# ARC-NLP at PAN 2023: Transition-Focused Natural Language Inference for Writing Style Detection

Izzet Emre Kucukkaya[1], Umitcan Sahin[1] and Cagri Toraman[1]

[1]*Aselsan Research Center, 06378, Ankara, Turkey.*

## Abstract

The task of multi-author writing style detection aims at finding any positions of writing style change in a given text document. We formulate the task as a natural language inference problem where two consecutive paragraphs are paired. Our approach focuses on transitions between paragraphs while truncating input tokens for the task. As backbone models, we employ different Transformer-based encoders with warmup phase during training. We submit the model version that outperforms baselines and other proposed model versions in our experiments. For the easy and medium setups, we submit transition-focused natural language inference based on DeBERTa with warmup training, and the same model without transition for the hard setup.

## Keywords

Multi-author, natural language inference, transition, writing style detection.

## 1. Introduction

Multi-author writing style analysis is an important task due to recent advancements in AI-assisted conversational systems and chatbots. Moreover, it can be used for author verification and plagiarism detection. In PAN 2023 [1], participants are asked to find all positions of writing style change on the paragraph-level for a given text [2]. In contrast to previous years of PAN, in this year, the style change happens in the same topic, which makes it challenging to detect writing style change by topic.

In this study, as Aselsan Research Center - Natural Language Processing team (ARC-NLP), we propose transition-focused natural language inference (NLI) for multi-author writing style detection. We formulate the task as a natural language inference problem where two consecutive paragraphs are paired and learning objective is to label whether they are written by the same author. We choose this approach since we already demonstrated that the NLI method performs well in similar tasks [3]. As backbone models, we employ different Transformer-based language models with warmup phase during training (e.g. BERT-like encoders [4]). Since the models have a limited length of input sequence (e.g. 512 tokens), we need to truncate the input paragraphs. For truncation, we focus on transitions between paragraphs, since transitions provide logical connections between paragraphs in documents. We are requested to send our models and algorithms as a docker image to the TIRA [5] system where the test set evaluation is made.

<table>
<tr><td>Author 1</td><td>I'm not arguing with you here, I'm simply trying to contextualize this for you. To the extent that they are there, it is with your consent. The state has passed laws making sure that vulnerable people (not saying he's one) don't get abused (not saying you're abusing him), and in casting a wide net to save as many vulnerable little birds as possible from hitting the floor after being kicked out of their nest wrongfully, the state has (as much from a lack of better options as from any other reason) created a circumstance where occasionally some not-so-vulnerable little bird can take advantage of someone else's nest.</td></tr>
<tr><td>Author 2</td><td>He's at my place half the time and his fiance(e)'s place the other half of the time. He's been (homeless) couch surfing for several years and only recently got engaged to his other partner. We don't have any current issues that would lead me to want this arrangement to stop, but I do want to protect my own legal rights.</td></tr>
<tr><td>Author 2</td><td>I don't think he, in particular, would do that, but I do want to retain my own legal rights wherever possible/appropriate.</td></tr>
</table>

**Figure 1:** An example of writing style change. The labels are [1, 0] for this instance, representing writing style change in two consecutive paragraphs.

## 2. Related Work

In PAN 2018 [6], the task is basically binary classification. Zlatkova et al. [7] develop an ensemble approach of the models including SVM, Random Forest, LightGBM etc. Hosseinia and Mukherjee [8] use parallel attention networks to focus on the hierarchical structure of the language.

In PAN 2019 [9], the task is to detect the number of authors in a given document. Nath [10] use two clustering algorithms based on the threshold and window merge. In addition, Zuo et al. [11] use K-means and hierarchical clustering algorithms.

In PAN 2020 [12], the task is to detect style changes between two consecutive paragraphs. Castro-Castro et al. [13] use a paragraph representation based on character, lexical, and syntactic features in a clustering algorithm. Iyer and Vosoughi [14] use a pre-trained BERT model, and train a random forest classifier of the embedding representation generated from the BERT model.

In PAN 2021 [15], the task is to determine the number of authors, and locate specific author changes. Strøm [16] apply a stacking ensemble on text embeddings. Deibel and Löfflad [17] use an LSTM-based algorithm.

For the task of the last year, in PAN 2022 [18], the winning solution Lin et al. [19] employ an ensemble of three Transformer-based language models using majority voting to obtain the final prediction. Furthermore, Jiang et al. [20] use base and large versions of the ELECTRA model, and report highly challenging scores.

