Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210316

Data-Centric Machine Learning: Improving Model Performance and Understanding Through Dataset Analysis

Hannes WESTERMANN [a,1], Jaromír ŠAVELKA [b], Vern R. WALKER [c], Kevin D. ASHLEY [d] and Karim BENYEKHLEF [a]
[a] *Cyberjustice Laboratory, Faculté de droit, Université de Montréal*
[b] *School of Computer Science, Carnegie Mellon University*
[c] *LLT Lab, Maurice A. Deane School of Law, Hofstra University*
[d] *School of Computing and Information, University of Pittsburgh*

Abstract. Machine learning research typically starts with a fixed data set created early in the process. The focus of the experiments is finding a model and training procedure that result in the best possible performance in terms of some selected evaluation metric. This paper explores how changes in a data set influence the measured performance of a model. Using three publicly available data sets from the legal domain, we investigate how changes to their size, the train/test splits, and the human labelling accuracy impact the performance of a trained deep learning classifier. Our experiments suggest that analyzing how data set properties affect performance can be an important step in improving the results of trained classifiers, and leads to better understanding of the obtained results.

Keywords. Classification, Evaluation, Data-centric Approach, Machine Learning, Legal Texts, Semantic Homogeneity

1. Introduction

Two fundamental components of a machine learning (ML) experiment are data and a model. The ML community appears to prefer putting more effort into tweaking the models while spending less time on important data considerations [3]. This means that researchers often invest considerable resources into developing novel models and approaches, achieving marginal improvements. At the same time, they pay much less attention to the properties of the data set (e.g., size, quality, train/test split), or to the effects these might have on the performance of the ML models. Potentially, this under-investigated area of research could lead to significant improvements of the models.

2. Related Work

The ML community has shown increased interest in exploring how data set properties affect trained classifiers. In [3], researchers investigated data-cascades, where issues with

[1]Corresponding Author: Hannes Westermann, E-mail: hannes.westermann@umontreal.ca

data labelling affected downstream systems. In [4], the authors estimated that ten of the most commonly used ML data sets contain an average of 3.4% errors in labelling. AI & Law researchers have investigated data set effects on model performance, including iterative masking of predictive sentences [11], ablating data about criminal charges or sentences [8] and enhancing lawsuit data with ODR data [12]. Researchers have investigated if models can transfer information from single or pooled data sets in different domains [6] or different contexts (languages, jurisdictions and domains) [5].

3. Experimental Design

We evaluated a trained classifier on three data sets annotated on a sentence level under three experimental settings. **Data** We used three publicly available data sets:

- 50 decisions by the U.S. Board of Veterans' Appeals **(BVA)**, containing 6153 sentences tagged with rhetorical roles [9].[2]
- 880 sentences from court opinions mentioning vague statutory terms **(StatInt)**, tagged with usefulness of sentences for statutory interpretation [7].[3]
- 50 opinions of the Supreme Court of India **(ISC)**, containing 9,380 sentences tagged with the rhetorical roles of the sentences [1].[4]

Model We embed each sentence using the Google Universal Sentence Encoder [2] (**GUSE**).[5] We input these embeddings into a two-layer dense neural network classifier (NN model). Full model specs and training procedure are available on github.[6]

Experiments *E1 - Sample-Size Sensitivity:* In this experiment we analyzed the impact of increasing the size of a data set, by first training a classifier on very little data, and then adding more data points each iteration. This allowed us to investigate how adding data to the training set impacts the performance of the classifier, and whether performance trends suggest that adding additional data could be beneficial.

E2 - Split Sensitivity: It is common practice to divide data sets into train and test splits. To investigate the impact of split selection, we split each data set into five folds. In each iteration, four of the folds were used as training split, and the remaining fold as test split. This allowed us to observe the particular split's impact on the performance of the trained NN model, and how much the scores vary on a per-label basis.

E3 - Error Sensitivity: The high-quality of the human labelling is an important concern in ML. To investigate the impact of labelling errors on classifier performance, we started with the original data and then replaced an increasing percentage of the labels with a randomly chosen incorrect label.

4. Results and Discussion

Experiment 1 - Figure 1 shows the evolution of F1-scores for each individual class when adding more and more data. The rates of improvement among the classes are not con-

[2] Dataset available at https://github.com/LLTLab/VetClaims-JSON
[3] Data set available at https://github.com/jsavelka/statutory_interpretation
[4] Data set available at https://github.com/Law-AI/semantic-segmentation
[5] https://tfhub.dev/google/universal-sentence-encoder/4
[6] https://github.com/hwestermann/jurix2021-data_centric_machine_learning

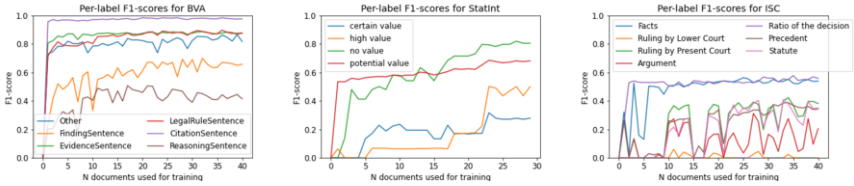

Figure 1. Evolution of per-label F1-score of classifier, as documents are added to training data one by one.

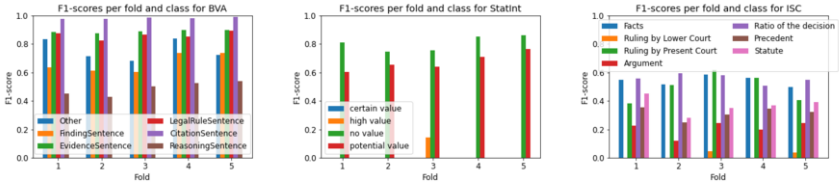

Figure 2. Variations in F1-score for individual classes per five different folds for training and test data.

