A Comparison of Language Dependent and Language Independent Model for Violence Prediction

Ben Greenawald, Elaine Liu, Gregory Wert, Mohammad al Boni,
and Donald E. Brown
University of Virginia, bgh5yd, yl9qr, gaw8pa, ma2sm, deb@virginia.edu

Abstract - **The changing nature of violence under the recent development of world political order, religious belief, and technology pose challenges to governmental organizations and non-governmental organizations to bring forth timely strategy in engaging with groups advocating violence. While some groups are well-known for their violence, as the characteristics of groups vary throughout time and across regions, the traditional decision-making process takes time and resources that organizations will not always have. As such, a scalable and effective technique to identify the violent groups becomes imperative. This paper applies text analysis techniques to differentiate violent and non-violent groups using English text from various value-based groups.**

The models presented in this paper achieved an accuracy of at least 71% and as high as 81%. Such high performance in accuracy along with other evaluation metrics demonstrates that text analysis provides a powerful solution for predicting the violent groups and the non-violent groups. In addition, by incorporating natural language processing tools, the language-dependent models show an improvement in accuracy over language-independent models. However, this improvement was only about 2%. The overall similar performance of language-dependent and language-independent models suggest the two approaches are comparable alternatives.

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

INTRODUCTION

The challenge of linguistic and cultural diversity posed by value-based groups often requires governmental and non-governmental organizations to make immediate action based on their understanding of whether these groups tend to be violent or not. Given that it is often unattainable for organizations to acquire a large amount of knowledge and expertise associated with a certain value-based group before engaging with that group, it becomes essential for organizations to equip themselves with a solution scalable and flexible enough to detect violent groups in a timely fashion.

Based on the premise that the behavior of value-based groups can be inferred from their use of language, researchers at the UVa Center for Religion, Politics, and Conflict (RPC) and the University of Virginia's Computational Linguistics (CL) team developed text-mining algorithms that accurately evaluated important characteristics of language usage by religious and non-religious value-based groups [1][2]. Built upon this early work, this paper differs in two important ways. First, it focuses specifically on the question of violent behavior rather than the more general characteristics of group behavior. Second, in addition to the language-dependent approach, this paper incorporates the language-independent approach, meaning the algorithms do not rely on special knowledge of the input language.

For this study, a value-based group is a group that operates under a common name, has a primary mission outside of making a profit, and has a publicly available statement or set of values that generally reflect a worldview and historical narrative. It should be noted that under this definition, individuals can qualify as value-based groups. Violence is defined as the intentional use of physical force, threatened or actual, that has a high likelihood of causing human injury or death. A violent group is defined as a group whose members perform acts that fall under the above definition of violence, and the group must claim responsibility for that action.

In this study, text was collected from 13 violent groups and 23 non-violent groups. These groups were selected to contain a multitude of regional and ideological diversity. Vector-space models are the only language-dependent models used in this paper. The results of the ensemble models are also presented in this paper. The main contributions of this work include: 1) proposing language dependent and language independent models for violence prediction based on English text, and 2) comparing and evaluating the proposed models to demonstrate the viability of using text in predicting violent groups.

PRIOR WORK

Text classification and natural language processing (NLP) have worked to automate the analysis and interpretation of language. They have sought to find statistical means of replicating the interpretations humans assign. The applications of this can be wide and narrow. For instance, past work has found success classifying documents for ambiguous tasks such as finding the intent of a document [3]. On the other hand, researchers have also accomplished more specific tasks such as detecting semantic change within publications [4]. Text classification has, in general, shown versatility in its applications and has thus been applied in a wide array of tasks.

One such task has been the analysis of political discourse. This area has sought to find common themes and divergences within groups and individuals texts and speeches. One way this has been done is by analyzing with topics undergo the most semantic change within a corpus [4]. Techniques such as latent semantic analysis have also proved useful, as they have been able to examine framing within political discourse [5]. Analysis can extend to groups as well. Through techniques such framing analysis, researchers have detected distinctions in the discourse of two groups focused on the same issues [6]. There appear to be various downfalls to these techniques within the literature, however. Existing techniques struggle at nuance. For instance, latent semantic analysis does not always distinguish between the desire to commit an action and confessing towards having committed it [7].

