Stacked Model based Argument Extraction and Stance Detection using Embedded LSTM model

Notebook for the Touché Lab on Argument Retrieval at CLEF 2022

Pavani Rajula[1], Chia-Chien Hung[1] and Simone Paolo Ponzetto[1]

[1]*Data and Web Science Group, University of Mannheim, Germany*

Abstract

In this paper, we present our submission approach for the third Touché lab at CLEF 2022 [1], shared task 2 on Argument Retrieval for Comparative Questions, which tackles answering comparative questions based on argument retrieval of text passages to support answering comparative questions in the scenario of personal decision making. The previous two Touché editions [2] [3] mostly focused on retrieving complete arguments and documents, while this edition is about whether argument retrieval can support decision-making directly by extracting the argumentative gist from documents, by classifying their stance with respect to the objects compared. In our approach, we performed tokenization and named entity recognition using a RoBERTa classifier. Followed by generating Boolean queries by categorizing the words in must and should terms. The top 100 results are retrieved from an Elasticsearch index that contains the corpus provided for the shared task. Result documents are then stripped of code, advertisements, and other noise. A Stacked model with SVM model, DistilBERT, and a learning meta-model binary classification [4] is performed on a sentence level. Extracted arguments are then scored based on a mix of BM25, the ratio of argument sentences in a document, and the similarity of the query and the sentences in the document. Finally, stance detection is then performed on a per-document-base using a word-embedding LSTM model. Our system, achieved a retrieval performance quality score of 0.492 mean nDCG@5 and relevance score of 0.582 mean nDCG@5 which is a slight improvement over the baseline scores.

Keywords

Comparative Questions, Argument Identification, Natural Language Processing, Stance Detection

1. Introduction

Having alternatives and multiple options for everything these days, people tend to search for them and compare, to get the best among the choices available. The web contains a vast number of opinions and objective arguments that can facilitate the comparative decision-making process, but a faceted view of different aspects of the search topic makes it difficult. It creates the need of developing an open-domain general system that could process such information, and generate insights that support the user in informing well-justified opinions. Such a system has several challenges which include assessing an argument's relevance to a query, deciding what is an argument's main gist in terms of the take-away, and estimating how well an implied stance is

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

prajula@mail.uni-mannheim.de (P. Rajula); chia-chien.hung@uni-mannheim.de (C. Hung); ponzetto@uni-mannheim.de (S. P. Ponzetto)
https://github.com/rpavani1998 (P. Rajula)

justified. As this problem drives the attention of many researchers, such events as the Touché Lab on Argument Retrieval at CLEF, foster research on argument retrieval and establish more collaboration and exchange of ideas and data sets among researchers and collaborate to develop and share retrieval approaches that aim to support social and personal decisions.

We present in this paper our adopted approach for participation as Team Olivier Armstrong in the third Touché lab on argument retrieval at CLEF 2022, Shared Task 2. The main objective of this task is to use argument retrieval to support decision-making directly by extracting the argumentative gist from documents. Followed by classifying their stance with respect to the objects compared. The goal is to retrieve relevant documents from the given document collection, rank them using different approaches, and finally find the stance. Our approach to the submission made to this task is presented in detail in this paper.

2. Related Work

The approaches submitted for the previous two Touché editions [2] [3] are the most relevant work, which have about 130 submissions from 44 teams who have participated. Various approaches were proposed, which involved initial document retrieval using ChatNoir [5] search engine by using the original query, query pre-processing, and various query expansion techniques such as synonyms from WordNet [6], word embeddings using sense2vec [7] or word2vec [8], argument retrieval techniques, predicting document relevance labels by using a random forest classifier, XGBoost [9], LightGBM [10] and also implemented multiple (re-)ranking algorithms. The conclusive paper gives an overview of the implemented systems and ideas.

Argument mining problem, claim and premise detection drives the attention of various researchers [11], An open-source API for the argument retrieval from the text is TARGER [12], DistilBERT-based Argumentation Retrieval [13] proposed by one of the teams participated in Touché 2021 and there was also a Stacked Model proposed for Argument Identification [4], which tackles the argument identification task by following two approaches: a classical machine learning approach Support Vector Machine (SVM) model [14] and a DistilBert-based approach [15].

