Tightening the Loop in Mixed-Initiative ML Engineering and Domain Annotation using Active Learning and Visual Analytics

Mert Erkul[1]*, Piriyakorn Piriyatamwong[1]*, Batuhan Tomekce[1]*, Manuel Morales Wyden[1], William A Baumgartner Jr[2], Elizabeth White[2], Michael Bada[2], Lawrence Hunter[2], and Mennatallah El-Assady[1]

[1]ETH Zurich [2]University of Colorado Anschutz Medical Campus

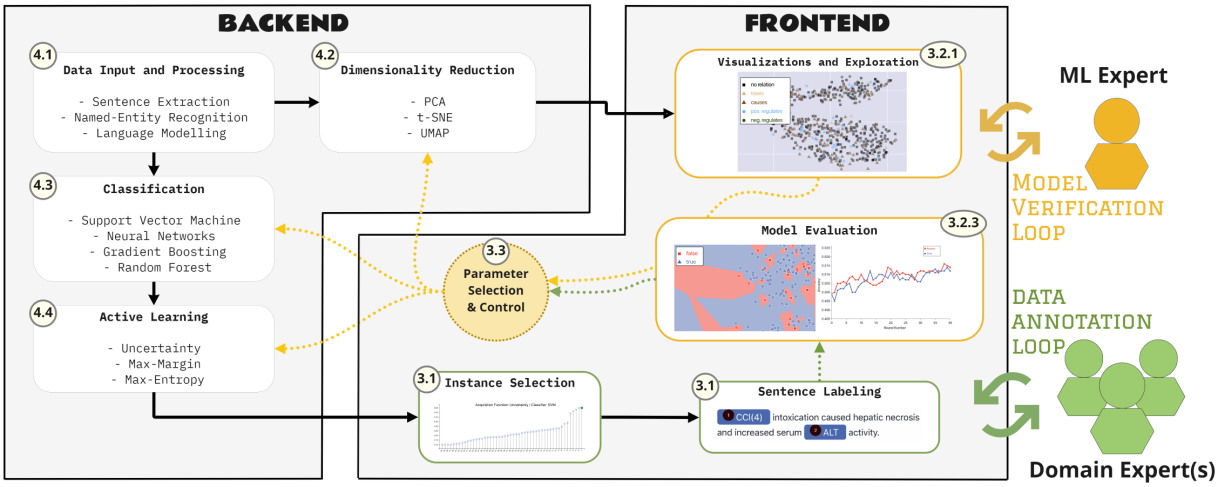


Figure 1: Overall dashboard workflow: domain experts label the data while ML experts explore, analyze, and select model parameters. The numbers in the circles show the corresponding Section in our paper.

Abstract

Annotating data is often a very expensive, yet an indispensable task for modern-day supervised deep learning (DL), requiring interaction between experts and systems. In this paper, we propose a mixed-initiative workflow and present a specific application, targeted towards both the annotators and the machine learning (ML) experts, aiming to bridge the gap between two different stakeholders. With our work, we contribute a dashboard utilizing Active Learning (AL) for annotators–domain experts in medicine–and ML experts containing the complete labeling workflow of relations between entities in biomedical sentences. Our dashboard offers various AL strategies, ML algorithms as well as dimensionality reduction techniques to explore samples and select items that have the biggest impact on the generalization performance after labeling and retraining, while providing a modular design for different use case scenarios.

Keywords: Mixed-initiative systems, active learning, visual analytics, biomedical text annotation

1 Introduction

Understanding the relations between genes, diseases, and chemicals, has been a key goal of modern-day biomedical research [19]. While many of these relationships, e.g., the genetic cause of a disease or how a chemical may regulate a particular gene, have been captured in structured resources, curation of this knowledge often lags behind the rate at which the scientific community produces it [3]. Building automated text mining systems to extract such knowledge from the biomedical literature is one solution to address this lag in knowledge curation. The development of such systems requires the coordination of domain experts trained in the biomedical sciences to label data for training computational models, and machine learning engineers capable of building systems based on the training data to automatically identify entities and relations between them in biomedical text. The work described here facilitates the collaboration of these different stakeholders by providing a generalizable platform that increases domain expert annotation efficiency by prioritizing the annotation of sentences in the data set.

