Computational Analysis of Religious and Ideological Linguistic Behavior

Seth Green, Megan Stiles, Katherine Harton, Samantha Garofalo,
and Donald E. Brown
University of Virginia, smg5b, mes5ac, klh8mr, smg7un, deb@virginia.edu

Abstract - **In today's global environment, effective communication between groups of diverse ideological beliefs can mean the difference between peaceful negotiations and violent conflict. At the root of communication is language, and researchers at the University of Virginia Center for Religion, Politics, and Conflict (RPC) hypothesize that the analysis of the performative character of a group's discourse (how words are used) provides valuable guidance for how to negotiate with groups of differing ideological beliefs. However, high pressure situations leave little time for an exhaustive analysis of this nature. To address this challenge, this paper expands on the signal processing approach of previous work in the literature, which evaluated the efficacy of a computational approach to applying performative analysis to predict linguistic rigidity. Significantly, this paper evaluates the generalizability of computational performative analysis by considering text from non-religious groups. These include groups focused on political and social agendas rather than religion. The key computational and analytical improvements described in this paper include an enriched judgment selection process and the extraction and analysis of pronoun usage. By examining the raw text of various religious and ideological groups, results show an improved accuracy of 97% for predicting linguistic rigidity, compared to the best predictive accuracy of 83% reported in previous work. These results strengthen the evidence for the hypothesis of the effectiveness of computationally implemented performative analysis as predictive of linguistic rigidity. The results also provide evidence that this approach is applicable to non-religious groups since the predictive accuracy is as consistent with these groups as it is for religious groups.**

Index Terms - Machine learning, Natural language processing, Religious conflict

INTRODUCTION

As the world becomes more globalized, the need to efficiently engage with groups from different cultures and ideologies becomes increasingly paramount. The ability to quickly analyze how best to interact with different groups is particularly important for military and humanitarian organizations who oftentimes have to make quick decisions on the ground. Unfortunately, working with some value-based groups requires a certain amount of background knowledge and expertise that these organizations may not have. To help combat this need for immediate expertise in the field, the UVa Center for Religion, Politics, and Conflict (RPC) hypothesizes that an analysis of the performative character of discourse, defined as "the capacity for language to encapsulate an action or identity," has the potential to make recommendations to US-based and international organizations in determining the most effective way of engaging with local groups on the ground [1]. The RPC team developed a manual method for measuring the performative character of discourse using a 1-9 linguistic rigidity scale, with a score of 1 indicating a rigid use of language and a score of 9 indicating a flexible use of language [1].

The general interpretation of the linguistic rigidity score is the average number of meanings that can be attributed to deep-value words in that group's discourse. For example, a word like "holy," which is deeply imbued with value judgement, could be used very narrowly by one group to refer only to those who have sacrificed their lives for the cause, whereas another group might use "holy" in a number of contexts, to imply compassion, wisdom, or trustworthiness, and possibly *also* some notion of self-sacrifice. In this example, the first group would be scored a 1, while the second group would likely be a 3 or a 4. The underlying hypothesis motivating this analysis is that a group's movement towards one end of the scale or the other indicates that their discourse is in a sense breaking down, and that this is an indicator of broader strife in the community or the culture at large. This information could be critical in helping military or humanitarian organizations determine which groups will be easier to work with, and which will not.

Previous work by Venuti et al. developed an automated approach to the RPC Team's manual method and achieved an accuracy rate of 83% [2]. The automated method used a combination of judgment identifiers and semantic density analysis to score the linguistic rigidity of religious text written by nine different religious groups or individuals. To further the work done by Venuti et al., the University of Virginia's Computational Linguistics (CL) team continued to improve the signals in the signal processing model to more closely replicate the manual process of the RPC team. The CL team also collected more diverse data to use in

training the models. The results from the signal improvements and data collection are reported in this paper.

RELATED WORK

The methods in this paper draw on work from religious and ideological linguistics, natural language processing and machine learning. First consider religious and ideological linguistics. Fundamentally, ideological negotiation brings its own set of challenges. Wade-Benzoni et al. argue that only after the fabric of a group is analyzed can real-world decisions between groups be made and understood [3]. In addition, peace between local and global actors need "'strategic' planning and implementation" [4]. The RPC team's hypothesis thoughtfully analyzes the use of language and how it might differ between these local and global participants.

