Evaluating dataset creation heuristics for concept detection in web pages using BERT

Michael Paris[1][0000−0003−2077−6984] and Robert Jäschke[1,2][0000−0003−3271−9653]

[1] Berlin School for Library and Information Science
Humboldt-Universität zu Berlin
{michael.paris,robert.jaeschke}@hu-berlin.de
[2] L3S Research Center Hannover

Abstract. Dataset creation for the purpose of training natural language processing (NLP) algorithms is often accompanied by an uncertainty about how the target concept is represented in the data. Extracting such data from web pages and verifying its quality is a non-trivial task, due to the Web's unstructured and heterogeneous nature and the cost of annotation. In that situation, annotation heuristics can be employed to create a dataset that captures the target concept, but in turn may lead to an unstable downstream performance. On the one hand, a trade-off exists between cost, quality, and magnitude for annotation heuristics in tasks such as classification, leading to fluctuations in trained models' performance. On the other hand, general-purpose NLP tools like BERT are now commonly used to benchmark new models on a range of tasks on static datasets. We utilize this standardization as a means to assess dataset quality, as most applications are dataset specific. In this study, we investigate and evaluate the performance of three annotation heuristics for a classification task on extracted web data using BERT. We present multiple datasets, from which the classifier shall learn to identify web pages that are centered around an individual in the academic domain. In addition, we assess the relationship between the performance of the trained classifier and the training data size. The models are further tested on out-of-domain web pages, to asses the influence of the individuals' occupation and web page domain.

Keywords: Dataset · generation · heuristic · bias · quality · web archive · classification.

1 Introduction

Dataset generation for a specific machine learning task is time-consuming when done by humans and takes even more time in the case of manual sample creation. Since the release of BERT [9] many variations of pre-trained NLP language models have been published, which alleviate the stress on optimizing the architecture for many NLP tasks and increased the focus on dataset generation. Specifically, aspects of dataset curation [23] gain growing attention as a consequence of this development. In this context, datasets used to train a concept classifier allow the model to learn a representation through annotated examples. Furthermore, tasks comprising the classification of an abstract concept rely on heuristics for their a priori definition to create such annotated examples. This increases the likelihood of unintended bias, depending on how the annotators or curators interpret these heuristics. The accompanying indefiniteness can lead to different training datasets, which in turn would lead to diverging classifications of the model on new data. Uncertainties have generally been analyzed as a model-dependent phenomenon, such that specific datasets were created to probe the behavior of BERT [7] or improve on the learned decision boundary [13]. On the one hand, improving the model through diversification of the data samples improves the quality of the decision boundary by means of a more detailed representation of the concept. On the other hand, this approach changes nothing about the architecture of the model, implying that heuristics determine the representation of the concept within the dataset learned by the model. In addition to the challenge of creating refined datasets through heuristics, a semantic concept drift may occur over time, thereby altering the initially captured concept. Especially for rapidly changing web concepts, adjusting individual data samples quickly becomes unfeasible.

In contrast to the current approach, in which datasets are used to analyze or improve a language model, we use BERT to measure how well different heuristics for annotation reflect a particular concept. We use raw web archives as data source, present a pipeline that simplifies the creation of datasets for NLP language models, and compare the performance and limitations of the heuristics.

Our approach allows researchers to leverage their domain knowledge of an existing dataset to train an NLP classifier and extract subsets relevant for their research objective. Specifically, our contribution comprises

1. *the creation of datasets using three different heuristics,*
2. *a pipeline for the application and comparison of these heuristics, and*
3. *insights into the dataset creation quality, measured using BERT.*[3]

This paper is organised as follows: Section 2 discusses related work, Section 3 presents the data creation process, training of the models, and approaches to investigate their performance, Section 4 presents the results, and Section 5 discusses possible explanations and pitfalls of these.

2 Related Work

The Web as a dataset resource has been frequently discussed and used by the community [21,11]. To a large degree the datasets created for the use in linguistics and natural language processing (NLP) rely on (semi-)structured and homogeneous data (e.g., Wikipedia) [25,30,16]. But most of the Web is only available in an unstructured form and is thus less accessible to non-technical research fields.