Stance classification is an active research area that has been studied in different domains. Determining the stand of the text toward a concrete entity or an abstract idea is quite challenging. Supervised learning is the basic and most common approach for most work on stance detection [16, 17]. Many studies have proposed different supervised ML algorithms such as classical algorithms, for instance, Naive Bayes (NB), SVM, decision trees, deep learning algorithms, RNNs, and LSTMs [18] [19] [17] , to detect the stance.

3. Datasets

Touchè organizers have provided 50 comparative questions (topics), for which documents are to be retrieved from the given corpus which is a collection document of about 0.9 million text passages. Stance data set [20] is also provided which has data dump from Stack Exchange[1]

[1] https://archive.org/details/stackexchange

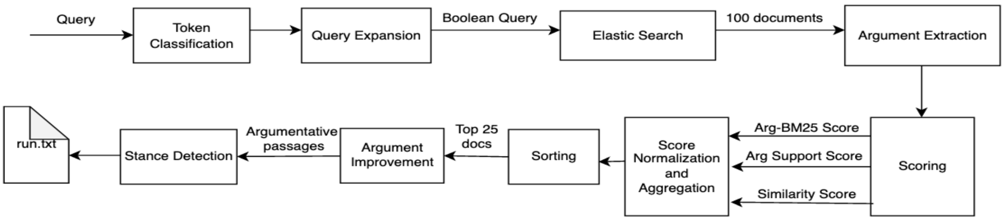


Figure 1: Architecture of submitted Approach

and L6 - Yahoo! Answers Comprehensive Questions and Answers version 1.0 (multi-part)[2], which contains questions from Yahoo and Stack Exchange, with the best answer and answer stance. Additional resources include a Subset of MS MARCO with comparative questions and a collection of text passages expanded with queries generated using DocT5Query. All these datasets are available on the task page.[3] For training, the stacked model datasets - Student Essays [21] and Web discourse [22] which are publicly available are used for this purpose.

4. Methodology

The approach is built with one of the previous year's team paper[23] as a baseline. The approach pipeline has multiple components which include query pre-processing, query expansion, document retrieval using elastic search, and further re-rank them through argument identification using a stacked model [4]. Followed by sorting the documents by multiple ranking scores and finally retrieving the top relevant arguments for stance detection by the embedded LSTM model. Our architecture of the proposed approach to building a search engine for answering comparative questions is presented in Figure 1. Each component in this architecture is explained in detail in the following sub-sections.

4.1. Query Parsing, Token Classification, and Query Expansion

Understanding the query and recognizing the query attributes would help in better query enrichment for document retrieval. Based on the approach suggested in the paper Towards Understanding and Answering Comparative Questions [20], we used the token-based Named Entity Recognition(NER) model using RoBERTa classifier [24] to tokenize the queries and label the objects, aspects, and predicates. And then query expansion is done by parsing the query, removing the stop words, and considering the labeled entities in the previous step, the words are classified into should and must categories and build a Boolean query. Entities - Objects and their synonyms fetched using WordNet are considered as must words and Aspect, Predict

[2] https://webscope.sandbox.yahoo.com/catalog.php?datatype=l
[3] https://webis.de/events/touche-22/shared-task-2.html

and its synonyms are considered as should words. By the end of this stage Boolean query is generated joined by 'AND' and 'OR' based on the category.

4.2. Document Retrieval

In this approach, we have used Elasticsearch client[4], an open-source search engine, which allows us to store, and search from the huge volume of data, the same is used to improve ChatNoir [25] All the documents from the given corpus with the collection of documents are indexed with the contents field provided in the data set along with document_id as the id. The body of the index consists of chatNoirUrl and processed content - removal of special characters, stop words, and lemmatized verbs and nouns with nltk [26]. For each query topic, the boolean query built in the previous step is used to search from the Elasticsearch index dump and retrieve the top 100 documents.