Natural language processing (NLP) is a range of techniques that allow machines "to analyze, understand, and derive meaning from human language in a smart and useful way" [5]. NLP has been applied to a diverse set of tasks, such as sentiment analysis, translation, and topic segmentation, and it sits at the intersection of many different fields, including text mining [5]. Text mining transforms text into data in order to discover relevant information that can be used for future analysis [6]. NLP is then the methodology that deciphers the textual data in order to gain understanding of the linguistic question at hand.

Of particular interest for this paper is the use of text mining and NLP for keyword selection. Especially as the amount of online text grows, keyword selection has been used in applications like search engines, text categorization, and topic detection [7]. An important step of keyword selection is how to represent each word and the choice of the term weight. Term frequency (TF) or binary representations are used, but the term frequency-inverse document frequency (TF-IDF) is a common metric and one described in more detail in the *Keywords* section of this paper. Term weighting is typically done using only the words within the documents being analyzed, but Gabrilovich et al. label text documents "with concepts from different knowledge bases" using outside corpora [8]. Given the range of ideological groups in the dataset, the CL team tests a similar idea in combination with the TF-IDF weighting method.

In order to make these linguistics concepts into something computational, various machine learning techniques have been used. The first of these is random forest, a tree-based method that segments the predictor space into regions [9]. Random forest can be applied to both regression and classification problems, and the method reduces variance found in other tree-based methods by decorrelating each tree grown [9]. Because of its structure, random forest is able to handle highly dimensional data. Both Palomino-Garibay et al. and Treeratpituk et al. use random forest in text classification [10] [11]. Another machine learning technique used in NLP is support vector machine (SVM). Similar to random forest, SVM can be used in both regression and classification problems. In a highly dimensional space, SVMs use kernels to draw non-linear boundaries between classes, relying on a few influential observations called support vectors [9]. Joachims empirically shows that SVMs work well for text categorization, mainly due to the method's ability to work in high dimensional spaces with sparse vectors [12]. Özgür et al. focused on the use of keywords in text categorization using SVM [13]. Lastly, boosting is an approach that improves prediction results by fitting "a decision tree to the residuals from the model" [9]. Instead of fitting the decision tree to the outcome values of the function, a boosted tree fits to the residuals from the previous tree [9]. In each iteration, the decision tree is added to the fitted function before it, allowing the approach to slowly learn without overfitting [9].

Given this research in the fields of linguistics, NLP, and machine learning, this paper describes applying a combination of these techniques to religious and ideological text.

DATA

The CL team obtained data for this work from a number of different sources. The first was the repository that Venuti et al. had collected [2]. This repository consists of 3,285 documents from nine different groups of various affiliations. The CL team collected additional sources of data using custom-built web scrapers using the *Beautiful Soup* package in Python [14]. This increased the total number of documents to 4,568, which consisted of 15 different groups from various affiliations. The CL team converted all the files to .txt files and then pre-processed them to be used for the models described below.

The RPC team manually scored a random sample of documents on their 1-9 scale for each group, which was used to calculate an overall group rank. The information in Table I shows all of the groups that were manually scored by the RPC team and therefore used to train the model created by the CL team. In addition, the table contains information on the total number of documents collected per group. Notice that these are not just religious groups, the data also contains text from environmental and political action groups, as well as other value-based groups. This created a richer data set that ranges across various political and ideological spectra.

PRE-PROCESSING

The CL team used pre-processing methods similar to those seen in [2]. Many of the new documents collected, shown in Table 1, were in PDF format, so an OCR was used to convert the documents to .txt files, as this is the format of the input into the models. From here, punctuation was removed, every character was converted to lowercase, and numbers were removed. In addition, processed tokens were created by tokenizing on any whitespace, and these tokens were stemmed. All of the parts of speech were tagged uniquely using the POS tagger, which became an important

aspect of this study, as will be seen in the sections below. Each of these methods were made possible by Python's *nltk* package [15]. A final pre-processing method that was an important part of previous studies was binning 10 documents together in order to fulfil the requirements for increased observations and text length needed to create the models [2]. All binned documents were written by the same group and the overall group score was used as ground truth for each document bin.

TABLE I
DATA SOURCES

Group	# Documents	Group Rank
ACLU	40	3
American Ethical Union	15	8
Bahai	73	6
Dorothy Day	774	4
Integral Yoga/Yogaville	59	6
ISIS	48	1
John Piper	579	4
Liberal Judaism	166	6
Malcolm X	15	2
Meher Baba	265	8
Pastor Anderson	228	1
Rabbinic	58	4
Sea Shepherds	606	2
Steve Shepherd	728	4
Unitarian	301	7
Westboro Baptist Church	422	1

SIGNALS

When taking unstructured data, such as free text, as input to a machine learning model, it is often necessary to engineer features or signals from the data. The following section describes the signals engineered by the CL team, as well as several signals from [2] that were included in the CL team's model.

