Explainable e-Discovery (XeD) Using an Interpretable Fuzzy ARTMAP Neural Network for Technology-Assisted Review

Charles Courchaine
National University
United States
charles@courchaine.dev

Tasnova Tabassum
Fitchburg State University
United States
ttabassu@student.fitchburgstate.edu

Corey Wade
National University
United States
corey@wademl.dev

Ricky J. Sethi
Fitchburg State University
National University
United States
rickys@sethi.org

***Abstract*—In the legal field, corporate civil matters often entail tens or hundreds of thousands of documents that need review for relevance. Technology-Assisted Review (TAR) systems utilize machine learning classification algorithms, such as logistic regression, SVM, and transformers, to retrieve all, or nearly all, relevant documents from the corpus under review. However, in these e-discovery scenarios, TAR systems are typically perceived as "black boxes" by practitioners; where the TAR system provides little or no insight into why a document is predicted to be relevant. This lack of explainability makes it difficult for attorneys to trust classifications from TAR systems, hinders litigants from participating fully as they cannot understand why documents are being classified as relevant, and relies on the costly interpretation of experts rather than the model itself for understanding.**

In contrast to these opaque methods, the Fuzzy ARTMAP algorithm is an explainable neural network architecture that is both *geometrically interpretable* and allows for the *extraction of fuzzy If-Then rules* from the model at any point in its training. This enables a practitioner or researcher multiple modes with which to understand what the model has learned up to that point, laying the foundation for Explainable e-Discovery (XeD).

In this paper, the explainable Fuzzy ARTMAP neural network is extended to include fuzzy subsethood to rank documents for active learning and is then evaluated for use in the TAR domain with several corpora. In addition to achieving desirable performance for a TAR system, it also enables *direct insight into how the algorithm decides relevance*. This is in contrast to existing approaches for explainable TAR which rely on extracting document snippets as post hoc explanations of why a document is relevant. Additionally, we demonstrate the model's interpretability with both textual and graphical representations of the learned model for a range of representations including tf-idf, GloVe, and Word2Vec.

***Index Terms*—TAR, Legal document review, Explainable AI, e-discovery, Fuzzy ARTMAP**

I. Introduction

A Technology-Assisted Review (TAR) system that can explain how and why document relevance predictions are made is a vital tool for enabling attorneys to meet their ethical obligations to clients and enable clients to fully participate in the legal process [1]. Despite the potential benefits of an explainable TAR system, current systems fail to deliver on why documents are classified as responsive and so these systems are still typically perceived as "black boxes" by practitioners [2]–[4].

While a few studies have attempted to bring explainability to TAR systems, they focused on extracting snippets from the documents as the mechanism of explanation rather than directly explaining the relevance model [2]–[4]. Instead, we looked at the explainable Fuzzy ARTMAP neural network. The model learned by the Fuzzy ARTMAP algorithm can be directly interpreted geometrically [5], [6] or as a set of fuzzy If-Then rules [7]–[9], depending on the features used.

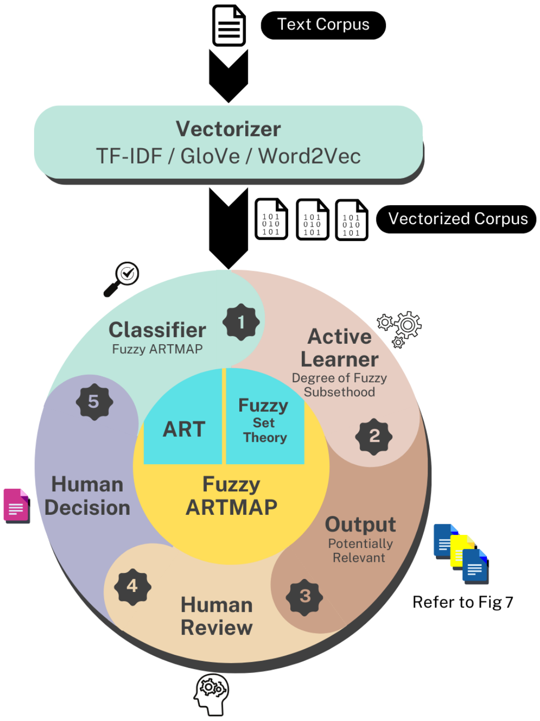


Fig. 1. Conceptual overview of TAR process with Fuzzy ARTMAP.