[3] Code and data are available at https://github.com/parismic/EvaluateHeuristics/.

Non-expert researchers working with unstructured data are therefore limited to data resources such as Twitter [4] or news articles [26]. These resources do not require a large overhead of frameworks or custom designed tools to yield datasets appropriate for specific research questions [18].

Web archives, for which various tools already exist, could be a rich resource for the curation of derived datasets, if the effort to accumulate research-related content from millions of web pages could be reduced [26]. Some tools aim to lower the threshold for such instances with a focus on reproducibility and best practices [15,3]. In many of these cases the desired collection of web pages depends on the textual information presented. This requires boiler plate removal approaches to extract the information from the heterogeneously structured pages on the Web. Advances in boiler plate removal allow high accuracy in the automatic extraction of main content sections [27].

The processing of common NLP tasks has fundamentally changed after the release of BERT [9], after which many other general NLP tools have followed. Without the requirement to deal with the architecture of the system, the importance of dataset quality becomes more pronounced [5]. For common NLP tasks it is desired to gauge the quality of the trained model on samples outside the distribution through contrast sets [12]. In a sense, the incomplete information about the decision boundary in [12] stems from the heuristics used for the creation of the dataset.

These heuristics are often manifested with methods for assembling datasets, from **1**) **existing resources** (e.g., author names from a digital library to label persons) [24], **2**) **manual annotation** [1], and **3**) **weak supervision** (e.g., structured information on Wikipedia for labeling financial events [10] or regular expressions to identify or extract samples). In those methods, a bias [20] of the dataset creator is carried into the heuristics, which should at least be coherent across the different dataset creation approaches. Further, bias enters the dataset on the level of the annotator or sample creator to an extent that the annotator can be identified on the basis of the sample itself [13].

This study aims to highlight another aspect of bias introduced to the dataset depending on the creation approach and definiteness of the concept itself. Specifically, we provide an exemplary case study for the concept of a *person-centric web page*.

3 Datasets & Experiments

This section details the dataset creation process for the concept of *person-centric* (*PC*) *web pages* in the academic context for the training, evaluation, testing, and validation of the classifiers. In general, we assume that a human observer can recognize whether a web page is *PC* and contains information about a person on the basis of the main content and not the style or navigation. We limit our investigation to content presented in natural language texts on web pages. The overall procedure will extract and boiler-plate web pages from a large web crawl and associate a class label with the boiler-plated content. The content together with the annotations will be used to fine-tune a pre-trained BERT model. This is followed by an evaluation of the model on test sets, which are subsets of all fine-tuning datasets to express the coherence between heuristics.

3.1 Datasets

We utilize several datasets to define and investigate what constitutes a *PC* web page in the academic context and to investigate the divergence between commonly used heuristics (i.e., *human annotation*, *weak annotation from existing resources*, and *weak annotation with regular expressions*) for creating training data for a binary classifier.

The datasets named in the remainder of this work as *DBLP*, *Manual*, *RegEx*, and *Wikidata* were created on the basis of the 2019-06 snapshot of the “German Academic Web” (GAW) [22] which will be referred to as *crawl*. The snapshot was created on the basis of a URL seed list containing all home pages of German academic institutions with the right to award a doctorate degree at that time. This *crawl* contains WARC records [14] of web pages reachable within 20 link hops from the domains of the seeds.

