

Opening the TAR Black Box: Developing an Interpretable System for eDiscovery Using the Fuzzy ARTMAP Neural Network

Charles Courchaine
National University
United States
charles@courchaine.dev

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

CCS CONCEPTS

• **Information systems** → *Evaluation of retrieval results; Information retrieval.*

KEYWORDS

TAR, Legal, eDiscovery, Fuzzy ARTMAP

ACM Reference Format:

Charles Courchaine and Ricky J. Sethi. 2023. Opening the TAR Black Box: Developing an Interpretable System for eDiscovery Using the Fuzzy ARTMAP Neural Network. In *Proceedings of European Conference on Information Retrieval (ECIR '23)*. ACM, New York, NY, USA, 3 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

Technology-assisted review (TAR) utilizes an information retrieval system to discover all, or nearly all, the relevant documents in a corpus and help reduce the human effort required to find these documents [7, 9, 20]. TAR systems are employed in high-recall tasks such as e-discovery, systematic literature reviews, evidence-based medicine, and information test collection annotation [9, 20]. These systems often employ a document classifier and an active learning component to select what documents a human should review [8, 21]. A 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 process [12]. Despite the 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 [7, 17].

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 [7, 17]. Instead, we looked at the explainable Fuzzy ARTMAP algorithm. The model learned by the Fuzzy ARTMAP algorithm can be directly interpreted geometrically [4, 19] or as a set of fuzzy If-Then rules [5, 6], depending on the features used.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ECIR '23, April 02-06, 2023, Dublin, Ireland

© 2023 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00

<https://doi.org/XXXXXXX.XXXXXXX>

We 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:

- performing a hyperparameter sweep to refine the parameters
- evaluating the system against the 20Newsgroups, Reuters-21578, RCV1-v2, and Jeb Bush emails corpora for recall, precision, and F_1 , and
- generating If-Then rules of document relevance

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 [20, 22] because legal matters are often confidential [7, 9] and their corpora are unavailable. The 20Newsgroups corpus is commonly used as a test corpus with ART-based algorithms [18, 19]; it and the Reuters-21578 corpus are also commonly used in evaluating text classification algorithms [1].

2 FUZZY ARTMAP

Adaptive Resonance Theory (ART) describes how the brain learns and predicts in a non-stationary world [14]. This theory models how brains can quickly learn new information without forgetting previously learned information. ART has been implemented in numerous neural network architectures for supervised, unsupervised, and reinforcement learning applications [3]. Fuzzy ART is a neural network algorithmic instantiation of ART that utilizes operators from fuzzy set theory; specifically, the fuzzy AND operator, to work with real-valued features [4]. The supervised version of the Fuzzy ART algorithm is the Fuzzy ARTMAP algorithm that maps between inputs and categories. By integrating fuzzy set theory and ART dynamics in the Fuzzy ARTMAP neural network algorithm, various interpretations of the learned model are possible. What the model learns may then be represented as fuzzy If-Then text-based rules or depicted geometrically [4, 13].

To take advantage of the geometric interpretation, however, the input must be complement encoded. Complement encoding is a normalization method when working with Fuzzy ARTMAP [4] in which the input vector \mathbf{x} is concatenated with its complement $\bar{\mathbf{x}}$, yielding an input of $\mathbf{I} = [\mathbf{x}, \bar{\mathbf{x}}]$. As a result, the categories learned by the Fuzzy ARTMAP algorithm can be interpreted as n -dimensional hyper-rectangles [4, 19]. When interpreting the model geometrically, the learned weights from the first half of the vector, the non-complement encoded portion, form one corner of the hyper-rectangle, and the second half of the vector, the complement-encoded portion, forms the other corner as illustrated in Figure 1.

3 METHOD

For the 20Newsgroups, Reuters-21578, RCV1-v2, and Jeb Bush emails corpora, we used tf-idf features with the smaller corpora and the 300-dimension versions of the GloVe and Word2Vec vectorizations with all of the corpora. All the topics in 20Newsgroups, 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 [22] due to memory constraints, retaining the general prevalence per topic. For each topic, the Fuzzy ARTMAP algorithm was trained with ten relevant documents and 90 non-relevant documents regardless of corpora size, and the review was run with batches of 100 for the smaller corpora and 1,000 for the larger corpora. The review of documents for each topic concluded when the algorithm predicted no more relevant documents in the unevaluated portion of the corpus. The Fuzzy ARTMAP algorithm was modified to report the degree of fuzzy subsethood [4, 16] associated with documents predicted as relevant, and this degree of fuzzy subsethood was then used to rank the documents for active learning. Based on the results of a sweep of the Fuzzy ARTMAP neural network algorithm hyperparameters, which evaluated different combinations of vigilance (ρ) and learning rates (β), vigilance was set to .95, and a fast learning rate of 1.0 was selected.

