Hyperparameter Optimization for Predicting the Tolerance Level of Religious Discourse

Donald E. Brown[1], Hope McIntyre[1], Peter J. Grazaitis[2], Riannon M. Hazell[2], and Nicholas Venuti[1]

[1] Data Science Institute, University of Virginia, Charlottesville, VA USA
brown@virginia.edu
https://dsi.virginia.edu/
[2] Human Research and Engineering Directorate, U.S. Army Research Laboratory, Aberdeen Proving Ground MD 21005

Abstract. To address the rising tide of religious violence as it affects U.S. Military deployments, the Army requires analytic methods that can be generalized to predict religious group violence around the globe and scalable to the number of potential groups in an area of operation. Current computational methods based on semantics and topics are lacking in predictive performance for the generalized problem and topic modeling performs poorly in predicting the tolerance level of new groups towards U.S. Military presence.The research in this paper aims to discover the association between religious speech and behaviors and provide a foundation for proactive engagement with these groups. The approach builds on the work from ethnolinquistics to model how things are said (performative analysis) rather than word meanings (semantic analysis). Recent research has developed computational approaches to streamline the manually intensive performative analysis of religious text. While producing promising results, these computational methods lack systematic optimization of the hyperparamters in the learning algorithms. Hence, we do not know the sensitivity of the results to parameter settings. This paper reports on results for predicting religious tolerance by optimizing the parameters in the signal processing algorithms and shows that the predictive power of performative approach is robust to parameter settings.

Keywords: Behavior Analysis, Military, Computational Linguistics

1 Introduction

The U.S. Army's role continues to expand into operations that are "characterized by ambiguity in the nature of the conflict, the parties involved, or the relevant policy and legal frameworks" [15]. These operations frequently require interactions with the indigenous populations with various communities, groups, state and non-state actors for effective mission execution and success. It is well recognized by the U.S. Military that ideological based groups including religious groups pose threats to mission success [14]. This paper describes research that focuses on the use of the language of religious non-state actors to predict their behaviors.

Many approaches to predicting religious attitudes use the literal or emotive attributes of religious speech as encoded in responses to questions and scenarios [1].

In contrast, some scholars of religions have argued for an approach to understanding religious discourse using the capacity for religious language to encapsulate an action or identity and the flexibility of this language for use in inter-group communications [12]. This approach is *performative* analysis.

However, application of these ideas to actual dialog and group interactions is difficult because of the time and effort required to manually process religious text from a variety of groups. To develop the classification for the documents used in this paper, we have had teams of students and faculty members review documents. This process takes approximately one hour for a 500-750 word document.

Recent work has applied machine learning with signal processing to help automate performative analysis [16]. While showing promising results in labeling the tolerance levels of religious groups, the signal processing used in [16] was not optimized. Hence, we have no idea how sensitive those results were to parameters in the models. This paper provides that sensitivity analysis and gives results for computational performative analysis of religious documents with optimized hyperparameters.

2 Related Work

Linguistic approaches to religious text have focused primarily on semantics to classify the goals and intentions of groups (e.g., see [3]. The resulting correlations with behaviors are often difficult to understand and interpret. For example, Hassner, et al., used Latent Semantic Analysis (LSA) to track temporal shifts in language usage for Iranian leaders. While topical shifts were identified in this approach, few predictive capabilities were developed through this methodology [9]. When researchers have employed quantitative methods for semantic analysis they have narrowly focused on specific incidences of violence that lack generality [17].

In contrast to semantic analysis, performative analysis considers word usage as signaling intentions. Barsalou [2] showed that shifts in definitions correlate with the representational flexibility of a concept. Additionally, Sagi, et al., [13] found that a diversity of contexts directly correlates with the performative characteristics of words.

Results by [5] showed that neither word stop-lists nor word stemming significantly improved performance on word-word co-occurrence statistics. Minimal context window size gave the best performance, while dimensionality reduction through singular value decomposition (SVD) also improved performance. Boussidan and Ploux [4] used a graph built from subsetted co-occurrence tables to create a map of lexical usages of words. While they got good results, the complexity of computing cliques limits the scalability of their method. [6] used a similar approach to detect 'amelioration' (a word losing a negative meaning) and 'pejoration.'

3 Data

The data for the analysis comes from online repositories for the groups and religious leaders shown in Table 1. Students and faculty members in the Global Covenant of Religion (GCR) [11] rated the documents produced by each group with a language flexibility score from 1-9 where 1 means the least flexible use of language and 9 means

very encompassing use of language. The GCR has a systematic approach for obtaining this score and their members can provide details [11]. This flexibility score serves as the response variable for supervised learning. Table 1 shows the scores, affiliation and number of documents for each group or religious leader.