An additional dataset $\mathit{Wikidata}_{Q5}$ for the validation of the classifier performance was created on the basis of URLs of the `official websites` (P856) associated with all entities of `instance` (P31) `human` (Q5) on Wikidata [28]. For all these entities the corresponding `occupation` (P106) was extracted and WARC records were created for the listed official websites on 2021-01-28 using the library scrapy-warcio.[4]

Pre-Processing We restricted the *crawl* to records of MIME type[5] *text/html* and HTTP response code *200*. After that, the HTML for all WARC records was extracted and processed using the boiler plate removal tool Web2Text [27] trained on the CleanEval [2] dataset. This process enables a robust extraction of the main textual content of a web page without the noise introduced by headers, side panels, navigation, etc. or any knowledge of the structure of the web page. This step is applied to all extracted WARC records and is followed by the removal of duplicates and the removal of identical text samples from the pairwise larger dataset, yielding non-overlapping datasets (i.e., for datasets $\tilde{D}_1$ and D_2, where $|\tilde{D}_1| > |D_2|$, $\tilde{D}_1$ is transformed to $D_1 = \tilde{D}_1 \setminus D_2$).

Dataset Enrichment and Annotation Due to an expected low frequency of the *PC* concept in a random subset of the *crawl*, an enrichment process was applied to increase its frequency. For that, a dataset was created such that from each seed institution in the *crawl* three annotators independently navigated to web presences of research groups of that institution and extracted common base URL paths of the staff. Specifically, this was done by collecting URLs of staff

[4] https://github.com/internetarchive/scrapy-warcio

[5] Multipurpose Internet Mail Extensions – a standard for naming different types of content.

members and extracting the longest common URL prefixes. The annotators were instructed to cover different fields of research to ensure a contextual diversity, such that the classifier would not instead focus on a particular research domain. However, diversification was not specified in any narrow terms to mimic an application scenario in which an unknown bias may affects the dataset. All URLs starting with an element from the list of common URL prefixes of staff members were extracted from the *crawl* and comprise the *enrichment* dataset. This process aims to increase the frequency of *PC* web pages without specifically excluding non-*PC* web pages.

In the four months following the completion of the crawl, the same annotators had to decide whether a page is *PC* on the basis of the displayed content in a web browser and label it as such. If all three annotators agreed, then the respective label was used as annotation, URLs not present at annotation time were removed from further processing and all others were considered to be *not PC*.

***Manual* dataset** The *Manual* dataset was created in a two-step process. First, 2,000 random records from the *crawl* and 1,274 records from the *enrichment* dataset were selected. Next, these 3,274 records were annotated, whereby the annotator agreement yielded a $\kappa_{\text{FLEISS}} = 0.844$. The resulting *Manual* dataset contains 1,407 non-*PC* and 606 *PC* samples, of which 68 originated from the 2,000 randomly selected records.

***RegEx* dataset** To construct a dataset which reduces annotation cost and leverages the structure of the *crawl* we constructed a dataset by utilizing common patterns in URLs of *PC* pages using regular expressions. The regular expressions act as a weak annotation mechanism to classify records of the *crawl* based on their URL. If any of the following regular expressions matched against substrings of the last path element[6] in the URL path, the record was annotated as *PC*: `mitarbeite`, `angestellte`, `group`, `gruppe`, `staff`, `~[a-z]`, `people`, `team`, `kolleg`, `lehrend`, `beschaeftigte`.

The non-*PC* records were extracted by only considering URLs which did not match any of the mentioned regular expressions *anywhere* in the URL. We select a subset such that the ratio of *PC* and non-*PC* records is equal to that ratio in the *Manual* dataset and refer to the resulting dataset as *RegEx*.

***DBLP* dataset** For many concepts there are datasets available which already provide a level of proximity to the desired concept for classification. In the case of *PC* web pages, DBLP [17] provides a frequently used dataset which extensively covers researchers from computer science and their associated web pages. We used these URLs in the DBLP dump of 2020-10-01 to identify and annotate records in the *crawl* as *PC*. All URLs associated with a person and contained within the *crawl* were selected and amounted to 1,859 weakly-annotated *PC*

[6] That is, the string confined by the last and second to last '/' in the SURT format of the URL.

Table 1: Total counts of the records in the datasets at different phases of the cleaning process. The difference between the sum of records in the train and test set in relation to the pre-processed dataset stems from the removal of identical text-label pairs within and between the datasets.