I. Keywords

Performative signals used in this analysis rely on the programmatic selection of keywords, or what the RPC team calls deep-value words. The identification of these words is the key to the manual linguistic analysis, performed by the RPC team, that this research is based upon.

Previous studies approached this problem by using a POS-tagger to identify the most frequently used adjectives and adverbs in a document and then used the ten most frequent as keywords. Consultation with the RPC team showed that a number of important keywords were not being identified through this method, so the CL team implemented another method, which did not rely on POS-tagging.

The keyword selection method is based on the common technique of the term frequency-inverse document frequency (TF-IDF) matrix. A TF-IDF score represents how common a word is *in a given document*, weighted by the inverse of how many documents *in the entire corpus* that word appears in. However, in this case, if the words are central to a group's discourse, then those words are likely to appear in most of the documents in the corpus, which would result in them being weighted down by a high IDF score. In contrast, the CL team sought to weigh the potential keywords by how common a word is *in general usage*, as opposed to the specific corpus of religious and ideological documents.

As many others have used Wikipedia as a general corpus, for example [8] and [16], the CL team also chose to use Wikipedia as a general usage corpus for keyword selection. To make the corpus more manageable, and to filter out redirects and disambiguations, the CL team subset Wikipedia to only articles over 30,000 characters (15,876 articles total), which were then pre-processed identically to the new documents (stemmed, lower-cased, etc.). A tabulation was then made of all terms that appeared in ten or more documents (87,709 terms total).

Equation (1) was then used to calculate TF-IDF scores for each pre-processed token in a given document *D*,

$$\frac{n_{id} \times 15876}{|\{w: n_{iw} \neq 0\}|} \tag{1}$$

where n_{id} is the number of occurrences of word i in document D, and n_{iw} is the number of occurrences of word i in Wikipedia article w. The potential keywords were then ranked by these scores and the top twenty were chosen.

When manually spot-checking individual documents with the RPC team, the new method picked out more of the correct keywords that the RPC team had chosen manually. The estimated recall was about 60%, as compared to about 25% for the previous method [2].

II. Performative Signals

The CL team incorporated several of the signals which were engineered in [2] into their model. Principally among these were the Average Semantic Density and Eigenvector Centrality features.

The CL team calculated Average Semantic Density by collecting context vectors for each occurrence j of each selected keyword w_i. The context vector for any word w_i "contains information about the specific contexts in which w_i appeared" [2]. In principle, these could be word embeddings or any other representation. The exact representation the CL team used is described in more detail in [2].

Average Semantic Density for a given document with N keywords is then calculated as an average of the cosine similarities between these context vectors, defined by (2).

$$\frac{1}{N}\binom{l}{2}^{-1} \sum_{i=1}^{N} \sum_{j<k}^{l} \frac{\langle c_{ij}, c_{ik} \rangle}{\|c_{ij}\| * \|c_{ik}\|} \tag{2}$$

Where c_{ij} is the context vector for occurrence j of w_i. By averaging these similarities between the contexts at each of the l occurrences of w_i, a scalar score is obtained (for each keyword w_i) which reflects the breadth or dearth of different contexts in which that word appears. The mean of these scores is then taken to get a scalar Average Semantic Density score for given document.

It should be noted that the complexity of computing the cosine similarities between n vectors is $\Theta(n^2)$. In this case, n would be $\binom{l}{2}$ which would quickly become computationally infeasible. Therefore, as in [2], we estimate it using a Monte Carlo simulation with 1,000 iterations.

For the Eigenvector Centrality signal, the CL team constructed a network graph using the *igraph* package in Python [17]. The nodes of the graph correspond to keywords and the edges were weighted by the cosine similarity between the corresponding context vectors of those keywords. Any edges whose weight was below a set threshold (0.524) were removed. The influence of keyword w_i can be quantified by the number of remaining edges connected to it. Eigenvector Centrality then measures "the influence of a network upon a node by looking at the number of influential nodes to which it is connected" [2].

