Towards Explainable Total Recall in TAR for eDiscovery Using Retraining of the Underlying Neural Network Model

Charles Courchaine[1], Corey Wade[1], Tasnova Tabassum[2], Stetson Daisy[2] and Ricky J. Sethi[1,2]

[1]National University, United States
[2]Fitchburg State University, United States

Abstract

Previous work has suggested that Fuzzy ARTMAP (FAM)-based Technology-Assisted Review (TAR) achieves viable levels of recall for eDiscovery (>75%), and produces models that are explainable via graphical and textual interpretation. However, these results also indicated room for improvement in recall performance, as FAM is sensitive to its training input. We evaluated the viability of improving recall performance through retraining the model based on documents evaluated as relevant from all prior review iterations. Retraining improved recall significantly, resulting in 72.9% of topic-vectorizer pairs being over the 95% recall threshold as compared to 42.2% without retraining. In addition, the FAM-based model continued to demonstrate self-stopping behavior even with retraining.

Keywords

Technology Assisted Review, Fuzzy ARTMAP, e-discovery, Stopping Problem, XAI

1. Introduction

Technology-assisted review (TAR) is a useful tool in high recall retrieval (HRR) tasks, such as systematic reviews and eDiscovery; in such areas, recall targets can range from 75-80% up to nearly 100% [1]. It is acknowledged that reaching high levels of recall is important in HRR but this is balanced with the level of effort required to reach high levels of recall [2, 3]. While existing efforts can reach these high levels of recall, they are typically black box implementations that provide little or no interpretability [4, 5, 6]. Previous work with Fuzzy ARTMAP-based TAR has indicated that it achieves levels of recall consistent with applicability to eDiscovery and, with tf-idf, GloVe, or Word2Vec features, allows the model to be interpreted in a variety of ways [7, 8]. While this initial implementation reached recall thresholds above 75%, there was significant room for improvement in the recall performance. It was hypothesized that retraining the model from scratch on previously reviewed relevant documents would increase

ALTARS'24: 3rd Workshop on Augmented Intelligence for Technology-Assisted Reviews Systems, March 28, 2024, Glasgow, Scotland

*Corresponding author.
charles@courchaine.dev (C. Courchaine); Corey@WadeML.dev (C. Wade); ttabassu@student.fitchburgstate.edu (T. Tabassum); sdaisy@student.fitchburgstate.edu (S. Daisy); rickys@sethi.org (R. J. Sethi)
0000-0002-5404-9438 (C. Courchaine); 0009-0007-7570-2372 (C. Wade); 0009-0000-9184-0600 (S. Daisy); 0000-0001-5254-3750 (R. J. Sethi)

recall performance, since Fuzzy ARTMAP is sensitive to its training input [9, 10]. In this paper, we present the early results of implementing up to three rounds of model retraining and its effect on recall performance.

2. Background

Fuzzy ARTMAP is a supervised classifier, one of many neural network algorithms derived from Adaptive Resonance Theory (ART), which *maps* inputs to category labels [10, 9]. ART describes how the brain learns and predicts in a non-stationary world [11]. One of the key features of models produced from this theory is that they can quickly incorporate new information without suffering from catastrophic forgetting [12]. In particular, Fuzzy ARTMAP utilizes the fuzzy AND operator from fuzzy set theory, instead of the binary set union operator, to work with values on the interval [0, 1] [10]. An important feature of the model produced by Fuzzy ARTMAP is that it can be presented as a set of fuzzy If-Then rules or graphically through a geometric interpretation [10, 13]. To employ the geometric interpretation, the input must be complement encoded. Complement encoding normalizes the input vector $\boldsymbol{x}$ by concatenating it with its complement $\overline{\boldsymbol{x}}$ (or $1 - \boldsymbol{x}$), yielding an input of $\boldsymbol{I} = [\boldsymbol{x}, \overline{\boldsymbol{x}}]$ [10]. With complement encoding, the categories in the Fuzzy ARTMAP model can be interpreted as n-dimensional hyper-rectangles [10, 14].