We developed a novel TAR system incorporating the Fuzzy ARTMAP neural network, as shown in the conceptual diagram in Fig. 1. We then performed an initial evaluation of the performance of the explainable Fuzzy ARTMAP algorithm in the TAR domain and found robust performance in terms of recall and precision [10].

Building on the strength of these initial results, we have now continued this foundational research by making the following contributions in this paper:

- modifying the Fuzzy ARTMAP algorithm to report the degree of fuzzy subsethood as a way to rank documents for active learning (see Fig. 1),
- performing a hyperparameter sweep [11] to refine the parameters and evaluating the system against the 20 Newsgroups, Reuters-21578, RCV1-v2, and Jeb Bush emails corpora for recall, precision, and F_1 (see Tables I and II),
- generating fuzzy If-Then rules to interpret the model (see Fig. 3),
- generating category descriptors to interpret the model (see Fig. 4), and
- generating graphical representations of the model (see Figs. 5 and 7)

While these corpora are not specific to the legal domain, the RCV1-v2 and Jeb Bush emails corpora are frequently used in e-discovery evaluations [12], [13] because legal matters are often confidential [2], [14] and their corpora are thus unavailable. The 20 Newsgroups corpus is commonly used as a test corpus with ART-based algorithms [6], [15]; 20 Newsgroups and the Reuters-21578 corpus are also commonly used in evaluating text classification algorithms [16].

The rest of this paper is organized as follows: Section II places our system in the context of related work. Then, in Section III, we give an overview of our approach. In Section IV, we give details of our experimental results and illustrate the explainability proof-of-concept. Finally, in Section V, we discuss the results and future work.

II. Related Work

A. Explainable TAR for e-Discovery

There are three relevant attempts at creating an explainable TAR for e-discovery system. The first attempt [2] evaluated two approaches to extract a snippet from a relevant document. Their first approach used the same document classification model to classify overlapping text snippets from the document and assign a probability of relevance. Their second approach used a rationale model, a secondary classification model based on annotated documents, to identify relevant snippets [2].

Building on the work of [2], the authors derived three metrics to determine a snippet's relevance [3]. These metrics included document-level relevance, a perturbation-based measure where the document is reclassified without the snippet, and a weighted average of the relevance assigned to the tokens in the snippet. These measures were combined in a weighted sum and a rank-based transformation of the scores. Each measure and combination of measures was evaluated and resulted in the weighted sum producing the highest snippet recall. However, this study required the availability of labelled snippets.

Addressing the typical lack of labelled snippet training data, [4] extended the previous work to identify snippets that predict document relevance without the benefit of labeled snippets. Two approaches were introduced, a snippet model which performs one pass of snippet selection through a document and an iterative snippet model which performs multiple iterations through a document reducing the snippet size by half each iteration. Snippet selection is the same for both models relying on an initial document-level scoring approach based on logistic-regression. The snippet model performs slightly better than the iterative approach and both perform better than the document-level approach.

These studies did not consider an active learning TAR system, instead using a fixed set of training documents. Removing the human-in-the-loop component, the core classifier model is not rebuilt based on the new judgements after each learning iteration, which would result in updating the snippet models in all of the previous approaches. This implies that the model and its explanations will not improve as a result of additional classification efforts. Selection of snippets as explanation are a *post hoc explanation* of the classifier model and relevance decision, which does *not* provide direct model interpretability.

B. Fuzzy ARTMAP

Adaptive Resonance Theory (ART) describes how the brain learns and predicts in a non-stationary world [17]. This theory models how brains can quickly learn new information without forgetting previously learned information. Various neural network algorithms have implemented ART across supervised, unsupervised, and reinforcement learning [18]. Fuzzy ART is a neural network algorithmic instantiation of ART which employs the fuzzy AND operator, from fuzzy set theory, instead of the binary set union operator, to work with real-valued features [5]. The Fuzzy ARTMAP algorithm is the supervised version of Fuzzy ART, *mapping* inputs to category labels. The integration of fuzzy set theory and ART dynamics in the Fuzzy ARTMAP neural network algorithm enable the model to be represented through different means. A model produced by Fuzzy ARTMAP may be represented as fuzzy If-Then text-based rules, as shown in Fig. 3, or depicted geometrically [5], [9], as shown in Fig. 5.