III. Judgments

In addition to signal processing methods used in previous work as described above, the CL team created a new judgment selection method in order to more accurately identify the proportion of judgment statements made in each document. According to the RPC team, a judgment is a statement that conveys a group's opinion on the outside world. Previous computational studies defined a judgment sentence as any sentence that contained a noun, a version of the verb "to be," and an adverb or adjective [2]. Based on the improved method of keyword selection, the CL team created an algorithm that identified a judgment sentence as any sentence that contained a keyword and a POS-tagged pronoun. The judgment signal was then calculated as the number of judgment sentences in the document bin divided by the total number of sentences in each document bin.

IV. Pronouns

The CL team also engineered additional signals that corresponded to the percentage of words in each document bin that were pronouns. The CL team took this one step further by breaking the different types of pronouns into seven different classes as shown in Table II.

TABLE II
PRONOUN SIGNALS

Pronoun Class	Pronouns
We	We, Our, Us
You	You, Your, Yourself
Me	Me, My, Myself, I
They	They, Them, Their
He	He, Him, Himself, His
She	She, Her, Herself, Hers
It	It, Itself, Oneself

The seven pronoun signals were then calculated by counting the number of each pronoun class used in each document bin divided by the total word count of the document bin.

Figures I and II in the *Results* section show the range of pronoun usage for "we" and "they" pronouns, when plotted against the group's ground truth linguistic rigidity score.

CLASSIFICATION APPROACHES

I. Classification vs. Regression

Previous research formulated this as a regression problem, based on the RPC team's description of their linguistic rigidity scale as a continuously valued score [2]. However, after further consultation, the CL team decided to reframe it as an ordinal classification problem because, while the values on the scale do indeed correspond to a continuously valued estimate of the number of contexts in which a group allows its deep-value words to be used, there are also a number of other performative traits of the speech which change as a group moves up or down the scale.

Formulating this as classification allows the model to more intuitively place a document in the class which most resembles its performative style, as opposed to placing it on a continuous scale. While there is ongoing debate about which method better approximates the RPC's manual methodology, as will be shown in the *Results* section, switching to classification showed significant boosts in predictive accuracy.

II. Support Vector Machines

SVMs are a powerful and popular classification method when dealing with noisy data. The CL team chose to use SVMs for this reason and to have a baseline against which to compare the results of this paper to the results in the research by Venuti et al. The CL team used the *kernlab* package in R, with a linear kernel, setting the regularization parameter C to 1 [20].

III. Random Forest

Random forests have also been shown to perform well in tasks such as this, especially when there is multicollinearity present in the predictors, which is the case in this research [2]. The CL team modeled using the *randomForest* package in R with a tree count of 500 and limiting the number of variables considered at each split to $m=\sqrt{p}$ [21].

IV. Gradient Boosted Trees

Gradient tree boosting has been shown to increase prediction accuracy by iteratively optimizing a loss function. The CL team modeled using the *xgboost* package in R and a 9-class "multi:softprob" objective, with a max depth of 5, and 5 rounds [22].

RESULTS

As previously mentioned, Figures I and II show the range of pronoun usage. These plots are informative, and encouraging, for two reasons. First, they show discernable patterns across the linguistic rigidity score. In particular, they show that the pronoun signals separate out the prediction space in a meaningful way. The word "they" is used frequently by groups on the low end of the scale (groups of high rigidity) while it is used infrequently by

groups on the high end of the scale (groups of low rigidity). In contrast, the use of "we" is used frequently for groups in the middle of the rigidity scale and infrequently for groups of high rigidity. Second, they show patterns which the RCP team has observed through manual analysis; specifically, that groups who are typically more challenging to deal with are "externally" focused (using "they" more than "we") while groups in the middle of the scale are more focused on examining and interpreting their own traditions (indicated by the higher "we" scores) [18] [19].