	URLs	URLs$_{\text{on seeds}}$	WARCs	Pre-proc.	Train		Test	
			PC		*PC*	non-*PC*	*PC*	non-*PC*
DBLP	234,291	4,439	2,759	1,859	1,669	3,881	121	281
Manual	-		800	2,013	484	1,126	121	281
RegEx	-		1,135,899	-	4,840	11,260	-	-
Wikidata	159,431	531	388	293	-	-	254	589

samples. Non-*PC* samples were constructed by employing the same method as for the *RegEx* dataset, whereby the ratio of *PC* and non-*PC* samples equal that in the *Manual* dataset.

Wikidata-derived datasets Wikidata provides an additional resource for the selection of *PC* web pages, as well as rich ontological structure to validate the scope of the trained classifiers. As such, it allows us to investigate the heuristics in terms of the entities' occupation and region. Analogous to the *DBLP* dataset, the existing WARC records in the *crawl* associated with the aforementioned URLs were extracted, comprising the weakly-annotated *Wikidata* dataset of 293 *PC* samples. To validate results and determine limitations of the classifiers, we categorized all occupations in $\mathit{Wikidata}_{Q5}$ as *academic*, if they or an immediate sub-class have an occupation of 1) researcher, 2) knowledge worker, 3) scientist, 4) scholar, or 5) university-teacher, thus allowing us to observe the preference of the classifier given that the *crawl* focuses on the academic web.

3.2 BERT as a measurement tool for heuristics

With the rise of powerful general-purpose language models following BERT [9] a shift occurred from the traditional NLP pipeline towards a dataset-focused approach. Allowing users to quickly fine-tune a pre-trained model for a given task alleviates the previously required considerations about the model architecture and allows users to focus on issues like dataset curation. In general, these models are evaluated on established datasets, thereby testing their performance on tasks defined through the dataset, while maintaining stable performance across different dataset for the same task. Due to this stability these models can be used to evaluate the coherence and performance of a dataset in capturing a concept, and in turn the heuristics underlying the annotation process.

Fine-tuning BERT We used hugginface's pretrained multilingual-cased BERT implementation [29] trained on cased text in the top 104 languages. We fine-tuned for 4 epochs, using batch size 32, with learning rate 2×10^5 on the Adam optimizer with weight decay [19] and maximum sequence length of 128 word pieces. The optimal number of epochs was determined by finding a model with minimal validation loss using the standard fine-tuning approach. From the four datasets (*Manual*, *DBLP*, *RegEx*, *Wikidata*) we constructed three training sets and three test sets as presented in Table 1. For each sample size the pre-trained model was fine-tuned for 10 different random seeds, used in the selection of the sub-samples and initialization of the system. This was done to determine the robustness of the model inherited from the training dataset by determining the fluctuation with varying training sample sizes.

3.3 Complementarity

We also investigate the improvement achievable by one classifier to another [6]. This provides another perspective on how the heuristics for constructing the training datasets complements another. Given the prediction results of two binary classifiers A and B, we can determine the complementary recall $R_{\text{comp}} = 1 - \frac{|B_{\text{wrong}} \cap A_{\text{wrong}}|}{|A_{\text{wrong}}|}$ and precision $P_{\text{comp}}(A, B) = 1 - \frac{|B_{\text{wrong}} \cap A_{\text{wrong}}|}{|A_{\text{wrong}}|}$ [8].[7]

4 Results

The following results present the performance of the fine-tuned classifiers based on the datasets described in Section 3.1. As a starting point, we present the F1 score as a function of the fine-tuning sample size as described in Section 3.2.

4.1 Robustness and Sample Size

We observe in Figure 1 that the F1 scores diverge with increasing sample size for the classifiers trained on the *DBLP* and *Manual* datasets when tested on the respective other. The classifier trained on the *RegEx* dataset displays a comparable performance for a sample size of 1,600 to the *DBLP* and *Manual* classifiers tested on the *Manual* and *DBLP* datasets, respectively. With an increasing sample size the *RegEx*-trained classifier outperforms the other classifiers when tested on *non-native*[8] test data. The general performance of the regular expression approach shows a score between 0.83 and 0.86 on *Wikidata*. A significant drop in the variance of the models' F1 scores occurs between sample sizes of 200 to 400 and 400 to 600 for the datasets *DBLP*, *Manual* and *RegEx*, respectively. The *Manual*-trained model generally performs better than the *DBLP*-trained model when tested on the respective other test set.