Recent advancements in natural language processing (NLP) have enabled the successful extraction of valuable lexical and semantic relations in documents [35]. Human-comparable performances in these tasks can be obtained by utilizing word embeddings, sentence embeddings, or contextual representations [28], which are extracted using pre-trained language models such as BERT [10]. Transfer learning (TL) strategies often use a fine-tuning stage to boost performance [10], requiring labeled samples for the downstream tasks.

Obtaining numerous labeled samples is challenging and task/domain-specific. Active learning [20] is an emerging field that aims to decrease the number of labeled samples that is required to achieve the necessary model performance. Therefore, by employing NLP, AL, and ML, we enhance our biomedical annotation dashboard with intelligent tools, to ease the labeling process of domain experts and to increase the performance of the relation classifier as much as possible with as few samples required to be labeled as possible. Our dashboard is designed to support annotators and machine

E-mails of the authors affiliated with ETH Zurich: [1]{merkul, ppiriyata, tbatuhan, mamorales, melassad}@ethz.ch

E-mails of the authors affiliated with University of Colorado Anschutz Medical Campus: [2]{william.baumgartner, elizabeth.white, mike.bada, larry.hunter}@cuanschutz.edu

* Authors contributed equally to the paper.

learning engineers, with two user interfaces supporting data exploration and downstream functionality such as model re-training and performance visualization as the labeling continues.

With the interactive design in mind, our dashboard provides an out-of-the-box labeling tool for understanding biomedical text data and the dynamics of training the base machine learning algorithm powered by active learning. Our main contribution with this paper is a modular workflow, deploying AL, ML, and dimensionality reduction techniques, while addressing the needs of multiple stakeholders, for simultaneous and concurrent intelligence augmentation and performance evaluation. This way, we tighten the loop between data annotation and model verification.

In the following subsections, we present the development of our smart biomedical labeling dashboard in terms of visualization design choices and the mechanisms behind it.

2 Literature Review

The dashboard and the workflow we propose contain three concepts that are investigated frequently in the literature, namely, annotation, textual data projection and active learning.

Annotation – Interactive machine learning has enabled new possibilities for creating labeling platforms. Regarding textual annotation tasks, various NLP-based solutions have been developed in recent years. Both *BRAT* [33] and *VIANA* [32] for instance, let users annotate (named entities or parts-of-speech tags) by directly clicking on words or dragging over various textual parts. Some platforms make suggestions on the parts to be labeled, e.g., with temporal events (e.g., *TimeLineCurator* [16]) or general entity extraction (e.g., *Anafora* [9]), use gamification to engage the users (e.g., *QuestionComb* [30]), or pre-select a label for the user (e.g., *GATE Teamware* [6]). As our task specifically focuses on labeling the relations between entities, we highlight the entities during selection, but follow a labeling functionality similar to *BRAT* and *VIANA*. We do not suggest any annotations to the user as this could lead to selection bias [14, 31].

Active Learning and Interaction – In order to take advantage of the AL benefits, there have also been advances to integrate AL to interactive labeling dashboards. One of them is *MONAI Label* [11] a platform using a combination of AL and ML to annotate 3D biomedical images. An established closed source example is *Prodigy* [13] using AL for interactive annotation tasks of different types of data. Another example is *PAL* [31], an extension to *BRAT* which adds active learning and pre-annotation on the input data but without showing any performance measures nor being directly applied to biomedical data. AL and interactivity using lower-dimensional representations of samples have been implemented in literature [4, 5]. Such tools allow the users to label by automatically selecting the sample with the highest acquisition value and demonstrating different aspects such as Voronoi tessellations [4], convex hulls, color maps, or butterfly plots [5] on the projected two-dimensional space, as an indicator of neighborhood structures.