While Fuzzy ARTMAP has significant benefits in terms of online learning and interpretability, it is sensitive to the ordering of training examples [9, 10]. One way to mitigate this limitation is through a voting strategy where three or five models are trained with inputs in different orders; however, this is for using Fuzzy ARTMAP as a traditional classifier [10]. In TAR, new training examples are available after each review iteration. These new examples, including documents judged as relevant and not relevant, are used to update the Fuzzy ARTMAP model without full retraining of the model [7, 8].

3. Procedure

In previous work, we evaluated Fuzzy ARTMAP performance in TAR and found robust recall performance; however, there were several instances where Fuzzy ARTMAP failed to achieve 75% or better recall [7, 8]. In this earlier work, the Fuzzy ARTMAP implementation stopped when it predicted no more relevant documents. As Fuzzy ARTMAP is sensitive to the input order of the training examples, we extended our previous work by clearing and retraining the model with documents previously evaluated as relevant. The clearing and retraining of the model occurred when the model predicted no more relevant documents, up to a set number of retraining events.

We set the number of retrainings at three, established through a mix of cost balancing and small empirical tests. In general, the baseline, with no retraining, had runs that took a range of times from sub-seconds to 13-minutes; with retraining, the range of times increased exponentially, more than tripling for most runs and some taking 4 to 9 hours.

Ultimately, this resulted in up to four models produced, the first (Model 0 in Table 1) was trained with ten relevant documents and ninety non-relevant documents. This model was used

Table 1
Model Iteration and Training Source Documents

Model	Initial Training Source	Review Iterations
Model 0	Randomly chosen 10 Relevant, 90 Non-Relevant documents	0 to *i*
Model 1	Random order of all evaluated Relevant documents in iterations 0 to *i*	*i*+1 to *j*
Model 2	Random order of all evaluated Relevant documents in iterations 0 to *j*	*j*+1 to *k*
Model 3	Random order of all evaluated Relevant documents in iterations 0 to *k*	*k*+1 to *stop*

and updated with the relevance judgements in each iteration until no more relevant documents were predicted in the i^{th} review iteration. A new model was then created from all the documents evaluated as relevant in the prior review iterations, presented in a random order to the model (e.g. Model 1 in Table 1). The process was repeated up to the limit of three additional models, and ended when the final model predicted no more relevant documents. The rest of the model implementation was kept the same as the baseline in [7, 8]. Briefly, the baseline implementation is recapitulated below.

We maintained the same hyperparameter values from our previous work, a baseline vigilance ($\overline{\rho_a}$) of 0.95 and a fast learning rate (β) of 1.0. For these early results, we evaluated the retraining modifications with the 20 Newsgroups and Reuters-21578 corpora, vectorized with tf-idf, GloVe in 300-dimensions, and Word2Vec in 300-dimensions. All topics were used from 20 Newgroups, and the 119 topics with relevant documents were used from the Reuters-21578 corpus, yielding 417 samples overall. Word vectors for each document were averaged to produce document representations [15]. For all vectorizations the representations were scaled to the [0,1] interval using the scikit-learn MinMaxScaler. Finally, the features were complement encoded prior to processing via Fuzzy ARTMAP [10]. The Fuzzy ARTMAP TAR implementation takes a continuous active learning (CAL) approach [16], and the model is updated after each review iteration. For each review iteration up to 100 documents were reviewed, based on the number of documents the model predicted as relevant. To rank the documents for active learning, the *degree of fuzzy subsethood* was used [7, 8, 17].