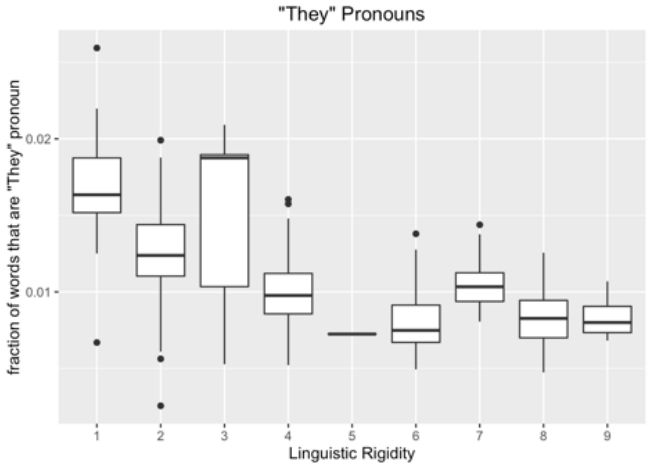

FIGURE I
"THEY" PRONOUN USAGE

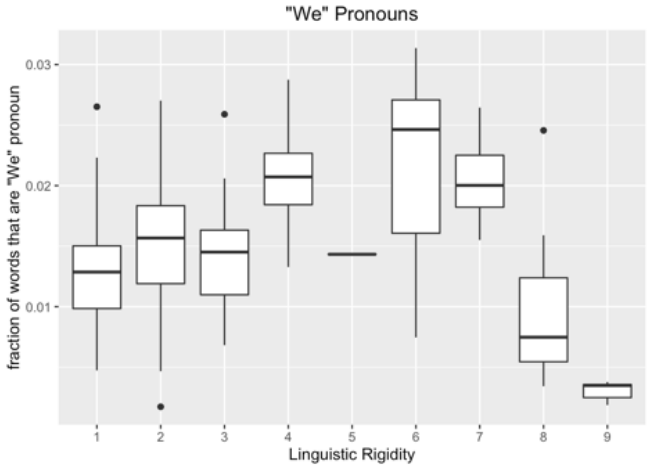

FIGURE II
"WE" PRONOUN USAGE

Equations (3) and (4) define model accuracy. The CL team defined a prediction to be accurate if it was within a margin of one, the same metric used by [2].

$$acc(bin, model) = \begin{cases} 1, |\hat{y} - y| \leq 1 \\ 0, otherwise \end{cases} \quad (3)$$

$$acc(model) = \frac{1}{|bins|} \sum_{bin \in bins} acc(bin, model) \quad (4)$$

As shown in Table III, the methods used by the CL team showed an improvement in classification accuracy compared to previous performative signal processing methods [2]. Table III shows these results using only the documents used in [2], and Table IV shows the results of the CL team's improved approach, using all of the documents listed in Table I.

The models using the improved signals increased the predictive accuracy of [2]'s models by over 13%. The three models, random forest, support vector machine, and boosted trees all had accuracy rates within 1%, with an average accuracy of 96.6%. In addition, the models retained their high accuracy with the addition of non-political and more diverse religious groups, with an average accuracy of 96.0%. This significantly broadens the applicability of this method by suggesting that the RPC team's hypothesis has implications beyond strictly religious group discourse.

TABLE III
RESULTS COMPARED TO PREVIOUS METHODS

Model	Accuracy	Venuti Score [2]
Random Forest	96.5	86.0
SVM	97.2	84.0
Boosted Tree	96.2	N/A
Artificial Neural Net	N/A	80.0
Average	96.6	83.3

TABLE IV
RESULTS USING FULL DATA SET

Model	Accuracy
Random Forest	96.1
SVM	96.3
Boosted Tree	95.4
Average	96.0

CONCLUSIONS AND FUTURE WORK

As our world becomes ever more global, the importance of peaceful communications between different ideological and religious groups will continue to increase. The ability to quickly and accurately assess a group and understand how to work with groups of differing ideologies is crucial to achieving peaceful outcomes. The performative signals, judgments, and pronoun usage of different groups have shown to predict a group's linguistic rigidity score with a high degree of accuracy, when using the RCP team's manual process of document scoring as ground truth. Furthermore, these signals continue to show predictive power when applied to non-religious ideological groups, extending the application of this research far beyond religious groups.

A more practical application of this work includes the ability to be able to track a group's use of language over time in order to identify potential spikes or deviations from the group's baseline. In order to accomplish this the CL team will run the model on a single-document level rather than on binned documents. This presents numerous challenges in that each input will contain fewer words, however, a successful analysis of single documents will have significant implications for real-time, on-the-ground analysis of ideological and religious groups.

ACKNOWLEDGMENT

The CL team would like to thank Professor Peter Ochs, Jonathan Teubner, Essam Fahim, and Syed Moulvi of the University of Virginia's Religious Studies department for their expertise and critical insights. The work for this paper is partially supported by a grant from the U.S. Army Research Laboratory.

REFERENCES

[1] Consultation with researchers at the University of Virginia Center for Religion, Politics, and Conflict (RPC). 1540 Jefferson Park Avenue, University of Virginia, Charlottesville, VA 22904-4126.