Projecting Textual Data – Projecting textual data on a lower-dimensional space for exploration and refinement has been applied frequently in the visual analytics literature. *Semantic Concept Spaces* [12] provide projections of the word embeddings for topic model refinement on a two-dimensional space. *DocuCompass* [18] framework offers the user to see documents on a two-dimensional space with a lens feature to investigate documents' characteristic labels while preserving distance measures of the documents with respect to similarity measures. For hierarchical topic exploration, *TopicLens* was proposed [22], offering a lens feature to recompute the topic model in a finer-grained structure for user selection, color coding, and clustering based on similarities of the identified topics in large documents. For visualization of semantic relations between word embeddings, *Word Embedding Visual Explorer* [25] was built, presenting global and pairwise projections on lower dimensions.

3 User Interface and Workflow Design

In order to cover the workflow of biomedical relation annotation, our dashboard is divided into two modes: (1) labeling mode and (2) discovery mode. These two modes have common components that are always visible and different components that are not visible in the other mode. The different components are developed to separate the ML expert view and domain expert view of our dashboard from each other to make the interaction as efficient as possible. In their individual subsections, we explain the different components and delineate the common components in the last subsection. An overview of the workflow from the dashboard can be seen in Figure 1.

3.1 Labeling Mode

In the labeling mode, annotators interactively label the sentences that have the highest acquisition values. The dashboard welcomes the user with a sorted Acquisition Value plot in Figure 2 and a section as seen in Figure 3 that directly refers the domain experts to the sample with the highest acquisition value, calculated using the selected ML model and AL strategy. We use vertical lollipop plots to emphasize the differences in acquisition values among the best 50 samples (see Figure 2). A "Next" button is also included, to continue labeling the next sample with the highest acquisition value. Furthermore, the user can interact with the points by hovering the mouse over them to see the corresponding sentence inside a tooltip.

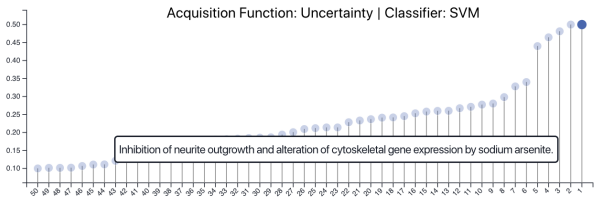


Figure 2: Acquisition Value plot example from the dashboard.

In the sentence labeling section shown in Figure 3, the pre-extracted entities are bolded and underlined, and the annotator can directly click on the text to declare one of the entities for labeling. Once two entities are chosen, a pop-up window asks the annotator to assign the relation label for the chosen entity pair. Afterward, the chosen relation label is automatically recorded in the backend database and the next sample is displayed.

3.2 Discovery Mode

Our discovery mode is for machine learning experts to learn about model dynamics and performance as the labeling process continues.

Dimensionality Reduction Plot – In the discovery mode, the machine learning expert is first presented with data exploration visualizations. The labeled sentences' embeddings are on a two-dimensional interactive zoomable plot, projected by the selected dimensionality reduction technique. The data point is linked to its sentence, and a tooltip shows the corresponding sentence when the mouse hovers. The selected sentence is highlighted in the plot. An example of the t-SNE reduction plot is shown in Figure 4a.

Sentence – On the right a table, that is displayed in the sentence panel, shows example sentences, their labels, and indices, which the users can use to discover the sentences in the dataset by using the 'Next' and 'Previous' buttons. Our dashboard keeps which sentence is selected, and the sentences can be selected either from the plots or from the table. Below the table, the user can find the selected