4.2 Context Dependence

To determine the dependence of the trained model on the context, we investigate the occupational dependence as a proxy for the context of the concept. Since

[7] Where 'wrong' refers to the falsely classified *PC* items for recall, and the falsely classified non-*PC* items for precision, respectively.

[8] Test data which does not originate from the same distribution as the training dataset.

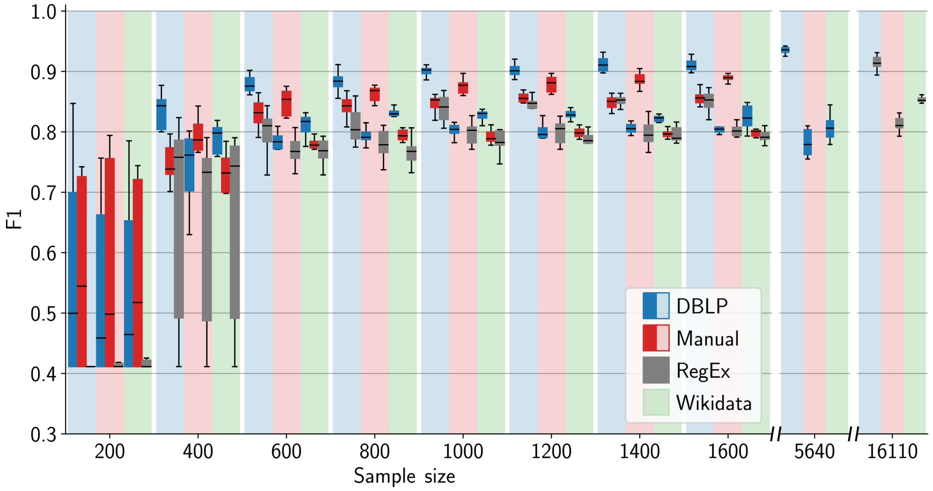

Fig. 1: Performance of the trained classifiers in terms of the F1 score as a function of the fine-tuning sample size and the source of the test data. For each training sample size a classifier was trained and evaluated for 4 epochs on 10 different seeds (solid colors, boxplot) and tested on 3 datasets (translucent colors, background strips).

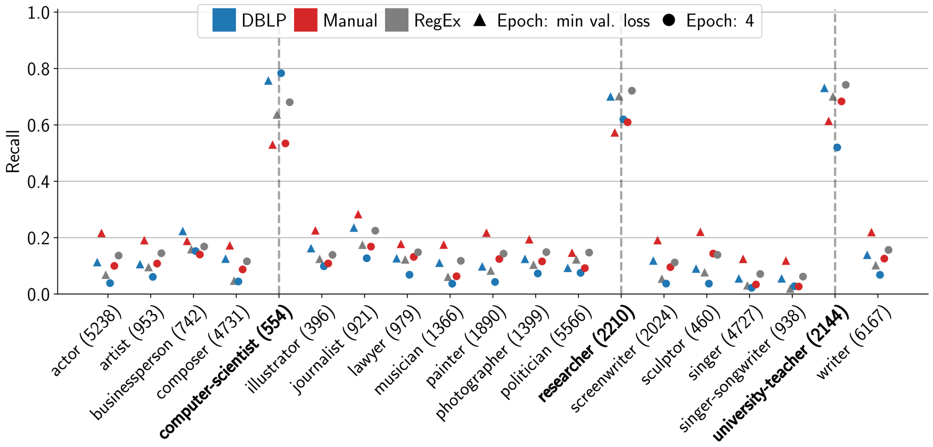

Fig. 2: Recall by occupation for the classifier after epochs: 4 (●) and epoch in which the validation loss reached the minimum (▲). The number following the occupation expresses the samples size for that occupation.

