Fuzzy Law: Towards Creating a Novel Explainable Technology-Assisted Review System for e-Discovery

1st Charles Courchaine
National University
United States
charles@courchaine.dev

2nd Ricky J. Sethi
Fitchburg State University
National University
United States
rickys@sethi.org

***Abstract*—In the legal field, Technology-Assisted Review (TAR) systems for e-discovery are typically perceived as "black boxes" by practitioners, providing little to no insight into how the system makes its classification predictions. The lack of explainability in TAR systems for e-discovery renders their decisions opaque, making it difficult for attorneys to trust their recommendations and thus to discharge ethical obligations to clients. In addition, litigants cannot fully participate in the process if they cannot understand the relevance judgments, and jurists cannot make well-informed judgments on discovery matters. The Fuzzy ARTMAP algorithm is an explainable neural network architecture that permits the extraction of fuzzy If-Then rules from the model at any point in its training, the model is also geometrically interpretable, allowing a researcher or practitioner to understand what the model has learned up to that point.**

This paper evaluates the explainable Fuzzy ARTMAP algorithm for use in the TAR domain. Not only does it achieve suitable document classification performance for a TAR system, as measured by recall and recall-at-effort, but it also enables *direct insight into how the algorithm decides relevance*. This is in contrast to existing approaches for explainable TAR which only rely on extracting document snippets as post hoc explanations of why a document is relevant.

In addition, the effect of different document features (tf-idf, word2vec, and GloVe) on recall performance is also evaluated. Performance is compared to AutoTAR, the state-of-the-art TAR algorithm which makes relevance predictions but is not able to provide any explanations about them. Experiments on the Reuters-21578 and 20Newsgroups corpora indicate robust recall performance overall and comparable or better metrics than AutoTAR in some circumstances.

***Index Terms*—TAR, AutoTAR, Legal, e-discovery, Fuzzy ARTMAP**

I. Introduction

Electronic discovery (e-discovery) is characterized as a high-recall retrieval (HRR) task, consisting of finding most or all of the documents relevant to a civil, criminal, or regulatory matter [1]. These matters may involve a substantial number of documents collected for review to determine their relevancy to the matter. HRR tasks thus focus on retrieving the majority of relevant information, and not just the few most relevant pieces; as such, in order to reduce the human effort required to review the documents in e-discovery, an information retrieval system is often utilized to help the human reviewers find and classify the documents; this process is known as technology-assisted review (TAR) [2]. There are, however, two significant distinctions between e-discovery and traditional information retrieval problem domains. First, the corpus in an e-discovery problem may be significant in size but it is finite and not intended to cover all possible information [3]. Second, a related distinction is that the classifier in e-discovery does not need to generalize to other corpora or queries [3], [4]. These differences require design choices that may not be appropriate for general information retrieval systems. In addition to the technical differences of e-discovery, the legal context creates additional impetus for explainability. Beyond the nominal benefits of explainable AI (XAI) facilitating transparency, explainability, and trustworthiness, the legal context of e-discovery adds additional desirability for the explainability of TAR systems. Understanding how and why TAR systems make predictions on the relevancy of documents is an essential enabler for attorneys to discharge their ethical obligations to clients and enable clients to participate in the judicial process fully [5]. Despite the benefits of an explainable TAR system, current systems fail to deliver on why documents are classified as responsive and are still typically perceived as "black boxes" by practitioners [6], [7].

To help address explainability in e-discovery TAR systems, we evaluate the performance of the explainable Fuzzy ARTMAP algorithm in the TAR domain. While other researchers have used or proposed ART algorithms for the unsupervised task of document clustering [8]–[11], little has been explored in the supervised classification task. Particularly within TAR, the Fuzzy ARTMAP algorithm does not appear to have been previously employed to the best of our knowledge and is a novelty of the present study. This study contributes to addressing this gap by examining the performance of Fuzzy ARTMAP for document classification and how the document representations impact recall performance.