**Table 1**
Number of instances on the train and validation sets.

| Difficulty | Train | Validation |
|---|---|---|
| Easy | 4200 | 900 |
| Medium | 4200 | 900 |
| Hard | 4200 | 900 |

**Table 2**
Number of pairs for each label. The 0 and 1 labels indicate that if the style change occurs.

| Difficulty | Train | | Validation | |
|---|---|---|---|---|
| | 0 | 1 | 0 | 1 |
| Easy | 1557 | 11347 | 377 | 2451 |
| Medium | 15001 | 13215 | 4013 | 3029 |
| Hard | 10092 | 9021 | 2159 | 1953 |

## 3. Task

Participants are asked to solve the the intrinsic style change detection task. For a given text, we find all positions of writing style change on the paragraph-level. For example, the document in Figure 1 is written by two authors, and there is a style change between first and second paragraph. The label of this transition is specified as 1. Furthermore, there is no style change between second and third paragraph where the label is 0. This example is chosen from the Easy split of the dataset. There are three difficulty levels:

- **Easy**: The paragraphs of a document consist of various number of topics.
- **Medium**: The topical variety is small. The need of the style detection instead of topic detection increases.
- **Hard**: All paragraphs in a document are on the same topic.

All the documents in this task are in English, and contain different numbers of style changes and authors. Furthermore, writing style change only occurs in paragraph level. There is no need to investigate sentences separately.

## 4. Dataset

In this task, there are three different difficulty levels with their own train-validation-test sets. The numbers of documents on the train and validation splits are the same in all of the three difficulty levels, which are provided in Table 1.

The total number of the 0 and 1 labels in documents are reported in Table 2. These numbers indicate the number of samples in the natural language inference task derived from the actual task as mentioned in Section 5.

# 5. Proposed Method: Transition-Focused Natural Language Inference



**Figure 2:** An illustration of the proposed approach that uses natural language inference with transition-focused truncation.

## 5.1. Main Approach

In this work, we formulate the task as natural language inference (NLI). To do so, we employ a Transformer-based language model that is based on the encoder structure. We prepare input by concatenating consecutive paragraphs and using the SEP token between them. We then place a binary classification layer of the CLS embedding, based on whether if the style change occurs (1) or not (0) between these to paragraphs.

Since the models have a limited length of input sequence (i.e. 512 tokens for BERT, RoBERTa, and DeBERTa; and 1024 tokens for BigBird), we need to truncate the input paragraphs before training NLI. For truncation, we focus on transitions between paragraphs (we refer it to as **Transition-Focused Truncation**), since transitions provide logical connections between paragraphs in documents. We also provide results for default truncation that focuses on the beginning of text (we refer it to as **Longest First Truncation**). The proposed NLI model and truncation approaches are illustrated in Figure 2. When input sequence length is 512, the last 256 tokens of the first paragraph and the first 256 tokens of the second paragraph are combined in transition-focused truncation as in Figure 2a, while first 256 tokens from both paragraphs are truncated as in Figure 2b.

## 5.2. Backbone Models

As the text encoder, we employ several Transformer-based language models in our preliminary experiments. Here, we report the highest performing four models.

**BERT [4]**    Bidirectional encoder representation for transformers, BERT, is an encoder architecture that utilizes an attention mechanism. It was pretrained on masked language modelling and next sentence prediction tasks.

**RoBERTa [21]**    A robust optimized BERT pre-training approach, RoBERTa, has the same architecture as BERT. However, the pretraining task of the next sentence prediction is removed. Furthermore, it has dynamically changing masking pattern applied to the training data with larger training batches.

**DeBERTa v3 [22]**    Decoding-enhanced BERT with Disentangled Attention, DeBERTa, has two additional techniques compared to the BERT, distangled attention and enhanced mask decoder. Due to the new adjustments, they state that it outperforms BERT and the other state-of-art models in many tasks.

**BigBird-RoBERTa [23]**    BigBird has a sparse attention mechanism that reduces this quadratic dependency to linear which enables it to handle sequences of length up to 8 times of what was previously possible using similar hardware. Since paragraphs can be too long in this task, we employ this model to cover more number of tokens in input.

### 5.3. Warmup

In preliminary experiments, we realize that our models converge different minima in the training of this task. In order to overcome this issue, we use the warmup with the warmup ration is 0.1. In warmup steps, the model trains with a very low learning rate and tries to find the global minima of the loss function, and hinders the inaccurate convergence.

## 6. Experiments

### 6.1. Experimental Setup

The input length in BigBird is 1024 tokens (512 for each paragraph), while 512 tokens for other models (256 for each paragraph). We set the following hyperparameters. Learning rate is 5e-5, number of epochs is 5, and batch size is 4. We used 3 NVIDIA GeForce RTX 2080 GPUs in training. The pre-trained models and the trainer framework are obtained from the HuggingFace library [24].

For evaluating model performances, we calculate Macro F1 Scores using the official evaluation script[1].