A proof-of-concept of one of these If-Then rules for the tf-idf vectorization was produced for predicting documents belonging to the pc.hardware category of the 20Newsgroups dataset, reproduced in Table 1. The tf-idf feature is in italics, and the level of prevalence is in bold. For this example, the level of prevalence was quantized into three levels: rarely, somewhat, and highly prevalent. Additionally, an example of the geometric interpretation is shown and discussed in Figure 1.

4 RESULTS AND DISCUSSION

Considering all corpora and vectorizations, the Fuzzy ARTMAP-based system achieved 100% recall 31% of the time, and achieved the suggested floor of 75% [15] or better recall 67% of the time, as seen

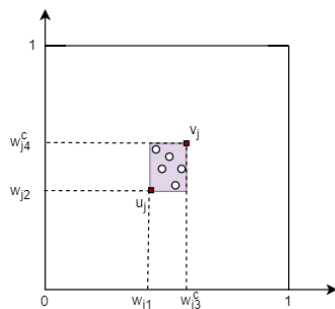


Figure 1: With complement encoding and a 2-dimensional input, the j^{th} category represented by weight vector w can be interpreted geometrically as a rectangle with corners u_j and v_j , with u_j corresponding to the first and second positions of the vector, and v_j corresponding to the complement encoded third and fourth positions. The circles inside the rectangle indicate inputs that fall within the category bounds.

Table 1: Excerpt of Rule Output for pc.hardware

| | |
|----------------------|--|
| Document is Relevant | |
| IF | <i>advance</i> is rarely prevalent in document |
| and | <i>apr</i> is rarely prevalent in document |
| and | <i>bogus</i> is rarely prevalent in document |
| and | <i>browning</i> is highly prevalent in document |
| and | <i>calstate</i> is rarely prevalent in document |
| and | <i>drive</i> is somewhat prevalent in document |
| ... | |

for median recall, precision, and F_1 in Table 2. Recall between the vectorizers for the Reuters-21578 and 20Newsgroups corpora was different by a statistically significant degree based on a Friedman test [11] with $p < .001$ ($\chi^3(2)=25.09$ and $\chi^3(2)=34.9$). A post-hoc Nemenyi test [11] indicated a difference between tf-idf and both GloVe and Word2Vec, with the average difference and statistical significance shown in Table 3. 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.

These results indicate generally robust recall performance, particularly with the tf-idf vectorization. Except for the Jeb Bush Emails, and the GloVe vectorization of 20Newsgroups, the median recall was 75% or better. In the more informal corpora of 20Newsgroups and the Jeb Bush Emails, the GloVe and Word2Vec features did not perform as well. However, this may be due to the corpus specificity of tf-idf compared with the off-the-shelf vocabulary of GloVe and Word2Vec. This suggests that generating corpus-specific GloVe and Word2Vec representations may perform better than the default vocabulary. Future research opportunities exist in optimizing the If-Then rule generation for the tf-idf vectorization and presenting textual and graphical explanations of Word2Vec and GloVe vectorizations. Additionally, exploring corpus-specific versions of Word2Vec and GloVe may bring recall in line with tf-idf, presenting a more efficient yet equally robust option.

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 [2]; in addition, there are also no qualitative or quantitative user studies of the existing prior attempts at explainability in e-discovery TAR [7, 17]. Therefore, this represents another likely productive area of future work.

Conclusion: This foundational research provides additional support for using the Fuzzy ARTMAP neural network as a classification algorithm in the TAR domain. While research opportunities exist to improve recall performance and explanation, the robust recall results from this study and the proof-of-concept demonstration of If-Then rules for tf-idf vectorization strongly substantiate that a Fuzzy ARTMAP-based TAR system is a potentially viable explainable alternative to "black box" TAR systems.

REFERENCES

[1] Berna Altunel and Murat Can Ganiz. 2018. Semantic Text Classification: A Survey of Past and Recent Advances. *Information Processing & Management* 54, 6 (Nov. 2018), 1129–1153. <https://doi.org/10.1016/j.ipm.2018.08.001>

Table 2: 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 3: Average Recall Difference