A requirement for geometric interpretation of the model is that the input must be complement encoded. Complement encoding is a normalization method in which the input vector $\boldsymbol{x}$ is concatenated with its complement $\overline{\boldsymbol{x}}$ (or $1 - \boldsymbol{x}$), yielding an input of $\boldsymbol{I} = [\boldsymbol{x}, \overline{\boldsymbol{x}}]$ [5]. As a result, the categories learned by the Fuzzy ARTMAP algorithm can be interpreted as n-dimensional hyper-rectangles [5], [6]. An example of this representation is shown in Fig. 5(b) where the category is displayed as a light gray rectangle.

III. Approach

A. Technology Assisted Review

We evaluated the Fuzzy ARTMAP neural network algorithm against the 20 Newsgroups, Reuters-21578, RCV1-v2, and Jeb Bush emails corpora. *Tf-idf* features were used with the smaller corpora and the 300-dimension versions of the *GloVe* and *Word2Vec* vectorizations were used with all of the corpora. All the topics in 20 Newsgroups, 120 topics in Reuters-21578, and 30 topics in both the RCV1-v2 and the Jeb Bush emails corpora were used for evaluation; the RCV1-v2 and the Jeb Bush corpora were down-sampled to 20% and 50% per [19] due to memory constraints, retaining the general prevalence per topic.

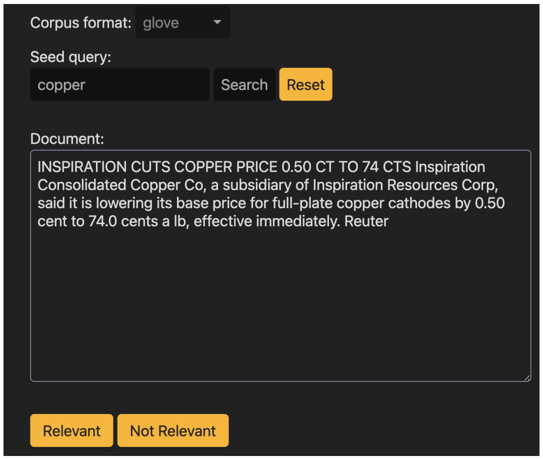
The tf-idf vectorization used was implemented in scikit-learn [20] with parameters based on [21] and resulted in 82,181-dimension vectors for 20 Newsgroups and 25,627-dimensions vectors for Reuters-21578. For GloVe [22], the 300-dimension vectors from the 6 billion token corpus were used. The gensim [23] implementation of Word2Vec [24] was used, based on the Google News 300-dimension vectors. For the GloVe and Word2Vec representations, the vectors for each word in the document were averaged to produce the overall document vector [25]. All document representations were scaled to the [0,1] interval using the scikit-learn MinMaxScaler, as this is the required feature range for the Fuzzy ARTMAP algorithm [5]. The features were complement encoded per [5] prior to processing via Fuzzy ARTMAP.

For our experiments, a continuous active learning approach was taken [14]. For each topic, a seed set of ten relevant documents and 90 non-relevant documents was used to initially train the Fuzzy ARTMAP algorithm, regardless of corpora size. Each review iteration consisted of up to 100 documents for the smaller corpora and up to 1,000 for the larger corpora, as only documents predicted relevant were returned in each review iteration.

As the Fuzzy ARTMAP classifier returns a selected class, not a set of real values indicating confidence among many classes, the Fuzzy ARTMAP algorithm was modified to report the degree of fuzzy subsethood [5], [26] associated with documents predicted as relevant. The *fuzzy set membership served as a proxy for confidence*, indicating how well the document matches the class. This *degree of fuzzy subsethood* was then used to **rank** the documents for active learning. The top ranked documents would be shown to a human-in-the-loop evaluator to determine if the documents are relevant or not relevant, as illustrated in Fig. 2. These human relevance judgements would then be used in an online learning mode to update the classifier model, rather than recreate the classifier model from scratch, as is the case in most TAR implementations [21]. This workflow is illustrated in Fig. 1. In our experiments the human-in-the-loop active learning evaluations are simulated using ground-truth labels instead of a human evaluator based on the framework in [21].

The review of documents for each topic concluded when the algorithm predicted no more relevant documents in the unevaluated portion of the corpus. This is in contrast to most other TAR approaches where the system leaves stopping an open question for the operator [21].

There are two significant parameters in the Fuzzy ARTMAP algorithm, the learning rate (β) and the baseline vigilance ($\overline{\rho_a}$). Based on the results of a sweep of the Fuzzy ARTMAP hyperparameters, which evaluated different combinations of baseline vigilance and learning rates, vigilance was set to .95, and a fast learning rate of 1.0 was selected.