The rest of this paper is organized as follows: Section II places our system in the context of related work. Then, in

978-1-6654-8045-1/22/$31.00 ©2022 IEEE

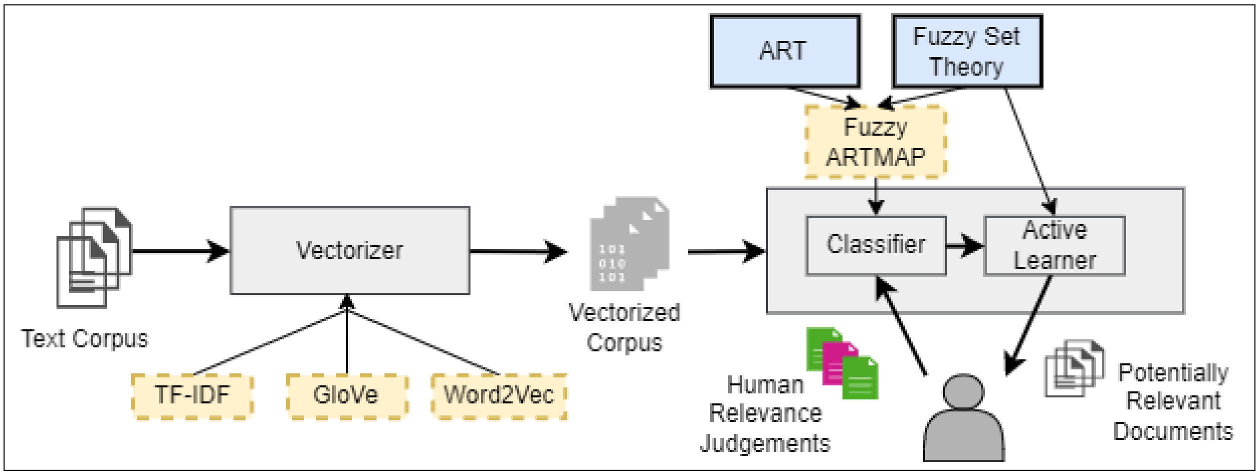

Fig. 1. Conceptual diagram of the Technology-Assisted Review system incorporating Fuzzy ARTMAP.

Section III, we give an overview of our approach. In Section IV, we give details of our experimental results. Finally, in Section V, we discuss the results and future work.

II. Related Work

Adaptive Resonance Theory (ART) describes how the brain learns and predicts in a non-stationary world [12]. This theory models how brains can quickly learn new information without forgetting previously learned information. The central motivating question in ART is how can an error be corrected with only local information, where no individual cell is aware of the error [12]. ART circuits emerge from answering this question in a methodical, rigorous manner, utilizing only local operations for feedback and minimax learning rules. ART has been instantiated in numerous models and algorithms [13]. The typical naming convention for these models is that the unsupervised versions have ART in the name, whereas supervised versions are denoted with ARTMAP.

The first implementation of ART as a neural network model was ART1. The ART1 implementation utilized binary crisp (classical) set operators, whereas the Fuzzy ART algorithm introduces fuzzy set theory operators; specifically, the fuzzy AND operator, to work with real-valued features [14]. The supervised version of the Fuzzy ART algorithm is the Fuzzy ARTMAP algorithm that enables a mapping between inputs and categories. By integrating fuzzy set theory and ART dynamics in the Fuzzy ARTMAP algorithm, various explainable features are yielded. What the model learns may be represented geometrically and in terms of fuzzy If-Then rules [14], [15].

In NLP, ART-based neural network algorithms have primarily been used for document clustering. ART1 was tested for clustering performance with the Reuters-21578 corpus and binary (one-hot) representation, with moderate success, exceeding the K-means performance lower bound and half the performance of the k-Nearest Neighbors upper bound [10]. Variants of the Fuzzy ART algorithm have been used with the 20Newsgroups corpus and tf-idf document representation [8], [11]. These variants matched or exceeded the performance metrics of the baseline comparisons, such as DBSCAN and Affinity Propagation. A hybrid of clustering and classification ART algorithms based on Fuzzy ART was proposed in [9]. The 20Newsgroups corpus and tf-idf representation were used to test the hybrid algorithm, without a baseline comparison, but reported F-1 measures of greater than 0.75. This study evaluates Fuzzy ARTMAP with a variety of document representations, in addition to tf-idf, following [4], and applies it in a transductive setting for use with specific corpora as opposed to a general document classifier.

