Interactive Dashboard for Text Labeling: Improving Quality, Reliability and Fairness of Automatic Annotators

Aisha Khatun and Mustapha Unubi Momoh

Abstract—This study aims to address the issue of unreliable labels produced by automatic annotators and machine learning models in text data annotation. The traditional tools are focused on model evaluation or interpretation and show several limitations. Our solution proposes an interactive dashboard to visualize the natural clusters in the data and assess the machine learning models used in text data annotation. The dashboard will include a scatter plot of data points, a side panel with details of each instance, options to view the scatter plot using different techniques, filters to view data points by labels and topics, and more. The goal of the dashboard is to help identify mislabeled samples, perform semantic de-duplication, and enhance the performance of frameworks used in text data annotation.

Index Terms—labeling, text analysis, NLP

1 Introduction

The goal of troubleshooting automatic annotators and machine learning models is to enhance their generalizability, reliability, and robustness, as well as to ensure the quality, reliability, and fairness of the labels they produce. Building trust in these models and their outputs is also crucial. Studies have shown that visualizing predictions is an effective way to interpret and diagnose a model [5, 11]. Inconsistent predictions that are not due to adversarial inputs often indicate a lack of robustness in the model. Consistent predictions and labeling by the model have various benefits, including building trust and reliability, enabling improved decision-making, reducing biases, and ensuring repeatability and comparability.

2 Motivation

The motivation behind this work stems from the lack of robust tools for diagnosing mislabeling in text datasets. The traditional tools are focused on model evaluation or interpretation and show several limitations. These methods tend to focus on uncovering the inner workings of specific model types, such as deep neural networks, but they often fall short in accommodating more complex scenarios involving multiple model types [11]. Additionally, most popular evaluation frameworks are non-agnostic and have been found to be challenging to use or require expert knowledge. For instance, interactive tools such as CNN explainer [10] and [8] are designed specifically for diagnosing convolutional neural networks (CNN) and are not compatible with models such as Bert. Frameworks like RNNvis [7] for NLP have limited interactivity as the visualizations produced are static and become increasingly difficult to comprehend for non-experts as model complexity increases. Other common NLP diagnostic tools such as Bertviz [9], Ecco [1], and Erudite [4] also have limitations in terms of compatibility with other language models, scalability for complex tasks, and ability to display the relationship between words and concepts, respectively.

3 Significance

Our solution will facilitate efficient and accurate text labeling in scenarios where labels are unclear and multiple manual attempts on datasets are necessary to identify mislabeling. Furthermore, our user-friendly framework will improve the assessment of machine learning models and frameworks used in text data annotation. Our tool is expected to enhance the performance of frameworks such as PIRATES [2] when integrated into the annotation process.

• *Aisha Khatun. E-mail: a2khatun@uwaterloo.ca*
• *Mustapha Unubi Momoh. E-mail: mmomoh@uwaterloo.ca*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

4 Proposed Solution

To solve the labeling inconsistencies, we propose an interactive dashboard to view the natural clusters in the data with the details of each data point. The intention is that users will be able to analyze the data in or near clusters and decide whether the labels they set were correct or need changing. Users will also be able to identify obvious outliers in clusters and decide to create new labels or re-think their labeling scheme. We first create embeddings for sentences using various techniques such as Doc2Vec [6] and Universal Sentence Embedding [3]. We find samples close to each sample in the dataset to help users perform semantic de-duplication if required. Next, we perform dimentionality reduction (to 2 dimensions) using various techniques such as PCA, SVD, and tSNE and plot the points to reveal clusters. Further we perform topic classification using NMF or LDA and gather topic words for each topic detected. All these data are then organized on the dashboard for users to gather insights from. We list the features for the dashboard in detail below.

1. The main view of the page includes a scatter plot of data points created using dimentionality reduction techniques. The dots are colored by given labels. These labels are assumed to be unreliable and the user can assess from the clusters whether to change some label or delete data points entirely.
2. Clicking each data point reveals a side panel that shows details of the instance. This includes the text, label, nearest neighbours (using cosine similarity of embeddings), topic words and topic word cloud.
3. We give users the option to view the scatter plot created using various techniques from a drop down menu which includes PCA, SVD, and tSNE. More techniques may be included if found.
4. User can select only the labels they wish to view using a checklist. This view filters the data to contain points only in the set of labels the user selects.
5. Another checklist lets users select by topics instead of labels. This will reveal whether certain topics contain a mix of labels or a subset of data points of a single label.
6. There can be a few more filters based on the attributes of the dataset such as sub-labels, data source, date curated etc. We can allow choice between various sentence embedding techniques used as well.

Overall, the visualization is intended to help text labeling for situations where labels are not well defined and requires multiple iterations on

the dataset to reveal whether each sample is correctly labeled. A visual aid using clustering and topic analysis can help identify mislabeled samples and perform semantic de-duplication where necessary.

ACKNOWLEDGEMENT

Aisha worked on Proposed Solution section, and on defining the project idea. Mustapha worked on Abstract, Introduction, Motivation, and Significance sections.

REFERENCES

[1] J. Alammar. Ecco: An open source library for the explainability of transformer language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 2021. 1

[2] S. C. Bayliss, H. A. Thorpe, N. M. Coyle, S. K. Sheppard, and E. J. Feil. PIRATE: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. *GigaScience*, 8(10), 10 2019. giz119. doi: 10.1093/gigascience/giz119 1

[3] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, and R. Kurzweil. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 169–174. Association for Computational Linguistics, Brussels, Belgium, Nov. 2018. doi: 10.18653/v1/D18-2029 1

[4] L. De Cicco, G. Cilli, and S. Mascolo. Erudite: A deep neural network for optimal tuning of adaptive video streaming controllers. In *Proceedings of the 10th ACM Multimedia Systems Conference*, MMSys '19, p. 13–24. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3304109.3306216 1

[5] J. Krause, A. Perer, and K. Ng. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, p. 5686–5697. Association for Computing Machinery, New York, NY, USA, 2016. doi: 10.1145/2858036.2858529 1

[6] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In E. P. Xing and T. Jebara, eds., *Proceedings of the 31st International Conference on Machine Learning*, vol. 32 of *Proceedings of Machine Learning Research*, pp. 1188–1196. PMLR, Bejing, China, 22–24 Jun 2014. 1

[7] Y. Ming, S. Cao, R. Zhang, Z. Li, Y. Chen, Y. Song, and H. Qu. Understanding hidden memories of recurrent neural networks, 2017. doi: 10.48550/ARXIV.1710.10777 1

[8] C. Olah, A. Mordvintsev, and L. Schubert. Feature visualization. *Distill*, 2017. https://distill.pub/2017/feature-visualization. doi: 10.23915/distill.00007 1

[9] J. Vig. Visualizing attention in transformer-based language representation models, 2019. doi: 10.48550/ARXIV.1904.02679 1

[10] Z. J. Wang, R. Turko, O. Shaikh, H. Park, N. Das, F. Hohman, M. Kahng, and D. H. Polo Chau. CNN explainer: Learning convolutional neural networks with interactive visualization. vol. 27, pp. 1396–1406. Institute of Electrical and Electronics Engineers (IEEE), Feb. 2021. 1

[11] J. Zhang, Y. Wang, P. Molino, L. Li, and D. S. Ebert. Manifold: A model-agnostic framework for interpretation and diagnosis of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):364–373, jan 2019. doi: 10.1109/TVCG.2018.2864499 1