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

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

- [2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Benetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion* 58 (June 2020), 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- [3] Leonardo Enzo Brito da Silva, Islam Elnabarawy, and Donald C. Wunsch. 2019. A Survey of Adaptive Resonance Theory Neural Network Models for Engineering Applications. *Neural Networks* 120 (Dec. 2019), 167–203. <https://doi.org/10.1016/j.neunet.2019.09.012>
- [4] Gail A. Carpenter, Stephen Grossberg, Natalya Markuzon, John H. Reynolds, and David B. Rosen. Sept./1992. Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analog Multidimensional Maps. *IEEE Transactions on Neural Networks* 3, 5 (Sept./1992), 698–713. <https://doi.org/10.1109/72.159059>
- [5] Gail A Carpenter and Ah-Hwee Tan. 1993. Rule Extraction, Fuzzy ARTMAP, and Medical Databases. In *Proceedings of the World Congress on Neural Networks*. Erlbaum Associates, Portland, OR, USA, 501–506.
- [6] Gail A. Carpenter and Ah-Hwee Tan. 1995. Rule Extraction: From Neural Architecture to Symbolic Representation. *Connection Science* 7, 1 (Jan. 1995), 3–27. <https://doi.org/10.1080/09540099508915655>
- [7] Rishi Chhatwal, Peter Gronvall, Nathaniel Huber-Fliflet, Robert Keeling, Jianping Zhang, and Haozhen Zhao. 2018. Explainable Text Classification in Legal Document Review a Case Study of Explainable Predictive Coding. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, Seattle, WA, USA, 1905–1911. <https://doi.org/10.1109/BigData.2018.8622073>
- [8] Rishi Chhatwal, Nathaniel Huber-Fliflet, Robert Keeling, Jianping Zhang, and Haozhen Zhao. 2017. Empirical Evaluations of Active Learning Strategies in Legal Document Review. In *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, Boston, MA, 1428–1437. <https://doi.org/10.1109/BigData.2017.8258076>
- [9] Gordon V. Cormack and Maura R. Grossman. 2015. Autonomy and Reliability of Continuous Active Learning for Technology-Assisted Review. *arXiv:1504.06868 [cs]* (April 2015). [arXiv:1504.06868 \[cs\]](https://arxiv.org/abs/1504.06868)
- [10] Charles Courchaine and Ricky Sethi, J. 2022. Fuzzy Law: Towards Creating a Novel Explainable Technology-Assisted Review System for e-Discovery. In *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, Osaka, Japan, 1218–1223. <https://doi.org/10.1109/BigData55660.2022.10020503>
- [11] Janez Demsar. 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. *The Journal of Machine Learning Research* 7 (Dec. 2006), 1–30.
- [12] Seth Katsuya Endo. 2018. Technological Opacity & Procedural Injustice. *Boston College Law Review* 59, 3 (March 2018), 822–875. <https://doi.org/bclr/vol59/iss3/2>
- [13] Stephen Grossberg. 2020. A Path Toward Explainable AI and Autonomous Adaptive Intelligence: Deep Learning, Adaptive Resonance, and Models of Perception, Emotion, and Action. *Frontiers in Neurobotics* 14 (June 2020), 36. <https://doi.org/10.3389/fnbot.2020.00036>
- [14] Stephen Grossberg. 2021. Toward Autonomous Adaptive Intelligence: Building Upon Neural Models of How Brains Make Minds. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 51, 1 (Jan. 2021), 51–75. <https://doi.org/10.1109/TSMC.2020.3041476>
- [15] Robert Keeling, Rishi Chhatwal, Peter Gronvall, and Nathaniel Huber-Fliflet. 2020. 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* 26, 3 (2020), 65.
- [16] Bart Kosko. 1986. Fuzzy Entropy and Conditioning. *Information Sciences* 40, 2 (Dec. 1986), 165–174. [https://doi.org/10.1016/0020-0255\(86\)90006-X](https://doi.org/10.1016/0020-0255(86)90006-X)
- [17] Christian J. Mahoney, Jianping Zhang, Nathaniel Huber-Fliflet, Peter Gronvall, and Haozhen Zhao. 2019. A Framework for Explainable Text Classification in Legal Document Review. In *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, Los Angeles, CA, USA, 1858–1867. <https://doi.org/10.1109/BigData47090.2019.9005659>
- [18] Dušan Marček and Michal Rojček. 2017. The Category Proliferation Problem in ART Neural Networks. *Acta Polytechnica Hungarica* 14, 5 (2017), 15.
- [19] Lei Meng, Ah-Hwee Tan, and Donald C. Wunsch II. 2019. *Adaptive Resonance Theory (ART) for Social Media Analytics*. Springer International Publishing, Cham, 45–89. https://doi.org/10.1007/978-3-030-02985-2_3
- [20] Eugene Yang, David D. Lewis, and Ophir Frieder. 2019. 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*. ACM, Montreal QC Canada, 153–162. <https://doi.org/10.1145/3322640.3326713>
- [21] Eugene Yang, David D. Lewis, and Ophir Frieder. 2021. On Minimizing Cost in Legal Document Review Workflows. *arXiv:2106.09866 [cs]* (June 2021). <https://doi.org/10.1145/3469096.3469872> [arXiv:2106.09866 \[cs\]](https://arxiv.org/abs/2106.09866)
- [22] Eugene Yang, Sean MacAvaney, David D. Lewis, and Ophir Frieder. 2021. Goldilocks: Just-right Tuning of BERT for Technology-Assisted Review. *arXiv:2105.01044 [cs]* (May 2021). [arXiv:2105.01044 \[cs\]](https://arxiv.org/abs/2105.01044)