Table 1. Data Sources

Group	Score	Affiliation	Number
Westboro Baptist	1	Baptist	419
Faithful Word Baptist	2	Baptist	228
Nouman Ali Khan	3	Sunni Muslim	88
Dorothy Day	4	Catholic	774
John Piper	4	Baptist	579
Steve Shepherd	4	Christian	728
Rabbinic Texts	6	Jewish	166
Unitarian Texts	7	Unitarian	276
Meher Baba	8	Spiritualist	265

The documents were randomly put in bins of 5 or more documents. We cleaned the text by removing punctuation, converting to lowercase, and converting all numbers to a single symbol. The tokens were stemmed, but stopwords (such as, “the”, “an”) were not removed. For Part-of-Speech (POS) labeling we applied the Maxent POS Tagger from the python Natural Language Toolkit (NLTK) package to the corpus [10]. Word counts were generated for each unique word/POS tag combination for each bin. The top 10 most frequent adjectives and adverbs were selected for each bin and used as the keywords. For each group we put 70% of the bins in a training set and the remaining 30% in a testing set.

4 Hyperparameter Optimization

There are two major components we used for the computational performative analysis of religious text: text signal processing and machine learning. The work in this paper focused on optimizing hyperparameters for signal processing or feature engineering. The details of the different signals acquired from religious text and the subsequent machine learning methods are given in [16]. Here we describe the optimization of the three major signals: co-occurrence window; context window; and network adjacency angle.

To capture the variations in linguistic flexibility of keywords within religious discourse, we implemented a context vector semantic density algorithm in python based on research by [13]. Using the pre-processed tokens as described in the data section above, a co-occurrence matrix, $X = [x_{ij}]$ was constructed with $i, j \in \{1, 2, \ldots, v\}$, where v is the size of the vocabulary. The x_{ij} elements of X capture how often each word appeared with other words in the vocabulary. This was done by iterating over each word in the bin and counting the words within a $\pm k$ sized co-occurrence window.

This co-occurrence window directly affects the representation of the distribution of the language within the corpus. A larger window includes more information about the words occurring around other words. This may better capture the linguistic signals but it may also add noise. Our optimization explored windows of size 2 to 6.

Once the co-occurrence matrix, X, was constructed, a distributional semantic matrix, D, was developed to reduce the computational load. D was obtained from X using truncated SVD. This reduced the column space of D to 50 components.

Next, context vectors were created from D. The context vectors for each keyword in the bin were developed by extracting the words within an *r*-sized context window surrounding the target and summing the rows of D for the words within the window. In contrast to the co-occurrence window which measures general word usage, the context window size impacts the specific measure of a word's usage. Without knowledge of how the proximity of a word affects the analysis of the variability of its usage, we varied the window size from 2 to 6.

Utilizing the distributional semantic matrix, D, we created a graph to estimate semantic density using the igraph package in python [7]. This was done by first creating a v-v adjacency matrix by computing the cosine similarity of each row h in D. Values in this matrix were converted to 0 or 1 if they were above or below the determined network adjacency angle threshold, respectively. This matrix was then used to create a graph where a node represents a word and an edge is assigned between two nodes if the associated value in the adjacency matrix is 1.

The network adjacency angle can significantly impact the accuracy of estimates of word usage variability. This is due to the fact that an inaccurate value of network adjacency angle can over or under estimate the relationship of words in the graph. To optimize network adjacency angle its value was varied by 15 degree increments over the range 15 - 75.

5 Results and Conclusions

To judge the relationship of the hyperparameters to linguistic flexibility we used two measures for linguistic flexibility: semantic density and eigenvalue centrality. Eigenvalue centrality is defined in [8]. The semantic density is computed for each target bin word. Let C_i be the set of context vectors for target bin word i. We then estimated the average cosine similarity of all the context vectors for that word by randomly sampling 2 context vectors from the set, C_i and calculated the cosine similarity between them. This was performed over $n = 1000$ iterations. The resulting values were averaged to produce the semantic density, SD, for that target word as shown below.

$$SD(w_i) = \frac{1}{n}\sum_{k=1}^{n}\frac{c_{ia_k}\cdot c_{ib_k}}{||c_{ia_k}|| * ||c_{ib_k}||}$$

where a_k, b_k are randomly chosen from C_i in iteration k.

Optimization results show that up to a point an increase of the co-occurrence and context vector window sizes results in an increase in the average semantic density of the keywords. For example, for random forests as the machine learning method the optimal

value occurs when both co-occurrence and context vector window are set to 5. These values allow for important signals to be held in the context vectors, while avoiding excess noise.

The response surface for average eigenvector centrality was multi-modal. Nonetheless, the average eigenvector centrality trends upward as the co-occurrence window size increases and the network adjacency angle decreases. An increase in the co-occurrence window size results in a greater connection between the words within the discourse, and the lower adjacency angle produces more edges between the nodes within the graphs. As eigenvector centrality is a bounded variable between 0 and 1, the limit of 1 occurs under the extreme conditions of 30 degrees and a co-occurrence window size of 6.