A. Fuzzy ARTMAP Interpretability

One of the primary methods of interpreting the model learned by the Fuzzy ARTMAP algorithm is through the use of If-Then rules [16], [17]. In [17] three different databases, the Pima Indian diabetes diagnosis, mushroom classification, and DNA promoter recognition are evaluated for rule extraction from Fuzzy ARTMAP. The Pima Indian database features are represented as real valued vectors, whereas the mushroom classification and DNA promoter databases are represented using binary vectors. Real-valued features, such as *age* and *diastolic blood pressure*, and binary features, such as *gill size is broad* and *has bruises*, can be interpreted directly; whereas the binary features of a DNA sequence require additional interpretation. However, even with additional feature interpretation, If-Then rules can be extracted from the Fuzzy ARTMAP model [17].

When using complement encoding, where the input vector $\mathbf{x}$ is concatenated with its complement $\overline{\boldsymbol{x}}$ yielding an input of $\boldsymbol{I} = [\boldsymbol{x}, \overline{\boldsymbol{x}}]$, the categories learned by the Fuzzy ARTMAP algorithm can be interpreted as n-dimensional hyper-rectangles [11], [14]. This geometric interpretation is the other primary method of interpreting the model learned by the Fuzzy ARTMAP algorithm [11], [14]. In this interpretation, the weights learned from the non-complement encoded portion of the input vector form one corner of the hyper-rectangle and the weights learned from the complement encoded portion of

TABLE I
Reuters-21578: AutoTAR Performance

Representation-Topic	Recall	Precision	F-1	Last Rel
tf-idf-earn	1.000	0.55	0.709	6863
tf-idf-money-fx	1.000	0.094	0.172	7266
tf-idf-crude	1.000	0.038	0.074	14555
glove-earn	1.000	0.238	0.385	15820
glove-money-fx	1.000	0.082	0.152	8316
glove-crude	1.000	0.049	0.094	11399
word2vec-earn	1.000	0.32	0.485	11778
word2vec-money-fx	1.000	0.048	0.092	14121
word2vec-crude	1.000	0.084	0.155	6718

TABLE II
Reuters-21578: Fuzzy ARTMAP Performance

Representation-Topic	Recall	Precision	F-1	Last Rel
tf-idf-earn	0.886	**0.593**	**0.711**	**5633**
tf-idf-money-fx	0.887	**0.144**	**0.248**	**4202**
tf-idf-crude	0.878	**0.124**	**0.217**	**3995**
glove-earn	0.914	**0.682**	**0.781**	**5057**
glove-money-fx	0.817	**0.277**	**0.413**	**2017**
glove-crude	0.887	**0.286**	**0.433**	**1750**
word2vec-earn	0.942	**0.662**	**0.778**	**5369**
word2vec-money-fx	0.844	**0.293**	**0.435**	**1967**
word2vec-crude	0.885	**0.28**	**0.426**	**1786**

the input vector form the other corner. Data within the hyper-rectangle are predicted to belong to the category associated with that region.

B. Explainable TAR for e-Discovery

There are two notable previous attempts at creating an explainable TAR for e-discovery system. The first attempt by [6] 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. The second approach used a rationale model, a secondary classification model based on annotated documents, to identify relevant snippets. These approaches used a logistic regression classifier and tf-idf over a private corpus of 688,294 documents [6]. The second attempt by [7], a similar group of researchers, builds on the work of [6]. Three metrics were calculated to determine the snippet's relevance: using the document model to predict relevance of the snippet, a perturbation-based measure where the snippet is removed and the document re-classified, and a weighted average of the relevance of the tokens in the snippet to the classifier. These three measures were combined in two ways: a weighted sum of the scores and a weighted sum of rank-based transformation of the scores. Each of the individual measures and combination of measures were evaluated, with the weighted sum of the scores generating the highest snippet recall. These studies did not consider an active learning TAR system, which would have mutated the document classifier; since they do not have that human-in-the-loop component, their classifier model is not rebuilt based on the new judgements after each learning iteration. That pattern would require retraining for any snippet specific models as well. Additionally, this is a post hoc ”explanation” of the classifier's decision, not a direct insight into why a document was classified as responsive.

III. Approach

The explainable TAR approaches described in Section II-B do not make any comparisons to AutoTAR since they focus exclusively on addressing explainability and not overall TAR results; as such, they are only broadly described, used a private corpus, and did not provide any metrics for comparison to state-of-the-art. Our proposed system not only achieves resulting metrics on a par with the state-of-the-art but it also provides direct insight into how the algorithm decides relevance for explainability.