4.3. Argument Extraction

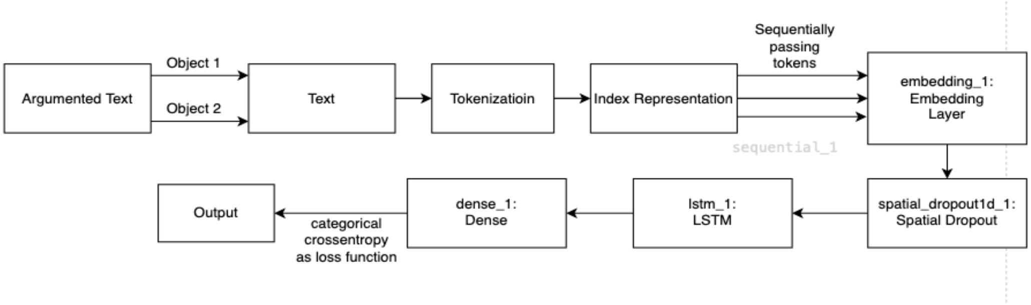
Argument identification is used to detect the comparative sentences in the document. To get full content, using Boilerpy3[5] and Trafilatura[6] and the URL from the Elasticsearch results is used to retrieve the HTML source page, removing tags, advertisements, and other cleaning processes. For sentences from the documents, binary classification is applied using the stacked model, which is modeled using Student Essays and Web discourse data sets, and then every sentence is classified as an argument or non-argument. The stacked model architecture consists of two main components: the base models, which include the trained SVM model and the trained transformer-based model (DistilBERT) in parallel, and the meta-model, which will learn from the outputs of the two models to produce the final prediction of a sentence. The proposed Stacked model compared to DistilBERT achieved better performance when trained with the Student Essays and Web Discourse datasets. [4]

4.4. Scoring and Sorting

Now in this step, the score is estimated using the best matching between the query and the arguments extracted from the documents to sort and rank them accordingly. Unlike the baseline approach, only three different scores are considered to evaluate the argument quality which are

- BM25 measure: Calculating on argumentative sentences of each document with respect to the original query, through re-indexing the retrieved documents by creating new ones that contain only argumentative sentences. Then the arg-BM25 score of each the document is calculated by querying the new argumentative documents with the original topic.
- Argument support score: Representing the ratio of argument sentences among all existent sentences in the document.
- Similarity score: Evaluating the similarity of two sentences based on the context and English language understanding using the SentenceTransformer library[27], which is

[4] https://www.elastic.co/
[5] https://github.com/jmriebold/BoilerPy3
[6] https://github.com/adbar/trafilatura



Argumented Text
Object 1
Object 2
Text
Tokenizatioin
Index Representation
Sequentially passing tokens
embedding_1: Embedding Layer
sequential_1
spatial_dropout1d_1: Spatial Dropout
lstm_1: LSTM
dense_1: Dense
categorical crossentropy as loss function
Output

Figure 2: Stance Detection Model Architecture

calculated using the similarity between the original query and every argumentative sentence in the document, and considers the average as the score.

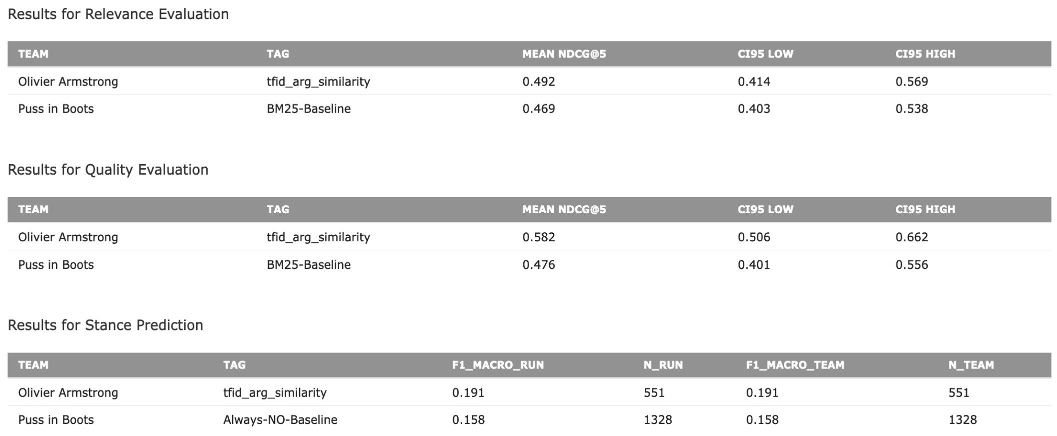
All these scores are normalized and all the scores are summed up using the respective weights then the documents are sorted and the top 25 highly relevant documents are fetched and ranked based on the descending scores.