4. Results

The effect of retraining on recall performance was substantial. The average difference in recall, and other metrics, between retraining and no retraining is shown in Table 2. Statistical significance was calculated using a one-way paired t-test, with $p < .001$ across all corpora and vectorizers for recall indicating statistically significant difference with retraining. The smallest improvement was six percentage points of recall, with most improvements in the 15 to 19 percentage point range, up to a maximum of 39 percentage points improvement in recall. This improvement in recall is further illustrated in Table 3, where *72.9% of topic-vectorizer pairs (304 of 417) achieved 95% recall or better* ***with*** *retraining as compared with 42.2%* ***without*** *retraining.* Commensurate with this improvement in recall performance is a decrease in precision, mostly between 13 and 21 percentage points. This was a statistically significant decrease in precision across corpora and vectorizers ($p < .001$), as calculated by a one-way paired t-test.

Table 2
Average Metric **Difference** Between Retraining and No Retraining

		$Recall_{diff}$	$Precision_{diff}$	F_{1diff}	$RE\text{-}75_{diff}$
Reuters-21578	GloVe	.19***	-.13***	-.17***	311***
	Word2Vec	.18***	-.14***	-.17***	299***
	tf-idf	.15***	-.21***	-.26***	275***
20 Newsgroups	GloVe	.39***	-.37***	-.02	1653***
	Word2Vec	.19***	-.21***	-.2***	1117**
	tf-idf	.06***	-.19***	-.22***	-1

**p <.01, ***p <.001
Increases are positive, decreases are negative

Table 3
Number of Topic-Vectorizer Pairs Over Recall Threshold for new model (with retraining) vs the baseline model (without retraining)

	0.75	0.80	0.85	0.90	0.95
Retrain Over Threshold	403 (96.64%)	391 (93.76%)	376 (90.17%)	345 (82.73%)	304 (72.90%)
Baseline Over Threshold	315 (75.54%)	272 (65.23%)	239 (57.31%)	212 (50.84%)	176 (42.21%)

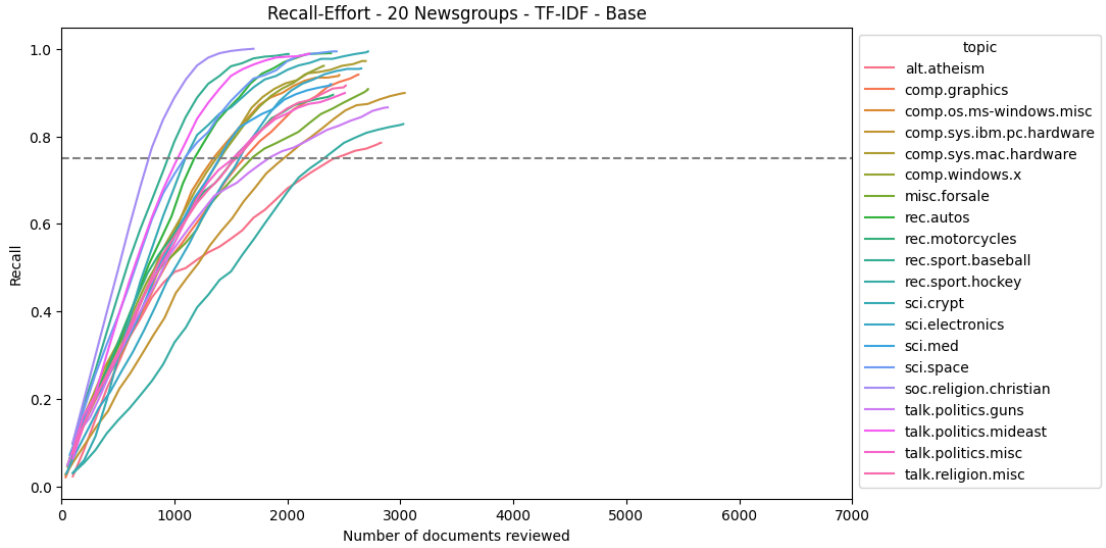

Figure 1: Number of documents to achieve recall, without retraining.