For model performance, we used a robust measure that allows an error margin of 1, as shown in Eq. 1 and Eq. 2. Let m be the machine learning method, i.e., random forests or support vector machines (SVM). Let $\hat{y}$ be the predicted flexibility score and y be the actual score. Finally, let B be the set of document bins. So, for $b \in B$ the accuracy, acc, is given by

$$acc(b, m) = \begin{cases} 1, & \text{if } |\hat{y} - y| \leq 1 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

$$acc(m) = \frac{1}{|B|} \sum_{b \in B} acc(b, m). \tag{2}$$

Finally, Table 2 shows the overall increase in accuracy from parameter optimization. The table shows the machine learning method (RF: random forest and SVM: support vector machines), the previous results reported in [16], and the results from optimized hyperparameters. These results come from three-fold cross-validation.

Table 2. Accuracy Comparison for Signal Processing

Method	Previous	Optimized
SVM	84%	86%
RF	86%	92%

The results show performative analysis can provide promising predictions of the flexibility evident in religious documents from a wide range of groups and the methods are not highly sensitive to parameter settings.The results in Table 2 indicate modest loss of accuracy from using default or arbitrarily chosen settings. This suggests that carefully chosen settings will put the performance in ranges not much different from those found by hyperparameter optimization. Clearly the robustness this approach suggests that the signals, i.e., the feature engineering when combined with the classification methods is doing the heavy lifting in predicting the flexibility of religious speech. Notice that the difference between the performance of optimized random forests and support vector machines is the same as the difference previous random forest results and hyperparameter optimized results.

These results suggest that future work should focus on finding new signals or relationships between signals. Clear areas for this new work are in the application of word embedding techniques which may produce signals not easily engineered or identified by feature engineering.

References

1. Abu-Nimer, M.: Conflict resolution, culture, and religion: Toward a training model of inter-religious peacebuilding. Journal of Peace Research 38(6), 685–704 (2001)
2. Barsalou, L.: Flexibility, structure, and linguistic vagary in concepts: Manifestations of a compositional system of perceptual symbols. Theories of memory 1, 29–31 (1993)
3. Blatter, B., Patel, V.: Exploring dangerous neighborhoods: latent semantic analysis and computing beyond the bounds of the familiar. In: AMIA 2005 Symposium Proceedings. pp. 151–155 (2005)
4. Boussidan, A., Ploux, S.: Using topic salience and connotational drifts to detect candidates to semantic change. In: Proceedings of the Ninth International Conference on Computational Semantics. pp. 315–319. Association for Computational Linguistics (2011)
5. Bullinaria, J.A., Levy, J.: Extracting semantic representations from word co-occurrence statis- tics: stop-lists, stemming, and svd. Behavior research methods 44(3), 890–907 (2012)
6. Cook, P., Stevenson, S.: Automatically identifying changes in the semantic orientation of words. In: LREC (2010)
7. Csardi, G., Nepusz, T.: The igraph software package for complex network research. InterJournal, Complex Systems 1695(5), 1–9 (2006)
8. Estrada, E., Rodriguez-Velazquez, J.A.: Subgraph centrality in complex networks. Physical Review E 71(5), 056103 (2005)
9. Hassner, R.E.: War on sacred grounds. Cornell University Press, Ithica, NY (2009)
10. Loper, E., Bird, S.: Nltk: The natural language toolkit. In: Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics- Volume 1. pp. 63–70. Association for Computational Linguistics (2002)
11. Marcus, B.: Global covenant of religion. https://sites.google.com/a/globalcovenant.org/global-covenant/team (2015)
12. Ochs, P.: The possibilities and limits of inter-religious dialogue. In: Omer, A., Appleby, R.S., Little, D. (eds.) Religion, Conflict, and Peacebuilding, pp. 488–534. Oxford University Press (2015)
13. Sagi, E., Kaufmann, S., Clark, B.: Semantic density analysis: Comparing word meaning across time and phonetic space. In: Proceedings of the Workshop on Geometrical Models of Natural Language Semantics. pp. 104–111. Association for Computational Linguistics (2009)
14. Tompkins, P.J.: Human factors considerations of undergrounds in insurgencies. U.S. Army Special Operations Command, Ft. Bragg, NC (January 2013)
15. United States Special Operations Command: Gray zone. https://info.publicintelligence.net/USSOCOM-GrayZones.pdf (September 2015)
16. Venuti, N., Sachtjen, B., McIntyre, H., Mishra, C., Hays, M., Brown, D.E.: Predicting the tolerance level of religious discourse through computational linguistics. In: 2016 IEEE Systems and Information Engineering Design Symposium (SIEDS). pp. 309–314. IEEE Press (2016)
17. Yang, M., Wong, S., Coid, J.: The efficacy of violence prediction: a meta-analytic comparison of nine risk assessment tools. Psychological bulletin 136(5), 740 (2010)