019.

[9] F. Galetta, F. Franzoni, A. Cupisti, E. Morelli, G. Santoro, F. Pentimone, Early detection of cardiac dysfunction in patients with anorexia nervosa by tissue doppler imaging, International journal of cardiology 101 (2005) 33–37.

[10] J. Parapar, P. M. Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2024: Early risk prediction on the internet (extended overview), in: Working Notes of the Conference and Labs of the Evaluation Forum CLEF 2024, CLEF 2024, CEUR Workshop Proceedings, Grenoble, France, 2024.

[11] M. Marion, S. Lacroix, M. Caquard, L. Dreno, P. Scherdel, C. G. L. Guen, E. Caldagues, E. Launay, Earlier diagnosis in anorexia nervosa: better watch growth charts!, Journal of eating disorders 8 (2020) 1–9.

[12] G. Salton, M. J. McGill, Introduction to Modern Information Retrieval, McGraw Hill, 1983.

[13] T. Basu, S. Goldsworthy, G. V. Gkoutos, A sentence classification framework to identify geometric errors in radiation therapy from relevant literature, Information 12 (2021) 139.

[14] A. Selamat, S. Omatu, Web page feature selection and classification using neural networks, Information Sciences 158 (2004) 69–88.

[15] T. Sabbah, A. Selamat, M. H. Selamat, F. S. Al-Anzi, E. H. Viedma, O. Krejcar, H. Fujita, Modified frequency-based term weighting schemes for text classification, Applied Soft Computing 58 (2017) 193–206.

[16] T. Basu, G. V. Gkoutos, Exploring the performance of baseline text mining frameworks for early

prediction of self harm over social media., in: Proceedings of International Conference of CLEF Association, 2021, pp. 928–937.

[17] H. Srivastava, N. S. Lijin, S. Sruthi, 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: Proceedings of Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, Bologna, Italy, 2022.

[18] S. Goswami, S. Pal, S. Goldsworthy, T. Basu, An effective machine learning framework for data elements extraction from the literature of anxiety outcome measures to build systematic review, in: Business Information Systems: 22nd International Conference, BIS 2019, Seville, Spain, June 26–28, 2019, Proceedings, Part I 22, Springer, 2019, pp. 247–258.

[19] T. Basu, C. Murthy, A supervised term selection technique for effective text categorization, International Journal of Machine Learning and Cybernetics 7 (2016) 877–892.

[20] 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).

[21] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, arXiv preprint arXiv:2004.05150 (2020).

[22] C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (1995) 273–297.

[23] D. W. Hosmer, S. Lemeshow, R. X. Sturdivant, Applied logistic regression, volume 398, John Wiley & Sons, 2013.

[24] Y. Freund, R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, Journal of computer and system sciences 55 (1997) 119–139.

[25] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 38–45.