In general, while F_1 performance did decrease by a practical and statistically significant amount ($p < .001$, except for 20 Newsgroups-GloVe), the drop was generally *more modest*, with the *improvement in recall offsetting the drop in precision.* Also interesting was the general increase in the **average difference** in the number of documents required for review to reach

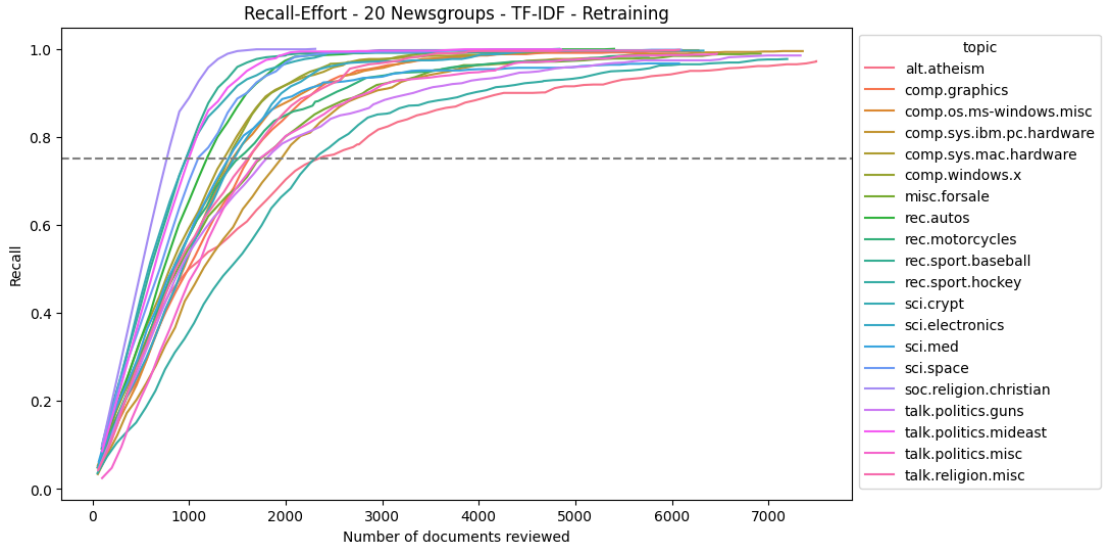

Figure 2: Number of documents to achieve recall, **with** retraining, showing an average improvement of 6 percentage points over Figure 1.

75% recall; illustrated in Table 2 under the RE-75$_{diff}$ metric. This is partly due to the number of un-retrained topic-vectorizer pairs that never reached 75% recall, compared with the retrained instances (24.46% vs. 3.36%). The increase is relatively modest for the Reuters-21578 corpus, around 300 documents or about 1.5% of the 19,044 documents with bodies. For GloVe and Word2Vec the difference is more substantial in 20 Newsgroups, around 1,400 documents or 8.5% of the 16,330 distinct posts. The effect of the retraining and number of documents reviewed is illustrated in the difference between Figure 1 and Figure 2; where more documents are reviewed, but higher recall is ultimately achieved in Figure 2 than Figure 1.

5. Conclusion

Utilizing full retraining of the Fuzzy ARTMAP model, even a modest number of times, improved recall significantly - **attaining 95% recall in over 72.9% topic-vectorizer pairs with retraining as compared to 42.2% without retraining**. Currently, the improvement in recall comes at the cost of precision (Table 2 - Precision$_{diff}$ and RE-75$_{diff}$); however, some of the increase in the number of documents to reach 75% recall is due to the no retraining model only reaching 75% recall 75% of the time compared with 96% of the time with retraining (Table 3). This improvement in recall is achieved *while retaining the characteristic that the algorithm eventually predicts no more relevant documents*. Additional research is required to further characterize this *stopping behavior*, evaluating it across more data sets, attempting to define the statistical and theoretical basis for the predictions of no more relevant documents, and tuning the number of retraining iterations to potentially target recall levels. There is additional research opportunity in improving precision along with the improvement in recall.