negative samples are unavailable, the recall is presented in Figure 2 instead of the F1 score. It illustrates the influence of the epoch given the test data of $\mathit{Wikidata}_{Q5}$. A strong discrepancy is displayed between the recall of computer-scientist, researcher, and university-teacher and all other occupations. Out of the distinct occupations, the *DBLP*-trained classifier performs best on the computer-scientist occupation. This occupation also presents the largest spread between the

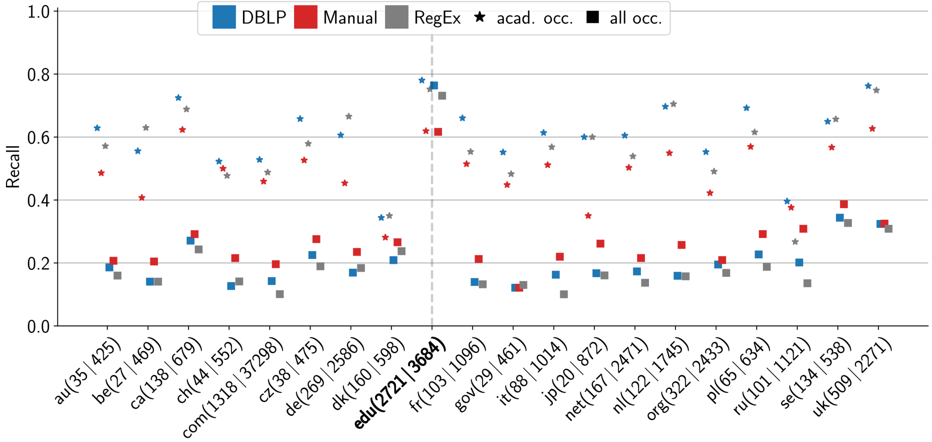

Fig. 3: Recall for the 20 most frequently occurring top-level domains depending on the underlying association with an academic occupation (see Section 3.1) in contrast to all occupations. The numbers after the domain express the sample size for the *academic* and *all* occupations, respectively.

examined classifiers. On a large scale, only a minor dependence is observed with regard to the training epochs for the same training dataset. An improvement can be observed at epoch = 4 (●) for the *RegEx*-trained classifier and a general deterioration for the *DBLP* and *Manual*-trained classifier relative to the epoch with minimal validation loss (▲).

4.3 Domain Dependence

Following the results in Figure 2, a limitation of the trained classifiers could also arise from regional differences associated with the top-level domain (TLD). Therefore, Figure 3 presents the TLD dependence of the recall with respect to the fine-tuning dataset. This is presented for the academic occupations (★) in contrast to all occupations (■). A clear shift in recall can be observed when *academic* occupations are classified independently of the TLD. This shift averages for all classifiers in *all* occupations at $R = 0.23$ and in the *academic* occupations at $R = 0.56$. Within the two different categories we can observe that in most TLDs the *Manual* and *DBLP* classifier perform best in *all* (■) and *academic* (★) occupations, respectively. Since the *GAW* is focused on Germany it contains a language bias. This could cause the performance of the trained classifier to vary in its ability to determine the correct results in another language setting. Another issue could be variations specific to each TLD, like the page structure or language, which could limit the applicability of this approach.

4.4 Complementarity

Since we would like to not only investigate the quality of the initial heuristics on the classifier's performance but also how much these diverge, it is necessary

Table 2: Complementarity measures on *Wikidata*. Classifiers were trained on 1,600 training samples.