Our high-level approach is outlined in Figure 1. Each corpus, 20Newgroups and Reuters-21578, was vectorized using each of the three vectorizers: *tf-idf* as implemented in scikit-learn and parameterized for AutoTAR reproduction based on [18] resulting in 82,181-dimensions for 20Newsgroups and 25,627-dimensions for Reuters-21578, *GloVe* [19] with 300-dimensions from the 6 billion token corpus, and *Word2Vec* [20] as implemented in gensim using the Google News 300-dimension vectors. For GloVe and Word2Vec representations, the vectors for each word in the document are averaged [4]. All document representations are scaled to the [0,1] interval using the scikit-learn MinMaxScaler, as this is the required feature range for the Fuzzy ARTMAP algorithm. The features are complement encoded per [14] prior to processing via Fuzzy ARTMAP.

These vectorizers were selected for both their common usage in NLP applications and their interpritability as features. As *tf-idf* vectorization represents the document as a weighted bag of words, the features can be interpreted as the degree of prevalence of a particular term in a document. The degree of prevalence or absence can be quantized to a textual description, such as highly or nominally prevalent or absent, as the predicate of an If-Then style rule, and the relevance or non-relevance of the document as the consequent portion. This representation is compatible with the If-Then style of read-out from the model learned by the Fuzzy ARTMAP algorithm. A proof-of-concept interpretation was generated using the labels from the *tf-idf* vectorizer to retrieve the meaning of the features from the model, and a three-tier quantization to label the level prevalence as *rarely*, *somewhat*, and *highly* prevalent; an excerpt of a rule generated from the proof-of-concept is shown in Table V. Documents vectorized using *GloVe* and *Word2Vec* can be interpreted geometrically as a point in a 300-dimension space. Complement encoded categories in a Fuzzy ARTMAP model can also be interpreted geometrically as producing hyper-rectangles of n dimensions [11], [14]. Accordingly, the document can be interpreted as a point within the relevant or non-relevant category hyper-rectangle where nearby words within the category could be used to provide an interpretation of the document, and the category overall.

Three topics from each corpus were selected. From the Reuters-21578 the topics were selected for a range of prevalence within the corpus, while maintaining a reasonable number of relevant documents at the low-end: earn (19.83%), money-fx (3.59%), crude (2.97%). The three topics from the 20Newsgroups corpus were selected for a range of coverage of language to ensure the words about each topic are different: sci.med (med), comp.sys.ibm.pc.hardware (pc-hardware), and misc.forsale (forsale).

The Fuzzy ARTMAP classifier is seeded using a training set of 10 relevant documents and 90 non-relevant documents. The classifier then assigns a label of relevant and non-relevant to each of the remaining documents in the corpus, along with a measure of the fuzzy set membership each document has to the class. As the Fuzzy ARTMAP classifier returns a selected label, not a set of real values indicating confidence among many classes, the fuzzy set membership serves as a proxy for confidence, indicating how well the document matches the class. This fuzzy set membership value is used to rank the documents labeled relevant. The 100 highest ranked documents are evaluated for the active learning component in which the documents would be shown to a human-in-the-loop evaluator to determine if the documents are, in fact, relevant or not relevant. These human relevance judgements are then used in an online learning mode to update the classifier model, rather than recreate the classifier model from scratch as is the case most TAR implementations [18]. In our experiments the human-in-the-loop active learning evaluations are simulated using ground-truth labels instead of a human evaluator. All relevance judgements in a batch are used. Batches are typically 100 documents, but they dwindle in size as the model returns fewer predicted relevant documents. Model training is only done with the most recent batch of judgements and updated in-place, not recreated.

The corpus, excluding previously judged documents, is evaluated again by the classifier and the process repeats iteratively. An interesting property of the Fuzzy ARTMAP classifier in this setting emerges here: at some point in the process, the classifier predicts that there are no more relevant documents in the corpus. This behavior is in contrast to AutoTAR and its logistic regression classifier which continues to rank documents regardless of confidence in relevance predictions.

Our implementation of the Fuzzy ARTMAP algorithm further modifies the approach of [14] by using a common engineering modification whereby the classification vector is used instead of a second Fuzzy ART (ART$_b$) module, due to the ART$_b$ module's vigilance often being set to 1 resulting in an equal number of ART categories as classes/input labels [13]. There are two significant parameters in the Fuzzy ARTMAP algorithm, the learning rate (β) and the baseline vigilance ($\overline{\rho_a}$). For the learning rate, following [14], the fast-commit slow-recode option was implemented, wherein a learning rate of 1 is used for initial learning and 0.75 for updates. Baseline vigilance was set at 0.95 to strike a balance between number of clusters learned and accuracy [14]. Finally, the degree of fuzzy set membership was utilized as a proxy for relevance ranking since the Fuzzy ARTMAP algorithm returns labels, not probabilities of classification.