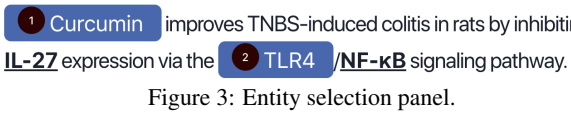
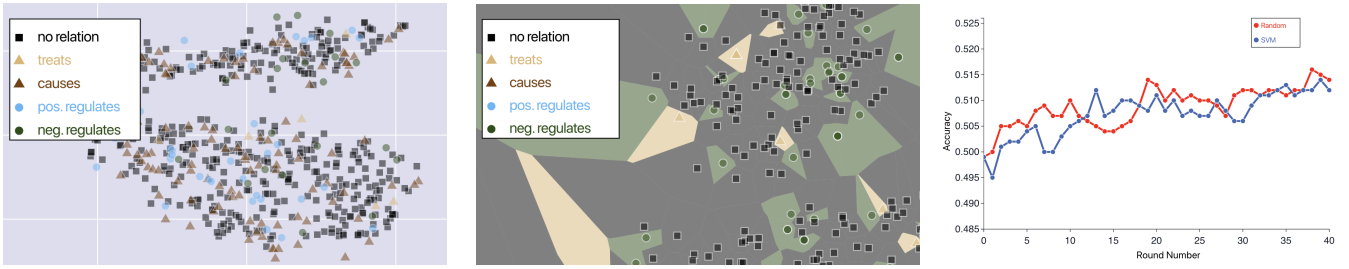


Figure 3: Entity selection panel.



(a) t-SNE exploration (b) t-SNE + Voronoi polygons on the test set (c) Accuracy plot against random labeling

Figure 4: Panels from the discovery mode. Machine learning experts can explore the data set through various dimension reduction techniques, see their decision boundaries in the Voronoi tessellations, and check the performance by simulating labeling rounds.

sentence and highlight its entities by hovering over the sentence.

Understanding the Classifiers – Furthermore, in the discovery mode, users can see and interact with the Voronoi tessellation of the test set samples projected on a 2-dimensional space. The polygons are filled with respect to three different modes, such that the Voronoi panel shows (0) actual labels, (1) predicted labels, or (2) predictions compared against the ground truth, based on user selection (see Figure 4b for the t-SNE Voronoi of mode 0). Furthermore, the user can visualize how the decision boundaries of the classifiers based on the Voronoi tessellation changes with additional labeled samples by going over different rounds of labeling that come from the simulations ran in the backend. By visualizing the different rounds, the user can observe changes in the decision boundaries and, by hovering over the point, learn which sentence it corresponded to.

Comparing Strategies – The performance of the selected model and the AL strategy is displayed on the right half of the panel, where the ML expert can compare AL strategy performance with random labeling based on different rounds of labeling and retraining. In Figure 4c, the accuracy of the Max-Entropy algorithm for SVM classifiers is displayed against random labeling over 40 simulation rounds. For further details on the simulations and performances, see Section 5.

3.3 Common Components in All Modes

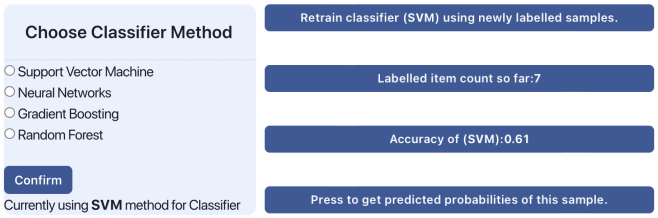
The user sees an interactive tutorial [2] when they enter the page for the first time, which can be reopened if desired. The tutorial contains GIFs to show users how to interact with the components, also highlighting the corresponding component. The icons showing the classifiers are interactive, their sizes enlarge on hover, and the user can choose the desired selection by clicking on them, as well as from the drawer by clicking on the menu icon on the top right.

The glossary component includes intuitive explanations for the interactive components regardless of the mode, for both the ML expert and the domain expert to read. Figure 5 shows a sample selection box and the query buttons located in the drawer. Here, multiple domain experts and machine learning engineers can synchronize on labeling status. They can see how many data points have been labeled since the model was last retrained, start a retraining process with the addition of the new labels, and display the predicted class probabilities for the selected sample.