Corpus format: glove

Seed query: copper

Document:

INSPIRATION CUTS COPPER PRICE 0.50 CT TO 74 CTS Inspiration Consolidated Copper Co, a subsidiary of Inspiration Resources Corp, said it is lowering its base price for full-plate copper cathodes by 0.50 cent to 74.0 cents a lb, effective immediately. Reuter

Fig. 2. Interface to perform initial search, view document, enter relevance.

B. Model Explanation

We constructed a proof-of-concept TAR implementation, as shown in Fig. 2. This implementation allows the user to select a representation for the corpus format, takes an initial keyword query to start the review, and then allows the user to input their relevance judgements for each document.

In this proof-of-concept, we implemented three explanatory approaches for the model learned by the Fuzzy ARTMAP neural network algorithm. A fuzzy If-Then rule interpretation for tf-idf, a textual description of the category for Word2Vec and GloVe representations, and a graphical representation for all three vectorizations.

For the tf-idf representation, fuzzy If-Then rules were generated based on a similar approach to [8]. The weight associated with each relevant antecedent feature, or word for tf-idf, is quantized into three levels of *rarely*, *somewhat*, and *highly* prevalent and is then output in a human readable text format as shown in Fig. 3. Because the focus is on understanding and not model consolidation, pruning operations from [8] were not performed.

As the features of GloVe and Word2Vec dimensions are not directly interpretable, the geometric interpretation is leveraged

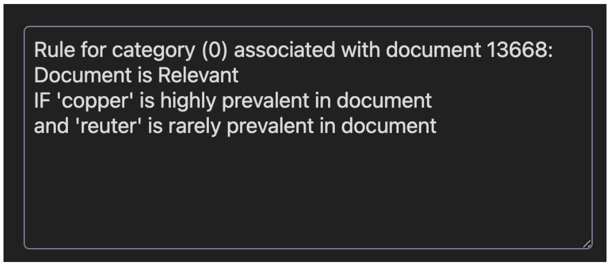
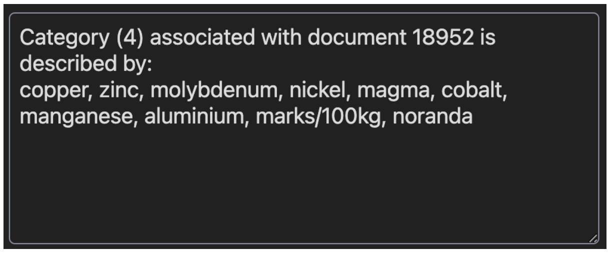

Rule for category (0) associated with document 13668:
Document is Relevant
IF 'copper' is highly prevalent in document
and 'reuter' is rarely prevalent in document

Fig. 3. If-then rules for tf-idf representation.



Category (4) associated with document 18952 is described by:
copper, zinc, molybdenum, nickel, magma, cobalt, manganese, aluminium, marks/100kg, noranda

Fig. 4. Top 10 closest descriptive words for GloVe-based category.

instead. A Fuzzy ARTMAP **category** is an n*-dimensional rectangle* in the feature space. To explain a GloVe or Word2Vec-based category **textually**, we find the center of the rectangle, then find the ten closest words to the center based on cosine-similarity resulting in the type of descriptor shown in Fig. 4.

Finally, for all representations, a fully graphical model was produced. Uniform Manifold Approximation and Projection (UMAP) [27] was used to reduce the high-dimensional spaces to a two-dimensional representation, as shown in Fig. 5. The corpus is reduced to a two-dimensional space using UMAP, and the UMAP model based on the corpus is used to project the category rectangle into the two-dimensional space. There, the first half of the category weights are used as one corner of the category rectangle and the other half of the weights are used to form the opposite corner, with the visualization of the category rectangle shown in Fig. 5(b).

IV. Results and Discussion

A. Performance

Considering all corpora and vectorizations, the Fuzzy ARTMAP-based system achieved 100% recall 31% of the time, and achieved the suggested floor of 75% [28] or better recall 67% of the time, as seen in Table I. Recall between the vectorizers for the Reuters-21578 and 20 Newsgroups corpora was different by a statistically significant degree based on a Friedman test [29] with $p<.001$ ($\chi^3(2)$=25.09 and $\chi^3(2)$=34.9). A post-hoc Nemenyi test [29] indicated a difference between tf-idf and both GloVe and Word2Vec, with the average difference and statistical significance shown in Table II. Based on the average difference, there is a practical significance to the tf-idf vectorization over GloVe and Word2Vec. No statistical or practical difference was present between GloVe and Word2Vec for the RCV1-v2 or Jeb Bush Emails corpora.