### 6.2. Baselines

We implement two baseline methods to compare with our approach.

**Random**    The output label array is generated randomly by sampling from 0 and 1, uniformly.

---

[1]https://github.com/pan-webis-de/pan-code/tree/master/clef23/multi-author-analysis

**Table 3**

Macro F1 Scores on the validation set. We submitted the highest performing models on the validation set given in bold. The test (leaderboard) results are given at the end of the table.

| Model | Easy | Medium | Hard |
|---|---|---|---|
| random | 0.401 | 0.504 | 0.505 |
| tfidf | 0.551 | 0.615 | 0.531 |
| bert-base | 0.980 | 0.807 | 0.753 |
| roberta-base | 0.790 | 0.363 | 0.344 |
| deberta-v3 | 0.966 | 0.591 | 0.344 |
| bigbird-roberta-base | 0.965 | 0.400 | 0.344 |
| bert-base-warmup | 0.982 | 0.786 | 0.755 |
| roberta-base-warmup | 0.662 | 0.766 | 0.643 |
| deberta-v3-warmup | 0.984 | 0.798 | **0.770** |
| bigbird-roberta-base-warmup | 0.962 | 0.805 | 0.742 |
| bert-base-transition | 0.918 | 0.766 | 0.344 |
| roberta-base-transition | 0.976 | 0.363 | 0.344 |
| deberta-v3-transition | 0.464 | 0.778 | 0.344 |
| bigbird-roberta-base-transition | 0.969 | 0.796 | 0.724 |
| bert-base-warmup-transition | 0.958 | 0.772 | 0.660 |
| roberta-base-warmup-transition | 0.981 | 0.804 | 0.344 |
| deberta-v3-warmup-transition | **0.987** | **0.812** | 0.746 |
| bigbird-roberta-base-warmup-transition | 0.958 | 0.804 | 0.759 |
| *Submitted models on the test set* | *0.982* | *0.810* | *0.772* |

**TF-IDF**    We use TF-IDF term weighting [25] to extract features using the English stopwords of NLTK library [26]. Additional features such as number of question marks, periods, apostrophes, parenthesis, and words are concatenated to the feature vector. Finally, we concatenate the TF-IDF feature vectors of consecutive two paragraphs, and train Support Vector Classifier (SVC) [27] for classification.

## 6.3. Experimental Results

We report the model performances on the validation splits (see Section 4) in Table 3. We divide the table into six parts. At the top, we provide the baseline scores. The second part consists of our proposed approach with four backbone models, described in Section 5. We use the same models with warmup during training in the third part. So far, we do not employ transition-focused truncation. Next, we provide the results of transition-focused truncation and lastly with warmup as well. In the last part, we provide the performance scores of the submitted models on the test set (leaderboard). We submitted the highest performing models on the validation set (given as bold). We have the following observations.

- Baseline models perform poor as expected. TF-IDF is based on bag-of-words model, which can show that writing style can not be detected by syntactical writing features.
- DeBERTa is the highest performing backbone model for NLI in all setups.
- Using warmup during training can increase the performance in some cases, specifically for the medium and hard setups.

- Transition-focused truncation method improves the results in some cases. More importantly, we obtain the highest scores on the validation set for the easy and medium setups when we employ transition-focused DeBERTa with warmup. For the hard setup, the same model with default truncation performs highest. We submitted the highest performing methods to the shared task.

## 7. Conclusion

In this paper, we propose transition-focused natural language inference (NLI) for multi-author writing style detection. We truncate the input paragraphs by focusing on transitions between paragraphs, since transitions provide logical connections between paragraphs in documents. Transition-focused NLI performs highest in easy and medium setups. Moreover, we obtain the highest performances when backbone model is DeBERTa in all setups. We submitted the highest performing models on the validation set. Our models are placed in the second place for all subtasks (easy, medium, and hard) in the leaderboard.

As a future work, there can be some improvements to overcome the class imbalance problem. Furthermore, other large language models can be employed to encode the embedding vectors.

## References

[1] J. Bevendorff, I. Borrego-Obrador, M. Chinea-Ríos, M. Franco-Salvador, M. Fröbe, A. Heini, K. Kredens, M. Mayerl, P. Pęzik, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, , E. Zangerle, Overview of PAN 2023: Authorship Verification, Multi-Author Writing Style Analysis, Profiling Cryptocurrency Influencers, and Trigger Detection, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), Lecture Notes in Computer Science, Springer, 2023.

[2] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Overview of the Multi-Author Writing Style Analysis Task at PAN 2023, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS, 2023.

[3] C. Toraman, O. Ozcelik, F. Sahinuç, U. Sahin, ARC-NLP at checkthat!-2022: Contradiction for harmful tweet detection, in: 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. 722–739. URL: https://ceur-ws.org/Vol-3180/paper-59.pdf.