(a) Recall R_{comp}

B *A*	*DBLP*	*Manual*	*RegEx*
DBLP	0.0	23.8	23.8
Manual	52.9	0.0	36.8
RegEx	**56.2**	41.1	0.0

(b) Precision P_{comp}

B *A*	*DBLP*	*Manual*	*RegEx*
DBLP	0.0	52.8	52.8
Manual	39.3	0.0	53.6
RegEx	54.1	**64.9**	0.0

to determine the recall complementarity (cf. Section 3.3) of the predictions. We perform this on the *Wikidata* dataset, as this is a *non-native* dataset for either of the classifiers. Table 2 can be understood as the relative improvement of classifier A by classifier B. Under the same training conditions, the classifiers retain a significant discrepancy in between the potential improvement in recall (Table 2a) provided by *DBLP* towards *Manual* and *RegEx* and the inverse. Unlike in the case of complementary precision (Table 2b), in which a closer symmetry can be observed.

5 Discussion & Conclusion

This study aims to provide insights into the relationship between dataset coherence regarding a specific concept and creation heuristics measured with BERT. We found a divergence between creation heuristics (Figure 1), which is larger than the variation within a heuristic, but nonetheless all heuristics perform similar on any *non-native* dataset. We further observed that the bias of *DBLP*, being focused on computer science, is inherited by the classifier (Figure 2) and that the general definition of the *PC* concept associated with the *Manual* dataset yields the most reliable recall across all occupations and domains (Figure 3). This coincides with the bias observed during the annotation of the *Manual* dataset, in which samples presenting publication lists of and articles about a single person were labeled as *PC*. The most effortless approach, utilizing regular expressions, provides a surprisingly reliable solution to the task. But this comes at the cost of an in-depth domain knowledge of the URL structure in the crawl. As such knowledge is often present with researchers analysing specific web archives it could be translated to other tasks.

Some of the problems that lead to a reduction in performance in all classifiers can be found in the annotator agreement and in the boiler-plating mechanism, as well as in the fact that web data is quite noisy. In addition, URLs associated with a person in databases such as DBLP or Wikidata sometimes do not point to a *PC* web page, but to a more general home page or the page of a research group. Such an inaccuracy in the weak annotation stems from the assumption that all official and DBLP-listed web pages are associated with a human entity, and can be regarded as a drawback of the use of existing resources. Overall, we find that the usefulness of focused web archives are the user-made semantic decisions in the structure of URLs, which can be leveraged by experts. Such expert decisions could be used to update rules for evolving concepts, thereby mitigating the influence of concept drift. A follow-up study might use this work for the analysis of the web-related interactions between identified individuals.

Acknowledgments

Parts of this research were funded by the German Federal Ministry of Education and Research (BMBF) in the REGIO project (grant no. 01PU17012D).

References

1. Al-Smadi, M., Qawasmeh, O., Talafha, B., Quwaider, M.: Human annotated arabic dataset of book reviews for aspect based sentiment analysis. In: 3rd International Conference on Future Internet of Things and Cloud. pp. 726–730. IEEE (2015)
2. Baroni, M., Chantree, F., Kilgarriff, A., Sharoff, S.: Cleaneval: a competition for cleaning web pages. In: Proc. of the International Conference on Language Resources and Evaluation. LREC, European Language Resources Association (2008)
3. Ben-David, A., Amram, A.: Computational methods for web history. The SAGE handbook of web history pp. 153–167 (2019)
4. Blank, G.: The digital divide among Twitter users and its implications for social research. Social Science Computer Review **35**(6), 679–697 (2017)
5. Bommasani, R., Cardie, C.: Intrinsic evaluation of summarization datasets. In: Proc. of the Conference on Empirical Methods in Natural Language Processing. pp. 8075–8096. EMNLP, Association for Computational Linguistics (2020)
6. Brill, E., Wu, J.: Classifier combination for improved lexical disambiguation. In: Annual Meeting of the Association for Computational Linguistics. pp. 191–195. Association for Computational Linguistics (1998)
7. Câmara, A., Hauff, C.: Diagnosing BERT with retrieval heuristics. In: European Conference on Information Retrieval. pp. 605–618. Springer (2020)
8. Derczynski, L.: Complementarity, F-score, and NLP evaluation. In: Proc. of the International Conference on Language Resources and Evaluation. pp. 261–266. LREC, European Language Resources Association (2016)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. NAACL, Association for Computational Linguistics, Minneapolis, Minnesota (2019)
10. Ein-Dor, L., Gera, A., Toledo-Ronen, O., Halfon, A., Sznajder, B., Dankin, L., Bilu, Y., Katz, Y., Slonim, N.: Financial event extraction using wikipedia-based weak supervision. Proceedings of the Second Workshop on Economics and Natural Language Processing (2019)
11. Ferrari, A., Spagnolo, G.O., Gnesi, S.: Pure: A dataset of public requirements documents. In: International Requirements Engineering Conference (2017)
12. Gardner, M., Artzi, Y., Basmova, V., Berant, J., Bogin, B., Chen, S., Dasigi, P., Dua, D., Elazar, Y., Gottumukkala, A., Gupta, N., Hajishirzi, H., Ilharco, G.,