The whole dashboard is designed to be full-screen scrollable (using fullPage.js [1]) to ensure that the user focuses on the desired panels, especially in the discovery mode.

4 Biomedical Text Data and ML Methods

In the following section, we explain the dataset and the methods used in our current implementation. However, the workflow of the dashboard is designed in a way that it could be enhanced with further methods or applied to other datasets without major changes to the user interface described above.



(a) classifier choice (b) query buttons

Figure 5: Sample options from the common drawer.

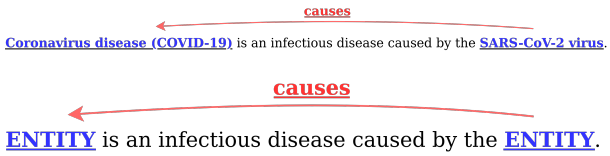


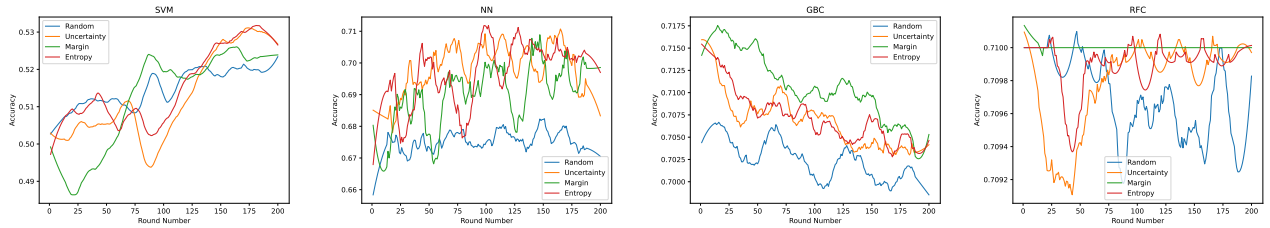
Figure 6: Sample sentence, before and after entities are replaced with placeholders.

4.1 Dataset, Preprocessing, and Feature Engineering

Dataset – Our dataset consists of 12,128 samples from *PubMed* abstracts or *PubMed Central* articles. Each sample contains a sentence, an entity pair from the sentence, and its relation label. The possible biomedical ontologies that form the entities are chemicals, diseases, or genes. There are 5 possible relation labels: (0) no relation, (1) treats, (2) causes, (3) positively regulates, and (4) negatively regulates. Entities and labels were mined from in-house annotation efforts as well as the ChemProt [23] and GeneReg [8] corpora.

Preprocessing – Entities in the sample sentences are extracted automatically. Sentences can have more than one entity pair and labels are specific to a given entity pair, which means that one extracted sentence from an article can create more than one data point. To simulate a deployment scenario, we split the dataset in a stratified fashion, so that the class distributions in "labeled" and "unlabeled" sets were preserved. We treated 6,066 of the samples as unlabeled and 6,062 as labeled, trained the models on the labeled set, conducted the simulations on the test set, formed by 1,000 samples randomly selected from the "unlabeled" set, and visualize the first 500 in the dashboard.

Feature Engineering – Using the original dataset, we create the feature vectors by replacing the pre-extracted entities with the placeholder "ENTITY" as in Figure 6. We then propagate these sentences with placeholders through BlueBERT, a pre-trained BERT model over *PubMed* abstracts and articles, which is precisely our data source [28], to obtain 768-dimensional vectors. Our main goal while using placeholders instead of exact entity words is to generalize for the contextual meaning in the sentences in an entity agnostic fashion.



(a) SVM (b) NN (c) GBC (d) RFC

Figure 7: Simulations comparing AL strategies against random sample selection.

4.2 Descriptive Overview and Dimensionality Reduction

To visualize the BlueBERT [28] embeddings of the labeled (already labeled and newly labeled) sentences in a 2-dimensional vector space, we offer three different dimensionality reduction algorithms to the user to select. The selected algorithms are **Principal Component Analysis (PCA)** [15], **t-Distributed Stochastic Neighbor Embedding (t-SNE)** [34], and **Uniform Manifold Approximation and Projection (UMAP)** [27] which helps the user understand different properties of the original high dimensional vector space.