[4] J. Devlin, M. Chang, K. Lee, K. Toutanova, 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, 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.

[5] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241.

[6] M. Kestemont, M. Tschuggnall, E. Stamatatos, W. Daelemans, G. Specht, B. Stein, M. Potthast, Overview of the Author Identification Task at PAN-2018: Cross-domain Authorship Attribution and Style Change Detection, 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_2.pdf.

[7] D. Zlatkova, D. Kopev, K. Mitov, A. Atanasov, M. Hardalov, I. Koychev, P. Nakov, An Ensemble-Rich Multi-Aspect Approach Towards Robust Style Change Detection: Notebook for PAN at CLEF 2018, in: 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/paper_142.pdf.

[8] M. Hosseinia, A. Mukherjee, A Parallel Hierarchical Attention Network for Style Change Detection: Notebook for PAN at CLEF 2018, in: 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/paper_91.pdf.

[9] E. Zangerle, M. Tschuggnall, G. Specht, B. Stein, M. Potthast, Overview of the Style Change Detection Task at PAN 2019, in: L. Cappellato, N. Ferro, D. E. Losada, H. Müller (Eds.), Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019, volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019. URL: https://ceur-ws.org/Vol-2380/paper_243.pdf.

[10] S. Nath, Style Change Detection by Threshold Based and Window Merge Clustering Methods, in: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019, volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019. URL: https://ceur-ws.org/Vol-2380/paper_163.pdf.

[11] C. Zuo, Y. Zhao, R. Banerjee, Style Change Detection with Feed-forward Neural Networks, in: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019, volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019. URL: https://ceur-ws.org/Vol-2380/paper_229.pdf.

[12] E. Zangerle, M. Mayerl, G. Specht, M. Potthast, B. Stein, Overview of the Style Change Detection Task at PAN 2020, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2696/paper_256.pdf.

[13] D. Castro-Castro, C. A. Rodríguez-Lozada, R. Muñoz, Mixed Style Feature Representation and B-maximal Clustering for Style Change Detection, in: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2696/paper_227.pdf.

[14] A. Iyer, S. Vosoughi, Style Change Detection Using BERT, in: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2696/paper_232.pdf.

[15] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Overview of the Style Change Detection Task at PAN 2021, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021, volume 2936 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 1760–1771. URL: https://ceur-ws.org/Vol-2936/paper-148.pdf.

[16] E. Strøm, Multi-label Style Change Detection by Solving a Binary Classification Problem, in: Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021, volume 2936 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 2146–2157. URL: https://ceur-ws.org/Vol-2936/paper-191.pdf.

[17] R. Deibel, D. Löfflad, Style Change Detection on Real-World Data using an LSTM-powered Attribution Algorithm, in: Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021, volume 2936 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 1899–1909. URL: https://ceur-ws.org/Vol-2936/paper-163.pdf.

[18] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Overview of the Style Change Detection Task at PAN 2022, 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. 2344–2356. URL: https://ceur-ws.org/Vol-3180/paper-186.pdf.

[19] T. Lin, C. Chen, Y. Tzeng, L. Lee, Ensemble Pre-trained Transformer Models for Writing Style Change Detection, in: 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. 2565–2573. URL: https://ceur-ws.org/Vol-3180/paper-210.pdf.

[20] X. Jiang, H. Qi, Z. Zhang, M. Huang, Style Change Detection: Method Based On Pre-trained Model And Similarity Recognition, in: 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. 2526–2531. URL: https://ceur-ws.org/Vol-3180/paper-205.pdf.

[21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[22] P. He, J. Gao, W. Chen, DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing, CoRR abs/2111.09543 (2021). URL: https://arxiv.org/abs/2111.09543. arXiv:2111.09543.

[23] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontañón, P. Pham, A. Ravula, Q. Wang, L. Yang, A. Ahmed, Big Bird: Transformers for Longer Sequences, in: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, Decem-

ber 6-12, 2020, virtual, 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/c8512d142a2d849725f31a9a7a361ab9-Abstract.html.

[24] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-Art Natural Language Processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp-demos.6.

[25] G. Salton, M. McGill, Introduction to Modern Information Retrieval, McGraw-Hill Book Company, 1984.

[26] S. Bird, E. Loper, NLTK: The Natural Language Toolkit, in: Proceedings of the ACL Interactive Poster and Demonstration Sessions, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 214–217. URL: https://aclanthology.org/P04-3031.

[27] C. Cortes, V. Vapnik, Support-Vector Networks, Mach. Learn. 20 (1995) 273–297. URL: https://doi.org/10.1007/BF00994018. doi:10.1007/BF00994018.