IV. Results

Tables I – IV report four metrics for each algorithm-corpus pairing: recall, precision, F-1, and Last Rel (the depth at which the last relevant document was found). These metrics are listed for each representation-topic pairing.

Although AutoTAR achieves 100% recall across the board, this is not necessarily sufficient or desirable as one could achieve 100% recall by returning all documents resulting in 0% precision and no effort savings; this is counter to a critical aspect of TAR, which is the reduction of human review effort. The Fuzzy ARTMAP algorithm overcomes this limitation and outperforms the AutoTAR algorithm on all other metrics, often by an order of magnitude.

The Fuzzy ARTMAP algorithm also attained recall greater than 80% on the entirety of the Reuters-21578 corpus and approximately 70% or greater on 80% of the 20Newsgroup corpus. Although it did not attain 100% recall, the trade-off between recall and precision makes it more reliable than AutoTAR, which sacrifices precision for total recall. In fact, a 70% recall floor for e-discovery is suggested by [21], and, in experimental settings, state-of-the-art TAR algorithms can consistently achieve over 80% recall [22], [23].

This 70% floor threshold is shown as a dashed line in Figure 2, which also shows an evaluation of recall-at-effort, or recall as a function of the number of documents retrieved, for both corpora. As can be seen there, the Fuzzy ARTMAP algorithm eventually reaches the maximum and stops predicting the presence of any more relevant documents in the corpus; as such, the lines in Figure 2 are abbreviated, either because 100% recall was achieved or the model predicted no more relevant documents remaining. Certain representation-topic pairings also begin returning results at much lower effort than AutoTAR; e.g., word2vec-money-fx, which exceeds 50% recall around 1,000 documents with Fuzzy ARTMAP instead requires around 1,500 documents effort with AutoTAR.

Looking beyond recall to precision and F-1 scores, Fuzzy ARTMAP outperforms AutoTAR on both corpora. In addition, although Fuzzy ARTMAP in general achieves somewhat less recall than AutoTAR, it inevitably finds the last reported relevant document much more quickly, indicating less review effort for the attained level of recall. A paired samples t-test was performed to compare the precision of AutoTAR to the Fuzzy ARTMAP implementation and there was a significant

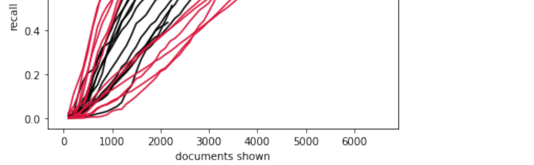

Fig. 2. Fuzzy ARTMAP recall-at-effort for both corpora.

TABLE III
20Newsgroup: AutoTAR Performance

Representation-Topic	Recall	Precision	F-1	Last Rel
tf-idf-pc-hardware	1.000	0.172	0.294	5791
tf-idf-med	1.000	0.068	0.127	14659
tf-idf-forsale	1.000	0.224	0.367	4447
glove-pc-hardware	1.000	0.067	0.127	14733
glove-med	1.000	0.061	0.115	16277
glove-forsale	1.000	0.066	0.124	15012
word2vec-pc-hardware	1.000	0.091	0.167	10958
word2vec-med	1.000	0.068	0.127	14631
word2vec-forsale	1.000	0.091	0.167	10908

TABLE V
Excerpt of rule output from Fuzzy ARTMAP for pc.hardware

Document is Relevant	
IF	*advance* is **rarely** prevalent in document
and	*apr* is **rarely** prevalent in document
and	*bogus* is **rarely** prevalent in document
and	*browning* is **highly** prevalent in document
and	*calstate* is **rarely** prevalent in document
and	*drive* is **somewhat** prevalent in document
and	*message* is **rarely** prevalent in document
and	*mfm* is **rarely** prevalent in document
and	*mitsubishi* is **somewhat** prevalent in document
and	*newsgroups* is **rarely** prevalent in document
and	*nscf* is **highly** prevalent in document
and	*number* is **rarely** prevalent in document
and	*sw1* is **rarely** prevalent in document
and	*sw2* is **somewhat** prevalent in document
...	

difference; t(34) = 3.76, p = .001, indicating an improvement in the precision of the Fuzzy ARTMAP implementation over AutoTAR. We show a proof-of-concept excerpt from one of the rules learned via Fuzzy ARTMAP for the pc.hardware newsgroup in Table V, which would be beyond the ability of AutoTAR and is more insightful than the other methods for explainability discussed in Section II-B.