TABLE I
Median Metrics by Corpus-Vectorizer

Corpus	Vectorizer	Recall	Precision	F_1
20 Newsgroups	GloVe	0.57	0.522	0.434
	Word2Vec	**0.772**	0.41	0.523
	tf-idf	**0.94**	0.367	0.53
Jeb Bush Emails	GloVe	0.622	0.07	0.125
	Word2Vec	0.593	0.055	0.098
RCV1-v2	GloVe	**0.764**	0.211	0.324
	Word2Vec	**0.752**	0.187	0.292
Reuters-21578	GloVe	**0.909**	0.384	0.526
	Word2Vec	**0.931**	0.514	0.624
	tf-idf	**0.92**	0.733	0.759

TABLE II
Average Recall Difference

	Reuters-21578	20Newsgroups
tf-idf-GloVe	0.085^{**}	0.451^{**}
tf-idf-Word2Vec	0.069^{*}	0.171^{**}

*p <.05, **p <.01

These results indicate generally robust recall performance, particularly with the tf-idf vectorization. Except for the Jeb Bush Emails, and the GloVe vectorization of 20 Newsgroups, the median recall was 75% or better. The precision-recall curve for Fuzzy ARTMAP is highly variable, with precision not monotonically decreasing as recall increases. Rather, precision varies after each review iteration as the model learns in each iteration. In the more informal corpora of 20 Newsgroups and the Jeb Bush Emails, the GloVe and Word2Vec features did not perform as well. This may be due to the corpus specificity of tf-idf compared with the off-the-shelf vocabulary of GloVe and Word2Vec. The difference in performance suggests that generating corpus-specific GloVe and Word2Vec representations may perform better than the default vocabulary. Accordingly, exploring corpus-specific versions of Word2Vec and GloVe may bring recall in line with tf-idf, presenting a more efficient yet equally robust option.

B. Interpretability

While If-Then rules and graphical representations are acknowledged methods of explainability, there are no agreed-upon quantitative metrics for the explainable artificial intelligence space generally [30]; in addition, there are no qualitative or quantitative user studies of the existing prior attempts at explainability in e-discovery TAR [2]–[4]. This represents another likely productive area of future work. However, we have illustrated three potentially powerful approaches to representing the model learned by Fuzzy ARTMAP for TAR. For a textual representation, fuzzy If-Then rules can be generated for tf-idf features (Fig. 3); for complex categories, pruning [8] could produce more compact rules while retaining performance and interpretability. With representations like GloVe and Word2Vec, where the features are not directly interpretable, we describe the category based on the ten words closest to the center of the learned category (Fig. 4).

Finally, for all representations a rich graphical representation is possible. In the graphical representation, the document being currently evaluated is shown as a red circle (Fig. 5(a)), while triangle points in the visualization represent documents that have already been judged relevant or not relevant by a human (Fig. 5(e)). This enables exploration of nearby documents that might be relevant as shown in Fig. 6, where the tool tip indicates an unevaluated document (the circle) near a relevant document (the triangle). Additionally, the relevant category is shown as a light gray rectangle covering a range of documents