Language classification has also been applied towards predicting violence. These predictions, however, have suffered in certain regards. One such case is predicting specific instances of violence. This has proven difficult because many incidences of violence are caused by specific environmental factors and are difficult to predict [8]. Similarly, researchers have struggled when trying to ascertain violent intent in communications. Here, they found that detecting features such as anger was difficult of the algorithms inability to classify unorthodox expressions of anger such as insults [9]. This negatively affected the algorithms ability to predict violent intent. However, analysis of violent communications through topic modeling has yielded categories which logically imply violence [9]. Recent work has found some success using diachronic modeling to semi-accurately predict future incidences of violence by groups based on past incidents of violence [10]. In general, studies over time or at more aggregate levels have shown greater success. For instance, research has shown that longitudinal analyses can be performed on individuals to examine changes in the level of aggressiveness within their texts and thus over time [5].

Alternatively, researchers have attempted to predict features related to violence. This work has examined linguistic rigidity: a measure for how flexible vocabulary within a document is. This work attempted to predict the linguistic rigidity of a document using performative signals and found some success in this approach [1]. Furthermore, these signals can attain very strong results using hyperparameter optimization [11]. From there, researchers have been able to identify patterns within the scores. They, for example, were able to notice how certain pronouns (such as they or we) are clustered around certain scores [2]. This has shown how computational techniques may infer trends less obvious to people, and these trends can then be analyzed for other purposes such as violence prediction.

Dataset

The dataset used was a collection of English documents compiled by the RPC team and extended by CL team [1][2]. A description of the dataset is shown in Table 1 below.

TABLE I

Word and Document count per group

Label	Group Name	Document Count	Total Word Count
Violent	Alqaeda	29	6,968
	Aqis	6	1,318
	ASL	7	878
	AST	235	744
	Boko Haram	18	1,758
	ISIS	65	405,469
	Nusra Front	11	1,194
	Sea Shepherds	605	2,043,349
	Stalin	3	23,363
	Taliban	6	1,298
	Unabomber	27	91,752
	KKK	93	453,391
	White Supremacists	571	3,256,459
	Total	**1,676**	**6,287,941**
Non-Violent	ACLU	39	64,386
	AEU	15	45,340
	Andrew Murray	1	264,111
	Bahai	73	126,166
	Dorothy Day	774	4,207,333
	Gandhi	12	257,683
	YV	59	177,118
	John Piper	579	4,440,264
	Liberal Judaism	166	947,084
	Malcolm X	15	255,095
	Mehr Baba	265	1,293,184
	Nauman Khan	84	1,158,921
	Pastor Anderson	228	7,419,541
	Peter Gomes	2	37,592
	Rabbinic	58	657,628
	Schizophrenia	70	142,536
	Shepherd	1,333	5,886,356
	Unitarian	301	3,152,747
	WBC	423	3,898,078
	Hamza Yusuf	8	395,814
	Yasir Qadhi	107	1,508,976
	Zakir Naik	28	69,101
	Suhaib Webb	20	77,425
	Total	**4,660**	**36,482,479**

Preprocessing

Any project involving text usually requires a fair bit of preprocessing on the raw data. Common actions include stemming (taking a word and returning only the stem of the word) and stop word removal (removing extremely common words). However, actions such as this would violate the assumptions of our language-independent models, as these techniques require knowledge of the language being used in the text. Thus, for the language independent models, the only preprocessing done was the removal of any non-

English characters, converting any character to lowercase, and the conversion from numbers to special characters ("NUM"). The conversion of numbers to special characters was a way to drastically reduce our initial feature space without losing information. In other words, "3" and "4" would have originally counted as two different features, however in the context of predicting violence, there really is no difference between "3" and "4", so we just converted both to the special character "NUM."

The preprocessing for the language-dependent models was much more thorough, since more complex, language-dependent methodologies could be used. Pre-processing for these models included (beyond what was done for language-independent) stemming the tokens, and performing part of speech tagging, which labels each token with its part of speech.

Feature Selection

Before analysis could begin, we needed to choose a feature space and then map our documents to feature vectors. This process was different for the language-independent models and the language-dependent model. For the language-independent models, we first chose our feature space, which in the context of this problem involves selecting which words (or tokens) will be used as predictors. Just choosing every possible word in the vocabulary seems like a viable option, but given that the vocabulary is extremely large, and most of the words in it are probably not very predictive, thus a mechanism was needed to whittle down the features.

We chose two different methods of feature selection, information gain and chi-squared [12]. Our final feature space was the intersection of these two methods. The first method was information gain. At a high level, this method provides an indication of how discriminatory a given feature is between the classes in the response. In other words, for a given token (say "God"), how well am I able to differentiate between violent and non-violent documents based on the presence, or lack thereof, of the word "God."