Khashabi, D., Lin, K., Liu, J., Liu, N.F., Mulcaire, P., Ning, Q., Singh, S., Smith, N.A., Subramanian, S., Tsarfaty, R., Wallace, E., Zhang, A., Zhou, B.: Evaluating models' local decision boundaries via contrast sets (2020), arxiv:2004.02709

13. Geva, M., Goldberg, Y., Berant, J.: Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets (2019), arXiv:1908.07898
14. International Internet Preservation Consortium (IIPC): The WARC Format 1.1. https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/#warc-file-name-size-and-compression, [Online; Last accessed 11 Mar 2021]
15. Kiesel, J., Kneist, F., Alshomary, M., Stein, B., Hagen, M., Potthast, M.: Reproducible web corpora: Interactive archiving with automatic quality assessment. Journal of Data and Information Quality **10**(4), 1–25 (2018)
16. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., et al.: DBpedia – a large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web **6**(2) (2015)
17. Ley, M.: DBLP: some lessons learned. Proc. of the VLDB Endowment **2**(2), 1493–1500 (2009)
18. Lin, J., Milligan, I., Wiebe, J., Zhou, A.: Warcbase: Scalable analytics infrastructure for exploring web archives. Journal on Computing and Cultural Heritage **10**(4), 1–30 (2017)
19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
20. Lu, K., Mardziel, P., Wu, F., Amancharla, P., Datta, A.: Gender bias in neural natural language processing. In: Logic, Language, and Security. Springer (2020)
21. Mohammad, S.M.: NLP scholar: A dataset for examining the state of NLP research. In: Proc. of the Language Resources and Evaluation Conference. ELRA (2020)
22. Paris, M., Jäschke, R.: How to assess the exhaustiveness of longitudinal web archives. In: Proc. of the Conference on Hypertext and Social Media. ACM (2020)
23. Peng, Y., Yan, S., Lu, Z.: Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. arXiv preprint arXiv:1906.05474 (2019)
24. Qian, Y., Zheng, Q., Sakai, T., Ye, J., Liu, J.: Dynamic author name disambiguation for growing digital libraries. Information Retrieval Journal **18**(5) (2015)
25. Strube, M., Ponzetto, S.P.: Wikirelate! computing semantic relatedness using wikipedia. In: AAAI. vol. 6, pp. 1419–1424 (2006)
26. Vlassenroot, E., Chambers, S., Di Pretoro, E., Geeraert, F., Haesendonck, G., Michel, A., Mechant, P.: Web archives as a data resource for digital scholars. International Journal of Digital Humanities **1**(1), 85–111 (2019)
27. Vogels, T., Ganea, O.E., Eickhoff, C.: Web2text: Deep structured boilerplate removal. In: Advances in Information Retrieval. pp. 167–179. Springer, Cham (2018)
28. Vrandečić, D., Krötzsch, M.: Wikidata: A free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (2014)
29. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Transformers: State-of-the-art natural language processing. In: Proc. of the Conference on Empirical Methods in Natural Language Processing. pp. 38–45. EMNLP, Association for Computational Linguistics (2020)
30. Zesch, T., Müller, C., Gurevych, I.: Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In: LREC. vol. 8, pp. 1646–1652 (2008)