V. Conclusion

Even with an untuned implementation of Fuzzy ARTMAP, performance on recall and recall-at-effort (Tables I – IV) is promising, with both meeting the 70% threshold for the entire Reuters-21578 corpus and for approximately 80% of the 20Newsgroup corpus. In addition, when compared with state-of-the-art, the Fuzzy ARTMAP implementation far exceeds in the precision, F-1, and Last Rel metrics, outperforming in all permutations. While the learning rate and vigilance parameters in this study were held constant, performing a parameter sweep of these values represents an area of opportunity for further improving the results. Finally, since existing explainable TAR methods build upon systems like AutoTAR, they only extract document snippets as post hoc explanations; as such, for comparison, we demonstrate the rules generated by Fuzzy ARTMAP in Table V, which are much more informative for attorneys and litigants. Expanding on the proof-of-concept If-Then rules and implementing a geometric interpretation of the *GloVe* and *Word2Vec* is another area of opportunity for future work.

TABLE IV
20Newsgroup: Fuzzy ARTMAP Performance

Representation-Topic	Recall	Precision	F-1	Last Rel
tf-idf-pc-hardware	0.872	**0.236**	**0.372**	**3683**
tf-idf-med	0.913	**0.287**	**0.436**	**3180**
tf-idf-forsale	0.907	**0.259**	**0.403**	**3492**
glove-pc-hardware	0.436	**0.203**	**0.277**	**2145**
glove-med	0.776	**0.32**	**0.453**	**2422**
glove-forsale	0.513	**0.229**	**0.317**	**2233**
word2vec-pc-hardware	0.679	**0.255**	**0.37**	**2662**
word2vec-med	0.846	**0.438**	**0.577**	**1928**
word2vec-forsale	0.694	**0.329**	**0.447**	**2104**

Acknowledgment

We would like to acknowledge Diyi Yang for the productive exchanges and guidance to some relevant references which were very helpful in our research.

References

[1] 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.

[2] G. McDonald, C. Macdonald, and I. Ounis, “Active learning strategies for technology assisted sensitivity review,” in *Advances in Information Retrieval*, G. Pasi, B. Piwowarski, L. Azzopardi, and A. Hanbury, Eds. Cham: Springer, 2018, vol. 10772, pp. 439–453.

[3] G. V. Cormack and M. R. Grossman, “Scalability of continuous active learning for reliable high-recall text classification,” in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. Indianapolis Indiana USA: ACM, Oct. 2016, pp. 1039–1048.

[4] 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.

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

[6] 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.

[7] 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.

[8] R. Forgac and R. Krakovsky, “Text processing by using projective ART neural networks,” in *2016 New Trends in Signal Processing (NTSP)*. Demanovska dolina, Slovakia: IEEE, Oct. 2016, pp. 1–5.

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

[10] L. Massey, “On the quality of ART1 text clustering,” *Neural Networks*, vol. 16, no. 5-6, pp. 771–778, Jun. 2003.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] G. A. Carpenter and A.-H. Tan, “Rule extraction, Fuzzy ARTMAP, and medical databases,” *Proceedings of the world congress on neural networks*, pp. 501–506, Jan. 1993.

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

[18] 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.

[19] 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.

[20] 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.

[21] M. R. Grossman and G. V. Cormack, “Technology-assisted review in electronic discovery,” in *Data-Driven Law*, E. Walters, Ed. New York City, NY, USA: Taylor & Francis Group, 2018.

[22] E. Yang, D. Grossman, O. Frieder, and R. Yurchak, “Effectiveness results for popular e-discovery algorithms,” in *Proceedings of the 16th Edition of the International Conference on Artificial Intelligence and Law*. London United Kingdom: ACM, Jun. 2017, pp. 261–264.

[23] J. Zou and E. Kanoulas, “Towards question-based high-recall information retrieval: Locating the last few relevant documents for technology-assisted reviews,” *ACM Transactions on Information Systems*, vol. 38, no. 3, pp. 1–35, Jun. 2020.