4.3 ML Classifiers and AL Strategies

Base Classifiers – We want to keep the dashboard as interactive as possible while also maintaining model performance. To this extent, also acknowledging that our training dataset is small, we try to avoid deep models with many parameters. Instead, we utilize classifiers that can successfully identify uncertainty. Trying to satisfy all constraints, we choose **Support Vector Machines (SVM)** [17], **Random Forest Classifiers (RFC)** [7], **Neural Networks (NN)** [26], and **Gradient Boosting Classifiers (GBC)** [21] as our base classifiers. These classifiers are used in our dashboard to infer prediction probabilities of the unlabeled sentences to calculate acquisition values and can be selectively changed by the user.

AL Strategies – In modern active learning, base classifiers are used to identify the uncertainty in unlabeled sentences; then different strategies are used to compute and assign "acquisition values" to the unlabeled sentences, that are considered to indicate which samples to label first [20]. We choose three of the most widely used techniques to assign acquisition values to the unlabeled sentences [24], namely **Uncertainty Sampling**, **Margin Sampling**, and **Max-Entropy Sampling**. We also allow stakeholders to select the desired strategy to calculate the acquisition value for the unlabeled samples in our dashboard.

5 Use Case and Results

Our dashboard finds direct application in tasks that need labeling of biomedical relations in text. The combination of ML experts and domain experts in our workflow thereby allows for more fine-tuning and quality control. In order to compare the Active Learning strategies and classifiers, we ran labeling and retraining simulations. We decided on a 200-round simulation where each round includes 25 additional samples added to the labeled dataset based on which strategy is employed. We ran every classifier and AL strategy combination and tested on a leave-out set with 1,000 samples with ground truth labels. The simulation plots can be seen in Figure 7.

The figures demonstrate the insignificance of performance differences. The imbalance in the dataset forces most of the models to always predict the majority class, even though we have utilized class weighting and different oversampling methods to make up for the imbalance. We also tried binary classification, comparing class 0 (no relation exists for the selected entity pair) against other classes, which did not demonstrate significant differences between random sampling and AL either. To satisfy the expectations, we kept the multi-class scenario in our dashboard, as well as the simulations.

Moreover, the results, especially observed with ensemble classifiers, are unexpected. Gradient Boosting performance decreases with more samples being labeled, whereas Random Forest does not demonstrate any changes. Support Vector Machine is the most stable, with all AL strategies surpassing random labeling with a small margin after round 125, while showing an upwards trend, which is expected as the classifier should theoretically generalize better with increased training set size. Neural Network accuracy scores against rounds are observed to be too noisy to interpret, with AL strategies showing better performance than random sampling.

Moreover, AL guarantees have been proven [29] for the scenario where the classifiers are re-calibrated using one sample each round. As we believed that this would hinder the interactivity of the dashboard and is also unrealistic to use with larger base classifiers in the future, we selected the top 25 samples each round before retraining, which might also constrain the simulation performance.

6 Conclusion

In this work we describe an annotation dashboard for sentences from the biomedical literature. The dashboard supports the user with two modes designed for domain experts and ML experts. It offers a wide variety of AL techniques and ML classifiers and allows the users to select different dimensionality-reduction algorithms to visualize the dataset. Furthermore, users can utilize the simulation result visualization, decision boundaries of classifiers, and retrain the models in the backend at their discretion.

As future research, we plan to investigate different classifiers such as Transformers and different explainability techniques. Extracting more word-level features from the sentences may also be helpful. Fine-tuning BlueBERT would be another direction that would also update the visualizations obtained via dimensionality reduction. Regarding AL, a future direction would be to explore committee learning as an additional sampling technique, where samples are selected based on their disagreement score calculated using several base classifiers' soft or hard voting results. To strengthen the communication between the domain experts and machine learning engineers, we would also like to create a "run simulation" button, which with a click, runs all simulations on newly added data.