[2] Venuti, Nicholas, Sachtjen, Brian, McIntyre, Hope, et al. "Predicting the Tolerance Level of Religious Discourse Through Computational Linguistics," presented at the *2016 IEEE Systems and Information Engineering Design Conference (SIEDS '16),* Charlottesville, VA, 2016.

[3] Wade-Benzoni, Kimberly A., Hoffman, Andrew J., Thompson, Leigh L., et al. 2002. "Barriers to resolution in ideologically based negotiations: The role of values and institutions." *Academy of Management Review* 27, pp. 41-57.

[4] Lederach, John Paul and Appleby, Scott R. "Strategic Peacebuilding: An Overview," in Daniel Philpott and Gerard Powers, Strategies of Peace (Oxford: Oxford University Press, 2010), pp. 19-44.

[5] "Algorithmia." 2016. http://blog.algorithmia.com/introduction-natural-language-processing-nlp/. Accessed: March 30, 2017.

[6] "Expert System: Natural language processing and text mining." 2016. http://www.expertsystem.com/natural-language-processing-and-text-mining/. Accessed: March 30, 2017

[7] Gupta, Vishal. and Lehal, Gurpreet S., 2009. "A survey of text mining techniques and applications." *Journal of emerging technologies in web intelligence*, *1*(1), pp.60-76.

[8] Gabrilovich, Evgeniy and Markovitch, Shaul. January 2007. "Computing semantic relatedness using wikipedia-based explicit semantic analysis." (In *IJcAI* Vol. 7), pp. 1606-1611

[9] James, Gareth, Witten, Daniela, Hastie, Trevor, et al. 2013. *Introduction to Statistical Learning*, pp 303-372.

[10] Palomino-Garibay, Alonso, Camacho-González, Adolfo T., Fierro-Villaneda, Ricardo A., et al. 2015. "A random forest approach for authorship profiling." *Cappellato et al. [8]*.

[11] Treeratpituk, Pucktada and Giles, C. Lee. June 2009. "Disambiguating authors in academic publications using random forests." In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries,* pp. 39-48.

[12] Joachims, Thorsten, 1998. "Text categorization with support vector machines: Learning with many relevant features." *Machine learning: ECML-98*, pp.137-142.

[13] Özgür, Arzucan, Levent Özgür, and Tunga Güngör. "Text categorization with class-based and corpus-based keyword selection." *International Symposium on Computer and Information Sciences*. Springer Berlin Heidelberg, 2005.

[14] Richardson, Leonard. *Beautiful soup documentation*, [Online] Available: https://www.crummy.com/software/BeautifulSoup/bs4/doc/.

[15] Bird, Steven, Edward Loper and Ewan Klein (2009), *Natural Language Processing with Python*. O'Reilly Media Inc. Available: http://nltk.org/book

[16] Pennington, Jeffrey, Richard Socher, Christopher D. Manning. "GloVe: Global Vectors for Word Representation." *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing.* 2014.

[17] Csardi, G. and Nepusz, T. *Package 'igraph'*. [Online]. Available: http://igraph.org/python/

[18] Íñigo-Mora, Isabel. "On the use of the personal pronoun we in communities." *Journal of Language and Politics.* Volume 3, Issue 1, 2004, pages: 27–52.

[19] Maitland, Karen and John Wilson. "Pronominal selection and ideological conflict." *Journal of Pragmatics*. Volume 11, Issue 4, August 1987, Pages 495-512.

[20] Karatzoglou, Alexandros, Alex Smola, and Kurt Hornik, *Package 'kernlab'*.[Online] Available: https://cran.r-project.org/web/packages/kernlab/kernlab.pdf

[21] Liaw, Andy and Matthew Wiener, R *Package 'randomForest'*. [Online] Available: https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

[22] Chen, Tianqi, Tong He, Michael Benesty, Vadim Khotilovich, and Yuan Tang, *Package 'xgboost'* [Online] Available: https://CRAN.R-project.org/package=xgboost

AUTHOR INFORMATION

Seth Green, M.S. Student, Data Science Institute, University of Virginia.

Megan Stiles, M.S. Student, Data Science Institute, University of Virginia.

Katherine Harton, M.S. Student, Data Science Institute, University of Virginia.

Samantha Garofalo, M.S. Student, Data Science Institute, University of Virginia.

Donald E. Brown, Director, Data Science Institute and William Stansfield Calcott Professor, University of Virginia.