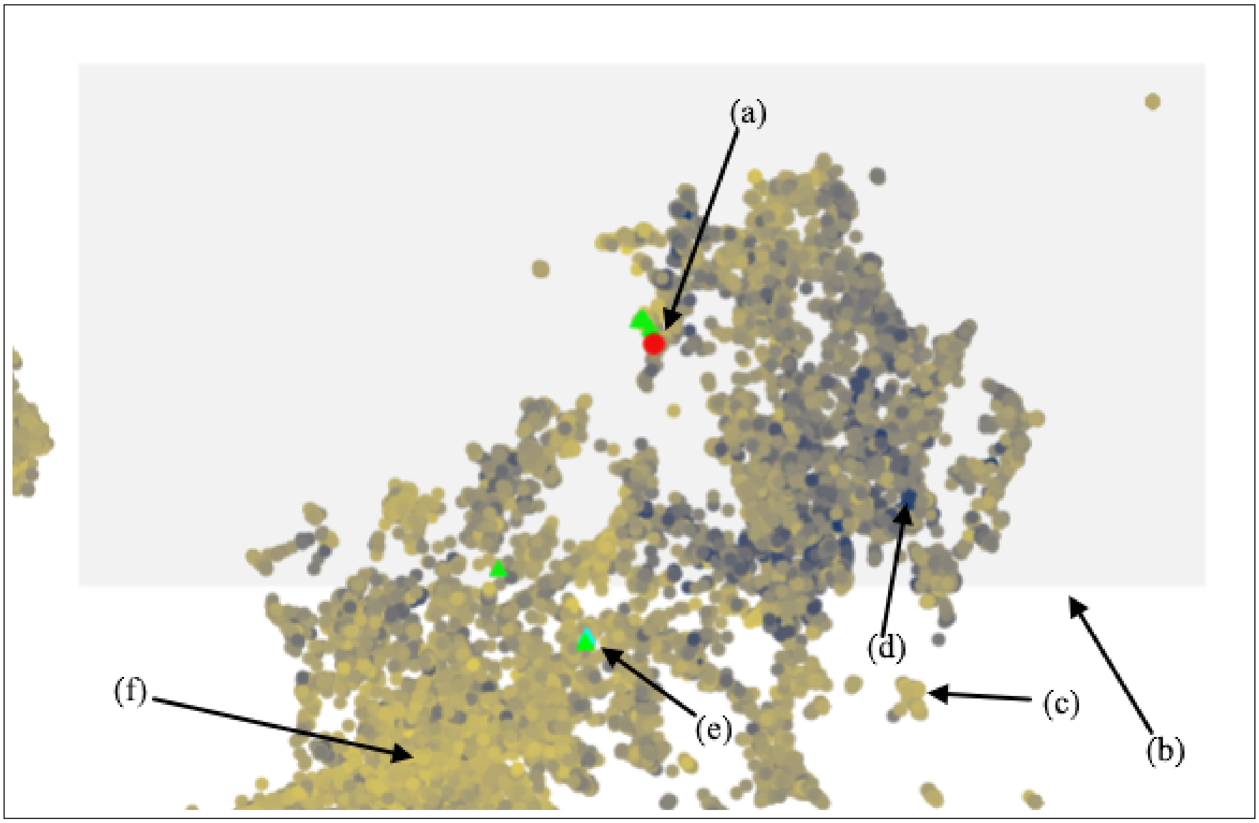
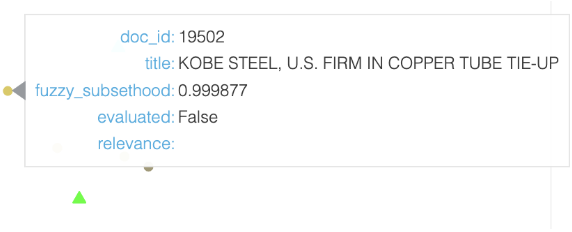


Fig. 5. Fuzzy ARTMAP Graphical Representation Explanation. (a) Current document under review. (b) The current relevant category the document is predicted to belong to. (c) A document predicted more likely to be relevant. (d) A document less likely predicted to be relevant. (e) A document whose relevance has already been evaluated. (f) All points (circles and triangles) represent documents in the corpus under review. Please see Fig. 7 for a cartoon illustration of the graphical representation.

indicating potentially relevant documents (Fig. 5(b)). The degree of relevance is shown as a color gradient from blue, for less relevant documents (Fig. 5(d)), to yellow, for more relevant documents (Fig. 5(c)), offering a perspective of where the more relevant documents might be in the corpus. Further, the display is interactive as Bokeh is used as the visualization framework allowing panning, zooming, saving, and hovering over the points to get more information about the represented documents (Fig. 6). Utilizing the Bokeh framework and the graphical representation, selecting a document other than the next highest predicted document to evaluate is possible and would support a hybrid explore/exploit interaction with the documents under review as opposed to typical TAR systems which just present the next document for review.



doc_id: 19502
title: KOBE STEEL, U.S. FIRM IN COPPER TUBE TIE-UP
fuzzy_subsethood: 0.999877
evaluated: False
relevance:

Fig. 6. A potentially relevant neighboring document.