4.5. Stance detection

Now that the top relevant documents are retrieved, the next step is to retrieve the relevant argument passages to detect their stance with respect to the query objects. In this approach, assuming that all the arguments with a good argument quality score for the query, is the most relevant passage, will be used for the stance detection. Argument quality [28] here is measured by the averaging list of arguments and query pairs, using the pre-trained neural network from the IBM Debater project [29] API service. Given a list of sentences and the query, The API returns a score ranging between 0, indicating that sentence is of the lowest quality to 1, indicating that sentence is of very high quality, for the query.

Number of studies [18] has shown that for conversation-based tasks, the LSTM approach outperforms other sequential classifiers and feature-based models. Long Short Term Memory networks are a special kind of RNN, capable of learning long-term dependencies. An LSTM recurrent unit tries to remember all the past knowledge that the network is seen so far and to forget irrelevant data. This is done by introducing different activation function layers called gates for different purposes. Each recurrent unit also maintains a vector called the Internal Cell State which conceptually describes the information that was chosen to be retained by the previous LSTM recurrent unit.

We experimented with the stance dataset provided by Touchè, which is annotated using predefined labels which are No stance, Neutral, Pro first object, and Pro second object. Figure 2 shows the model architecture used in this approach. The dataset is split into a training set and a test set using an 80:20 split. First, the argument passage is concatenated with both the objects separately with the same stance label for both. Pre-processing is performed on both the data split, based on the tokenization technique, where the text is tokenized and each token is transformed into an index-based representation. Then, each token sentence-based indexes will be passed sequentially through an embedding layer, this embedding layer will



Results for Relevance Evaluation

TEAM	TAG	MEAN NDCG@5	CI95 LOW	CI95 HIGH
Olivier Armstrong	tfid_arg_similarity	0.492	0.414	0.569
Puss in Boots	BM25-Baseline	0.469	0.403	0.538

Results for Quality Evaluation

TEAM	TAG	MEAN NDCG@5	CI95 LOW	CI95 HIGH
Olivier Armstrong	tfid_arg_similarity	0.582	0.506	0.662
Puss in Boots	BM25-Baseline	0.476	0.401	0.556

Results for Stance Prediction

TEAM	TAG	F1_MACRO_RUN	N_RUN	F1_MACRO_TEAM	N_TEAM
Olivier Armstrong	tfid_arg_similarity	0.191	551	0.191	551
Puss in Boots	Always-NO-Baseline	0.158	1328	0.158	1328

Figure 3: Results

output an embedded representation of each token which is passed through a spatial dropout layer [30], are passed through an LSTM neural net, stacked by a dense layer. The following settings for LSTM-based models were chosen: input layer size 500 (equal to the word embedding dimension), hidden layer size of 60, training for max 10 epochs with initial learning rate 1e-3 using ADAM [31] for optimization, dropout 0.2. Models were trained using categorical cross-entropy loss. The use of one, relatively small hidden layer and dropout help to avoid over-fitting. This trained model gave an accuracy of 0.86 and a loss of 0.34 on the test data set. For the prediction of the stance, the gist of the arguments retrieved is concatenated with the objects separately, given as input, and then predict using the trained model, which results in the same stance label for both the records. Then the labels are converted to the labels as per the labels specified in the Task and The final output is inserted into a text file in the format proposed by the Touché organization.

5. Evaluation

The shown architecture in Figure 1 presents our base approach, from which submission is made for task-2 Touchè 2022. In Table 1, we present the Precision, Recall, and F1 scores for each label obtained for the test dataset. A single-run submission has been made to the Touchè committee through the manual labeling of the documents with the help of a human assessor. Our results seem to point in the right direction, achieving a slight improvement over the baseline with a quality score of 0.492 mean nDCG@5 and relevance score of 0.582 mean nDCG@5. Our stance detection model performed better than the baseline model with F1_Macro score of 0.191. Figure 3 shows our system scores compared with baseline. Our system would have performed much better if more documents are initially retrieved and using multiple other methods for argument extraction or using more data for training the model. For stance detection, instead of considering top argument sentences as the gist of the document, if other techniques are used, the system would have given better results.