In terms of the usability of our dashboard, we would like to conduct experiments with annotators and ML experts. Based on their feedback, we plan to enhance our dashboard with additional tools, such as flagging options for users to indicate wrong entity extraction. Another additional future direction can be the incorporation of an RL agent to automatically select the best performing AL/ML combination for sample selection.

Acknowledgments

Partial support for this work was provided by the ETH AI Center. Additionally, another partial support for this work was provided by the National Center for Advancing Translational Sciences, National Institutes of Health, through the Biomedical Data Translator program, award #OT2TR003422. Any opinions expressed in this document are those of the Translator community at large and do not necessarily reflect the views of NCATS, individual Translator team members, or affiliated organizations and institutions.

REFERENCES

[1] fullPage.js. Retrieved from https://alvarotrigo.com/fullPage/, 2022.

[2] Intro.js. Retrieved from https://introjs.com, 2022.

[3] W. A. Baumgartner Jr, K. B. Cohen, L. M. Fox, G. Acquaah-Mensah, and L. Hunter. Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics*, 23(13):i41–i48, 2007.

[4] R. Beckmann, C. Blaga, M. El-Assady, M. Zeppelzauer, and J. Bernard. Interactive visual explanation of incremental data labeling. In J. Bernard and M. Angelini, eds., *EuroVis Workshop on Visual Analytics (EuroVA)*. The Eurographics Association, 2022. doi: 10.2312/eurova.20221073

[5] J. Bernard, M. Hutter, M. Zeppelzauer, D. Fellner, and M. Sedlmair. Comparing visual-interactive labeling with active learning: An experimental study. *IEEE Trans. on Visualization and Computer Graphics*, 24(1):298–308, 2018. doi: 10.1109/TVCG.2017.2744818

[6] K. Bontcheva, H. Cunningham, I. Roberts, A. Roberts, V. Tablan, N. Aswani, and G. Gorrell. Gate teamware: A web-based collaborative text annotation framework. *Language Resources and Evaluation*, 47, 12 2013. doi: 10.1007/s10579-013-9215-6

[7] L. Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, oct 2001. doi: 10.1023/A:1010933404324

[8] E. Buyko, E. Beisswanger, and U. Hahn. The genereg corpus for gene expression regulation events—an overview of the corpus and its in-domain and out-of-domain interoperability. In *Proc. of Int. Conf. on Language Resources and Evaluation (LREC)*, 2010.

[9] W.-T. Chen and W. Styler. Anafora: A web-based general purpose annotation tool. In *Proc. of NAACL HLT Demonstration Session*, pp. 14–19. Association for Computational Linguistics, June 2013.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, June 2019. doi: 10.18653/v1/N19-1423

[11] A. Diaz-Pinto, S. Alle, A. Ihsani, M. Asad, V. Nath, F. Pérez-García, P. Mehta, W. Li, H. R. Roth, T. Vercauteren, D. Xu, P. Dogra, S. Ourselin, A. Feng, and M. J. Cardoso. Monai label: A framework for ai-assisted interactive labeling of 3d medical images, 2022. doi: 10.48550/ARXIV.2203.12362

[12] M. El-Assady, R. Kehlbeck, C. Collins, D. Keim, and O. Deussen. Semantic concept spaces: Guided topic model refinement using word-embedding projections. 2019. doi: 10.48550/ARXIV.1908.00475

[13] Explosion. Prodigy. Retrieved from https://prodi.gy, 2022.

[14] K. Fort and B. Sagot. Influence of pre-annotation on pos-tagged corpus development. In *The Fourth ACL Linguistic Annotation Workshop*, pp. 56–63, July 2010.