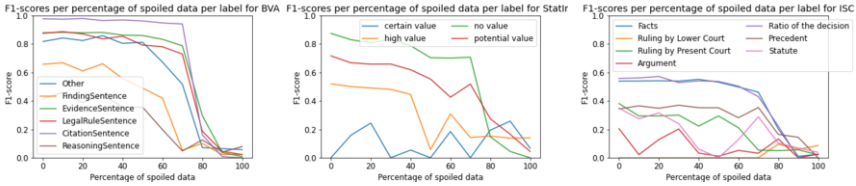

Figure 3. Evolution of per-class F1-scores when replacing set percentage of labels with random label.

sistent. Certain classes reach a high performance quickly, while others take considerably longer. The experiment shows the varying importance of larger data sets. The needed sample size, and whether adding data increases performance, can vary significantly depending on the class and the dataset.

Experiment 2 - Figure 2 shows the scores per class, across the training folds. Some classes' scores differ considerably across the folds. For some use cases, a single train-test split may not produce reliable results when working with legal data sets of limited sizes.

Experiment 3 - Figure 3 shows the effects of randomly mislabeling a portion of the training data set, on a per label basis. Some of the classes start to lose performance quite rapidly, while the others are more resilient to the random errors. In real-world scenarios, human annotators may more likely make systematic errors, increasing the impact.

Class difficulty and semantic homogeneity: It appears that certain labels are significantly more difficult for the classifier to learn than others. In Figure 1, some labels (such as "Citation" for BVA) quickly achieve a high level of performance and then improve more slowly, while other classes require more data to achieve high performance and continually improve (such as "Finding" for BVA). The latter classes also have a more variable performance across folds (Figure 2) and depend more upon high-quality data (Figure 3). Interestingly, this "difficulty" of classes does not fully correspond to the frequency of certain labels appearing in a data set. Rather, it seems related to what we refer to as the *semantic homogeneity* of a class, i.e., how semantically similar the sentences are within a particular class. In [10] we grouped sentences based on semantic similarity

in an embedding space (as determined by Euclidian distance in the GUSE embedding space). For each sentence in a certain class, we explored how many on average of the top 20 most similar sentences were also of that same class. Looking at the table presented in [10], it appears that classes with higher semantic homogeneity are easier to learn for the classifier, and vice-versa. The reason could be that the classifier can more easily find decision boundaries for sentences grouped into clear semantic clusters.

5. Conclusions and Future Work

We trained a classifier on three publicly available data sets, altering the size, training/test split and data labelling quality, to investigate the effects of these properties on ML classifier performance. We observe significant variations in performance over the experiments. These experiments could provide guidance in deciding to continue collecting data, and whether to focus on certain classes during data collection. Our work could represent the initial step in developing a methodology to assess properties of a data set.

Acknowledgments

We are grateful for support from the U. de Montréal Cyberjustice Laboratory, LexUM Chair on Legal Information, and Autonomy through Cyberjustice Technologies project.

References

[1] Bhattacharya, P., Paul, S., Ghosh, K., Ghosh, S. & Wyner, A. "Identification of Rhetorical Roles of Sentences in Indian Legal Judgments." *Jurix 2019*. pp. 3-12 (2019).
[2] Cer, D., Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. St John, N. Constant et al. "Universal sentence encoder." *arXiv preprint arXiv*:1803.11175 (2018).
[3] Sambasivan, N. et al. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. *2021 CHI Conference* pp. 1-15 (2021).
[4] Northcutt, C. G., Athalye, A. & Mueller, J. "Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks." *arXiv:2103.14749* [cs, stat] (2021).
[5] Savelka, J., Westermann, H., Benyekhlef, K. et al. "Lex Rosetta: transfer of predictive models across languages, jurisdictions, and legal domains." In *ICAIL 2021*, pp. 129–138. (2021).
[6] Savelka, J., Westermann, H. & Benyekhlef, K. "Cross-Domain Generalization and Knowledge Transfer in Transformers Trained on Legal Data." *ASAIL@ Jurix.* (2020).
[7] Šavelka, J., Xu, H., & Ashley, K. "Improving Sentence Retrieval from Case Law for Statutory Interpretation." *Proc. 17th Int'l Conf. on Artificial Intelligence and Law*, pp. 113-122. (2019).
[8] Tan, H., Zhang, B., Zhang, H., & Li, R. "The Sentencing-Element-Aware Model for Explainable Term-of-Penalty Prediction". In *CCF Int'l Conf. on NLP and Chinese Computing*. pp. 16-27. (2020).
[9] Walker, V. R., et al. "Automatic Classification of Rhetorical Roles for Sentences: Comparing Rule-Based Scripts with Machine Learning." *Proceedings of ASAIL 2019* (2019).
[10] Westermann, H., Savelka, J., Walker, V., Ashley, K. & Benyekhlef, K. "Sentence Embeddings and High-Speed Similarity Search for Fast Computer Assisted Annotation of Legal Documents." *Jurix*. (2020).
[11] Zhong, L., Zhong, Z., Zhao, Z., Wang, S., Ashley, K. D., & Grabmair, M. "Automatic summarization of legal decisions using iterative masking of predictive sentences." *ICAIL 2019*, pp. 163-172. (2019).
[12] Zhou, X., Zhang, Y., Liu, X., Sun, C., & Si, L. "Legal Intelligence for E-commerce: Multi-task Learning by Leveraging Multiview Dispute Representation." In *Proc. 42nd Int'l ACM SIGIR*, pp. 315-324. (2019).