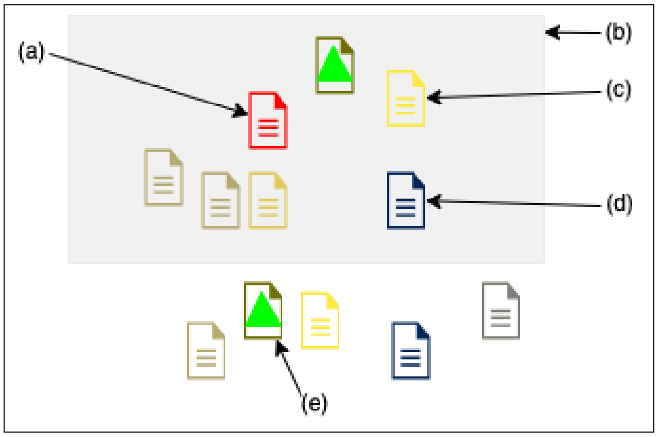


Fig. 7. A simplified cartoon explanation of the graphical representation shown in Fig. 5. The category descriptor is represented by the grey rectangle. The yellow-coloured documents are predicted to be more relevant and documents with a triangle are ones whose relevance has already been determined, either as relevant or not relevant. Relevant documents can be internal or external to the rectangle, as the rectangle only shows the category for the current document under review. Relevance of the documents is not a function of their location within the displayed rectangle, and may be part of another category (rectangle). (a) Current document under review, indicated by the red color. (b) The rectangle representing the category to which the current document is predicted to belong. (c) A document predicted more likely to be relevant, indicated by the yellow color. (d) A document predicted less likely to be relevant, indicated by the blue color. (e) A document whose relevance has already been evaluated, indicated by the enclosed green triangle.

V. Conclusions and Future Work

This foundational research provides additional substantiation for using the Fuzzy ARTMAP neural network as a classification algorithm in the TAR domain. The Fuzzy ARTMAP neural network demonstrates generally robust recall performance with a variety of representations and corpora. Additionally, multiple interpretations of the model, from fuzzy If-Then rules for tf-idf representations, categorical descriptions for GloVe and Word2Vec representations, and a graphical interpretation for all representations, illustrate a variety of viable alternatives to ”black box” TAR systems. Furthermore, as Fuzzy ARTMAP is not a pre-trained model, there are no inherent biases or ethical concerns around training or data disclosure as long as unique models are used per legal matter.

Research opportunities exist in improving recall performance through the use of corpus-specific vectorizations of GloVe and Word2Vec. Furthermore, the representation of the model in two-dimensional space might further be optimized through modification of the UMAP parameters, or exploring other dimensionality reduction techniques like principal component analysis. Evaluating other distance metrics to spatially collocate relevant and potentially relevant documents in the graphical representation is another potential improvement to the visual explanation. Finally, user studies are required to understand the viability and applicability of the model explanations to help reach the idea of ”procedural justice” in e-Discovery [1].

References

[1] S. K. Endo, “Technological opacity & procedural injustice,” *Boston College Law Review*, vol. 59, no. 3, pp. 822–875, Mar. 2018.

[2] R. Chhatwal, P. Gronvall, N. Huber-Fliflet, R. Keeling, J. Zhang, and H. Zhao, “Explainable text classification in legal document review a case study of explainable predictive coding,” in *2018 IEEE International Conference on Big Data (Big Data)*. Seattle, WA, USA: IEEE, Dec. 2018, pp. 1905–1911.

[3] C. J. Mahoney, J. Zhang, N. Huber-Fliflet, P. Gronvall, and H. Zhao, “A framework for explainable text classification in legal document review,” in *2019 IEEE International Conference on Big Data (Big Data)*. Los Angeles, CA, USA: IEEE, Dec. 2019, pp. 1858–1867.

[4] C. Mahoney, P. Gronvall, N. Huber-Fliflet, and J. Zhang, “Explainable Text Classification Techniques in Legal Document Review: Locating Rationales without Using Human Annotated Training Text Snippets,” in *2022 IEEE International Conference on Big Data (Big Data)*. Osaka, Japan: IEEE, Dec. 2022, pp. 2044–2051.

[5] G. A. Carpenter, S. Grossberg, N. Markuzon, J. H. Reynolds, and D. B. Rosen, “Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps,” *IEEE Transactions on Neural Networks*, vol. 3, no. 5, pp. 698–713, Sept./1992.

[6] L. Meng, A.-H. Tan, and D. C. Wunsch II, *Adaptive Resonance Theory (ART) for Social Media Analytics*. Cham: Springer International Publishing, 2019, pp. 45–89.