[15] K. P. F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. doi: 10.1080/14786440109462720

[16] J. Fulda, M. Brehmer, and T. Munzner. Timelinecurator: Interactive authoring of visual timelines from unstructured text. *IEEE Trans. on Visualization and Computer Graphics*, 22(1):300–309, 2016. doi: 10.1109/TVCG.2015.2467531

[17] M. Hearst, S. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. vol. 13, pp. 18–28. 1998. doi: 10.1109/5254.708428

[18] F. Heimerl, M. John, Q. Han, S. Koch, and T. Ertl. Docucompass: Effective exploration of document landscapes. pp. 11–20, 2016. doi: 10.1109/VAST.2016.7883507

[19] M. Jackson, L. Marks, G. May, and J. Wilson. The genetic basis of disease. *Essays In Biochemistry*, 62:643–723, 12 2018. doi: 10.1042/EBC20170053

[20] P. F. Jacobs, G. Maillette de Buy Wenniger, M. Wiering, and L. Schomaker. Active learning for reducing labeling effort in text classification tasks. In L. A. Leiva, C. Pruski, R. Markovich, A. Najjar, and C. Schommer, eds., *Artificial Intelligence and Machine Learning*, pp. 3–29. Springer Int. Publishing, 2022.

[21] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[22] M. Kim, K. Kang, D. Park, J. Choo, and N. Elmqvist. Topiclens: Efficient multi-level visual topic exploration of large-scale document collections. *IEEE Trans. on Visualization and Computer Graphics*, 23(1):151–160, 2017. doi: 10.1109/TVCG.2016.2598445

[23] M. Krallinger, O. Rabal, S. A. Akhondi, M. P. Pérez, J. Santamaría, G. P. Rodríguez, G. Tsatsaronis, A. Intxaurrondo, J. A. López, U. Nandal, et al. Overview of the biocreative vi chemical-protein interaction track. In *Proc. of BioCreative Challenge Evaluation Workshop*, vol. 1, pp. 141–146, 2017.

[24] D. D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. In W. W. Cohen and H. Hirsh, eds., *Machine Learning Proc. 1994*, pp. 148–156. Morgan Kaufmann, 1994. doi: 10.1016/B978-1-55860-335-6.50026-X

[25] S. Liu, P.-T. Bremer, J. J. Thiagarajan, V. Srikumar, B. Wang, Y. Livnat, and V. Pascucci. Visual exploration of semantic relationships in neural word embeddings. *IEEE Trans. on Visualization and Computer Graphics*, 24(1):553–562, 2018. doi: 10.1109/TVCG.2017.2745141

[26] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133, 1943.

[27] L. McInnes, J. Healy, N. Saul, and L. Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861

[28] Y. Peng, S. Yan, and Z. Lu. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proc. of Workshop on Biomedical Natural Language Processing (BioNLP)*, pp. 58–65, 2019.

[29] B. Settles. Active learning literature survey, 2009.

[30] R. Sevastjanova, W. Jentner, F. Sperrle, R. Kehlbeck, J. Bernard, and M. El-assady. Questioncomb: A gamification approach for the visual explanation of linguistic phenomena through interactive labeling. *ACM Trans. Interact. Intell. Syst.*, 11(3–4), aug 2021. doi: 10.1145/3429448

[31] M. Skeppstedt, C. Paradis, and A. Kerren. Pal, a tool for pre-annotation and active learning. *Journal for Language Technology and Computational Linguistics*, 31(1):91–110, 2017.

[32] F. Sperrle, R. Sevastjanova, R. Kehlbeck, and M. El-Assady. Viana: Visual interactive annotation of argumentation. 2019. doi: 10.48550/ARXIV.1907.12413

[33] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. brat: a web-based tool for NLP-assisted text annotation. In *Proc. of the Demonstrations at Conf. of the European Chapter of the Association for Computational Linguistics*, pp. 102–107. Association for Computational Linguistics, Apr. 2012.

[34] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[35] C. Wang, X. He, and A. Zhou. SphereRE: Distinguishing lexical relations with hyperspherical relation embeddings. In *Proc. of Annual Meeting of the Association for Computational Linguistics*, pp. 1727–1737. Association for Computational Linguistics, July 2019. doi: 10.18653/v1/P19-1169