Table 1
Stance Detection Model Performance Results

Label	Precision	Recall	F1-score
0 - No stance	0.846	0.917	0.880
1 - Neutral	0.967	0.894	0.929
2 - Pro first object	0.746	0.909	0.820
3 - Pro second object	0.902	0.780	0.836

6. Conclusion

In this paper, we as Team Olivier Armstrong presented our solution to the shared task 2 of argument retrieval for answering comparative questions at Touchè 2022. We proposed an approach for document and argument retrieval, based on several parts of existing systems that have shown acceptable performance previously. Our major contributions supervised learning embedded-LSTM model for stance detection and proposed stacked model for argument extraction is performing better compared to the previously proposed DistilBERT. Both the trained models, have performed well and have acceptable results. Our proposed approach outperforms the baseline. In future work, more enhanced and advanced techniques can be used for better query enrichment, argument extraction, argument quality, and stance detection.

References

[1] A. Bondarenko, M. Fröbe, J. Kiesel, S. Syed, T. Gurcke, M. Beloucif, A. Panchenko, C. Biemann, B. Stein, H. Wachsmuth, M. Potthast, M. Hagen, Overview of Touché 2022: Argument Retrieval, in: M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, V. Setty (Eds.), Advances in Information Retrieval. 44th European Conference on IR Research (ECIR 2022), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2022. URL: https://webis.de/publications.html#bondarenko_2022c.

[2] A. Bondarenko, M. Fröbe, M. Beloucif, L. Gienapp, Y. Ajjour, A. Panchenko, C. Biemann, B. Stein, H. Wachsmuth, M. Potthast, M. Hagen, Overview of Touché 2020: Argument Retrieval, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névéol, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 11th International Conference of the CLEF Association (CLEF 2020), volume 12260 of *Lecture Notes in Computer Science*, Springer, Berlin Heidelberg New York, 2020, pp. 384–395. URL: https://link.springer.com/chapter/10.1007/978-3-030-58219-7_26. doi:10.1007/978-3-030-58219-7_26.

[3] A. Bondarenko, L. Gienapp, M. Fröbe, M. Beloucif, Y. Ajjour, A. Panchenko, C. Biemann, B. Stein, H. Wachsmuth, M. Potthast, M. Hagen, Overview of Touché 2021: Argument Retrieval, in: K. Candan, B. Ionescu, L. Goeuriot, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 12th International Conference of the CLEF Association (CLEF 2021), volume 12880 of *Lecture Notes in Computer Science*, Springer, Berlin Heidelberg New York, 2021, pp.

450–467. URL: https://link.springer.com/chapter/10.1007/978-3-030-85251-1_28. doi:10.1007/978-3-030-85251-1_28.

[4] A. Alhamzeh, M. Bouhaouel, E. Egyed-Zsigmond, J. Mitrović, L. Brunie, H. Kosch, A stacking approach for cross-domain argument identification, in: C. Strauss, G. Kotsis, A. M. Tjoa, I. Khalil (Eds.), Database and Expert Systems Applications, Springer International Publishing, Cham, 2021, pp. 361–373.

[5] M. Potthast, M. Hagen, B. Stein, J. Graßegger, M. Michel, M. Tippmann, C. Welsch, Chatnoir: A search engine for the clueweb09 corpus, in: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '12, Association for Computing Machinery, New York, NY, USA, 2012, p. 1004. URL: https://doi.org/10.1145/2348283.2348429. doi:10.1145/2348283.2348429.

[6] G. A. Miller, Wordnet: A lexical database for english, Commun. ACM 38 (1995) 39–41. URL: https://doi.org/10.1145/219717.219748. doi:10.1145/219717.219748.

[7] A. Trask, P. Michalak, J. Liu, sense2vec - a fast and accurate method for word sense disambiguation in neural word embeddings (2015).

[8] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, 2013. arXiv:1301.3781.

[9] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 785–794. URL: https://doi.org/10.1145/2939672.2939785. doi:10.1145/2939672.2939785.

[10] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 3149–3157.

[11] H. Wachsmuth, M. Potthast, K. Al-Khatib, Y. Ajjour, J. Puschmann, J. Qu, J. Dorsch, V. Morari, J. Bevendorff, B. Stein, Building an argument search engine for the web, in: Proceedings of the 4th Workshop on Argument Mining, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 49–59. URL: https://aclanthology.org/W17-5106. doi:10.18653/v1/W17-5106.

[12] A. Chernodub, O. Oliynyk, P. Heidenreich, A. Bondarenko, M. Hagen, C. Biemann, A. Panchenko, TARGER: Neural argument mining at your fingertips, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Florence, Italy, 2019, pp. 195–200. URL: https://aclanthology.org/P19-3031. doi:10.18653/v1/P19-3031.