[7] G. A. Carpenter and A.-H. Tan, “Rule extraction, Fuzzy ARTMAP, and medical databases,” in *Proceedings of the World Congress on Neural Networks*. Portland, OR, USA: Erlbaum Associates, Jan. 1993, pp. 501–506.

[8] ——, “Rule extraction: From neural architecture to symbolic representation,” *Connection Science*, vol. 7, no. 1, pp. 3–27, Jan. 1995.

[9] S. Grossberg, “A Path Toward Explainable AI and Autonomous Adaptive Intelligence: Deep Learning, Adaptive Resonance, and Models of Perception, Emotion, and Action,” *Frontiers in Neurorobotics*, vol. 14, p. 36, Jun. 2020.

[10] C. Courchaine and R. Sethi, J., “Fuzzy Law: Towards Creating a Novel Explainable Technology-Assisted Review System for e-Discovery,” in *2022 IEEE International Conference on Big Data (Big Data)*. Osaka, Japan: IEEE, Dec. 2022, pp. 1218–1223.

[11] R. J. Sethi, *Essential Computational Thinking: Computer Science from Scratch*. Cognella Academic Publishing, 2020.

[12] E. Yang, D. D. Lewis, and O. Frieder, “A regularization approach to combining keywords and training data in technology-assisted review,” in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*. Montreal QC Canada: ACM, Jun. 2019, pp. 153–162.

[13] ——, “On minimizing cost in legal document review workflows,” *arXiv:2106.09866 [cs]*, Jun. 2021.

[14] G. V. Cormack and M. R. Grossman, “Autonomy and reliability of continuous active learning for technology-assisted review,” *arXiv:1504.06868 [cs]*, Apr. 2015.

[15] D. Marček and M. Rojček, “The category proliferation problem in ART neural networks,” *Acta Polytechnica Hungarica*, vol. 14, no. 5, p. 15, 2017.

[16] B. Altınel and M. C. Ganiz, “Semantic text classification: A survey of past and recent advances,” *Information Processing & Management*, vol. 54, no. 6, pp. 1129–1153, Nov. 2018.

[17] S. Grossberg, “Toward Autonomous Adaptive Intelligence: Building Upon Neural Models of How Brains Make Minds,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 1, pp. 51–75, Jan. 2021.

[18] L. E. Brito da Silva, I. Elnabarawy, and D. C. Wunsch, “A survey of adaptive resonance theory neural network models for engineering applications,” *Neural Networks*, vol. 120, pp. 167–203, Dec. 2019.

[19] E. Yang, S. MacAvaney, D. D. Lewis, and O. Frieder, “Goldilocks: Just-right tuning of BERT for technology-assisted review,” *arXiv:2105.01044 [cs]*, May 2021.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, “Scikit-learn: Machine learning in python,” *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[21] D. Li and E. Kanoulas, “When to stop reviewing in technology-assisted reviews: Sampling from an adaptive distribution to estimate residual relevant documents,” *ACM Transactions on Information Systems*, vol. 38, no. 4, pp. 1–36, Oct. 2020.

[22] J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1532–1543.

[23] R. Řehůřek and P. Sojka, “Software framework for topic modelling with large corpora,” in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50.

[24] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in Neural Information Processing Systems*, vol. 26. Lake Tahoe, Nevada: Curran Associates, Inc., 2013, pp. 3111–31 119.

[25] A. Carvallo, D. Parra, H. Lobel, and A. Soto, “Automatic document screening of medical literature using word and text embeddings in an active learning setting,” *Scientometrics*, vol. 125, no. 3, pp. 3047–3084, Dec. 2020.

[26] B. Kosko, “Fuzzy entropy and conditioning,” *Information Sciences*, vol. 40, no. 2, pp. 165–174, Dec. 1986.

[27] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” Sep. 2020.

[28] R. Keeling, R. Chhatwal, P. Gronvall, and N. Huber-Fliflet, “Humans Against the Machines: Reaffirming the Superiority of Human Attorneys in Legal Document Review and Examining the Limitations of Algorithmic Approaches to Discovery,” *Richmond Journal of Law & Technology*, vol. 26, no. 3, p. 65, 2020.

[29] J. Demsar, “Statistical Comparisons of Classifiers over Multiple Data Sets,” *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, Dec. 2006.

[30] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Information Fusion*, vol. 58, pp. 82–115, Jun. 2020.