References

[1] E. Yang, S. MacAvaney, D. D. Lewis, O. Frieder, Goldilocks: Just-right tuning of bert for technology-assisted review, arXiv:2105.01044 [cs] (2021).

[2] D. Li, 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 38 (2020). doi:10.1145/3411755.

[3] E. Yang, D. Lewis, Heuristic stopping rules for technology-assisted review (2021). doi:10.1145/3469096.3469873.

[4] R. Chhatwal, P. Gronvall, N. Huber-Fliflet, R. Keeling, J. Zhang, H. Zhao, Explainable text classification in legal document review a case study of explainable predictive coding, 2018. doi:10.1109/BigData.2018.8622073.

[5] C. J. Mahoney, J. Zhang, N. Huber-Fliflet, P. Gronvall, H. Zhao, A framework for explainable text classification in legal document review, IEEE, 2019, pp. 1858–1867. doi:10.1109/BigData47090.2019.9005659.

[6] C. Mahoney, P. Gronvall, N. Huber-Fliflet, 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), IEEE, Osaka, Japan, 2022, pp. 2044–2051. doi:10.1109/BigData55660.2022.10020626.

[7] C. Courchaine, R. J. Sethi, Fuzzy law: Towards creating a novel explainable technology-assisted review system for e-discovery, IEEE, 2022, pp. 1218–1223. URL: https://ieeexplore.ieee.org/document/10020503/. doi:10.1109/BigData55660.2022.10020503.

[8] C. Courchaine, T. Tabassum, C. Wade, R. J. Sethi, Explainable e-discovery (xed) using an interpretable fuzzy artmap neural network for technology-assisted review, 2023, pp. 2761–2766.

[9] L. E. B. da Silva, I. Elnabarawy, D. C. Wunsch, A survey of adaptive resonance theory neural network models for engineering applications, Neural Networks 120 (2019) 167–203. URL: https://doi.org/10.1016/j.neunet.2019.09.012. doi:10.1016/j.neunet.2019.09.012.

[10] G. A. Carpenter, S. Grossberg, N. Markuzon, J. H. Reynolds, D. B. Rosen, Fuzzy artmap: A neural network architecture for incremental supervised learning of analog multidimensional maps, IEEE Transactions on Neural Networks 3 (1992) 698–713. URL: http://ieeexplore.ieee.org/document/159059/. doi:10.1109/72.159059.

[11] S. Grossberg, Toward autonomous adaptive intelligence: Building upon neural models of how brains make minds, IEEE Transactions on Systems, Man, and Cybernetics: Systems 51 (2021). doi:10.1109/TSMC.2020.3041476.

[12] S. Grossberg, Competitive learning: From interactive activation to adaptive resonance, Cognitive Science (1987). doi:10.1111/j.1551-6708.1987.tb00862.x.

[13] S. Grossberg, A path toward explainable ai and autonomous adaptive intelligence: Deep learning, adaptive resonance, and models of perception, emotion, and action, 2020. doi:10.3389/fnbot.2020.00036.

[14] L. Meng, A.-H. Tan, D. C. W. II, Adaptive Resonance Theory (ART) for Social Media Analytics, Springer International Publishing, 2019, pp. 45–89. doi:10.1007/978-3-030-02985-2_3.

[15] A. Carvallo, D. Parra, H. Lobel, A. Soto, Automatic document screening of medical literature

using word and text embeddings in an active learning setting, Scientometrics 125 (2020). doi:10.1007/s11192-020-03648-6.
[16] G. F. Cormack, M. F. Grossman, Autonomy and reliability of continuous active learning for technology-assisted review, arXiv (2015).
[17] B. Kosko, Fuzzy entropy and conditioning, Information Sciences 40 (1986) 165–174. doi:10.1016/0020-0255(86)90006-X.