[13] A. Alhamzeh, M. Bouhaouel, E. Egyed-Zsigmond, J. Mitrović, Distilbert-based argumentation retrieval for answering comparative questions, in: CLEF, 2021.

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

[15] 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.

[16] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, C. Cherry, SemEval-2016 task 6: Detecting stance in tweets, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), Association for Computational Linguistics, San

Diego, California, 2016, pp. 31–41. URL: https://aclanthology.org/S16-1003. doi:10.18653/v1/S16-1003.

[17] S. Gottipati, M. Qiu, L. Yang, F. Zhu, J. Jiang, Predicting user's political party using ideological stances, in: A. Jatowt, E.-P. Lim, Y. Ding, A. Miura, T. Tezuka, G. Dias, K. Tanaka, A. Flanagin, B. T. Dai (Eds.), Social Informatics, Springer International Publishing, Cham, 2013, pp. 177–191.

[18] E. Kochkina, M. Liakata, I. Augenstein, Turing at semeval-2017 task 8: Sequential approach to rumour stance classification with branch-lstm, CoRR abs/1704.07221 (2017). URL: http://arxiv.org/abs/1704.07221. arXiv:1704.07221.

[19] K. Dey, R. Shrivastava, S. Kaushik, Topical stance detection for twitter: A two-phase LSTM model using attention, CoRR abs/1801.03032 (2018). URL: http://arxiv.org/abs/1801.03032. arXiv:1801.03032.

[20] A. Bondarenko, Y. Ajjour, V. Dittmar, N. Homann, P. Braslavski, M. Hagen, Towards Understanding and Answering Comparative Questions, in: K. S. Candan, H. Liu, L. Akoglu, X. L. Dong, J. Tang (Eds.), 15th ACM International Conference on Web Search and Data Mining (WSDM 2022), ACM, 2022, pp. 66–74. URL: https://dl.acm.org/doi/10.1145/3488560.3498534. doi:10.1145/3488560.3498534.

[21] C. Stab, I. Gurevych, Annotating argument components and relations in persuasive essays, in: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin City University and Association for Computational Linguistics, Dublin, Ireland, 2014, pp. 1501–1510. URL: https://aclanthology.org/C14-1142.

[22] I. Habernal, I. Gurevych, Argumentation mining in user-generated web discourse, Computational Linguistics 43 (2017) 125–179. URL: https://aclanthology.org/J17-1004. doi:10.1162/COLI_a_00276.

[23] A. Alhamzeh, M. Bouhaouel, E. Egyed-Zsigmond, J. Mitrović, Distilbert-based argumentation retrieval for answering comparative questions, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum (CLEF 2021), number 2936 in CEUR Workshop Proceedings, Aachen, 2021, pp. 2319–2330. URL: http://ceur-ws.org/Vol-2936/#paper-209.

[24] 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.

[25] J. Bevendorff, B. Stein, M. Hagen, M. Potthast, Elastic chatnoir: Search engine for the clueweb and the common crawl, in: ECIR, 2018.

[26] E. Loper, S. Bird, Nltk: The natural language toolkit, CoRR cs.CL/0205028 (2002). URL: http://dblp.uni-trier.de/db/journals/corr/corr0205.html#cs-CL-0205028.

[27] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, CoRR abs/1908.10084 (2019). URL: http://arxiv.org/abs/1908.10084. arXiv:1908.10084.

[28] S. Gretz, R. Friedman, E. Cohen-Karlik, A. Toledo, D. Lahav, R. Aharonov, N. Slonim, A large-scale dataset for argument quality ranking: Construction and analysis, CoRR abs/1911.11408 (2019). URL: http://arxiv.org/abs/1911.11408. arXiv:1911.11408.

[29] R. Bar-Haim, Y. Kantor, E. Venezian, Y. Katz, N. Slonim, Project Debater APIs: Decomposing the AI grand challenge, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational

Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 267–274. URL: https://aclanthology.org/2021.emnlp-demo.31. doi:10.18653/v1/2021.emnlp-demo.31.

[30] Y. Gal, Z. Ghahramani, A theoretically grounded application of dropout in recurrent neural networks, 2015. URL: https://arxiv.org/abs/1512.05287. doi:10.48550/ARXIV.1512.05287.

[31] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014. URL: https://arxiv.org/abs/1412.6980. doi:10.48550/ARXIV.1412.6980.