The second methodology used is the chi-squared statistic (2). For a given token *t* in the vocabulary and class *c*, the chi-squared statistic measures the dependence between *t* and *c*. In other words, is there some relationship between the presence of *t* and the outcome *c*. Since chi-squared is a normalized statistic, it can be measured against the chi-square distribution to determine the strength of the relationship. At a 95% confidence level, any feature corresponding to a chi-squared level greater than 3.841 will be significant. Indeed, we filtered out any insignificant features using this threshold.

For each of the previous measures, we computed the 10000 most significant features. Then, our final feature set became the intersection of these measures.

Feature Engineering - Language Independent

Now that the feature space was selected, we needed to map every document to a feature vector. To do this, we used a modification of the extremely popular text frequency-inverse document frequency encoding (TF-IDF). TF-IDF has two components. The first is text frequency, denoted *tf*, which involves counting up how often each feature appears in a given document. So if the word "God" appears 3 times for a given document, then the TF value for that document and the feature "God" is 3. However, we actually used sublinear TF, denoted *wt.* This modification accounts for the fact that just because a single terms appears twice as often in a document, this does not mean that the term is twice as important for the document. For a given document *d* and token *t*, the sublinear TF is defined in (1) [13].

$$wf(t,d) = 1 + \log(tf(t,d)) \\ \text{if } tf(t,d) > 0\text{, } 0 \text{ otherwise} \qquad (1)$$

The IDF component, denoted *idf*, represents how often a given term *t* appears in other documents. The reasoning here is that if a given token appears in almost every document, then it is not very effective at identifying any single document. The inverse document frequency for a given token *t* is defined as (2).

$$idf(t) = \log(\frac{N}{df(t)}) \qquad (2)$$

where *N* is the total number of documents in the corpus and *df(t)* is the total number of documents containing term *t*. Finally, the TF-IDF score for a given document *d* and a given token *t* is the multiplication of *wf(t,d)* and *idf(t)*. Doing this for each document gave us the set of feature vectors. It is important to note that now, part of this calculation depends on the text being in English, and easily works for any language.

Feature Engineering - Language-dependent

The feature vectors for the language-dependent models were replicated from the CL team [2]. We will not go into great detail on their feature engineering method, but will do a brief overview. The first set of features were keywords, which are similar to the TF-IDF described above, but used a global corpus, like Wikipedia, giving a sense of how a word performs in the language as a whole. The next set of features were performative signals, which map each keyword to vectors which attempt to quantify the context in which that keyword appears. Then come the judgment features. A judgment sentence is defined to be a noun, followed by some conjugation of the verb "to be," followed by an adverb or adjective. The final set of features are pronoun features, essentially measuring how often pronouns are used throughout a given document. The combination of these features creates the feature space for the language-dependent models. In contrast to the features defined in the previous section, all of these features require some knowledge of the language being used.

METHODOLOGIES

Initially, we attempted to fit and validate models using a simple cross-validation approach, however, this lead to a problem. The results obtained were being inflated based on superficial features. In other words, when the model was classifying documents from a given group, it was heavily relying on other documents it had seen from that group during training. This is not a realistic approach since the entire premise of this project was that the model will have never seen documents from a given group. Thus, we proposed a new validation form. Leave one group out-cross validation, or LOGO-CV, leaves all documents from a given group out during training, and uses those documents as the testing set. Such a methodology is much more realistic in the context of this problem. It should also be noted that we performed oversampling on the minority class during any given run in order to have a perfectly even class distribution during training.

The primary metrics of evaluation used were the F1 score, which is the harmonic mean between precision and recall, and the raw accuracy, which is the total number of correctly guessed observations over the total number of observations. So in order to perform the analysis, for each group we ran LOGO-CV, keeping track of the accuracy and the F1 score for each run. F1 scores were stored separately depending on the whether the left out group was violent or not, so we ended up with an F1 score for violent groups, and one for non-violent groups. Averaging the accuracies and F1 scores gave us overall performance for a given model.

We tested all models from the CL team, as well as three additional models. For each model (other than logistic regression and the ensemble models), a grid search was performed in order to determine good values for the hyperparameters. For any model that outputs probabilities (namely, logistic regression), a simple cutoff of 0.5 was used to determine classification.

I. Logistic Regression (LR)

This is considered a standard model when performing any sort of classification problem. Logistic regression, although it was not fit by Green et al. since they had a multiclass problem, whereas we had only a classification problem. We used a custom implementation of logistic regression developed in Java.

II. Support Vector Machines (SVMs)

We chose to use the sklearn implementation of SVMs in Python with a sigmoid kernel and a cost parameter or 100 [14]. All other parameters were kept as default.

III. Random Forest (RF)

Again, the sklearn implementation in Python was used with the number of trees set at 1000 and the max depth of a single tree at 10 [14]. All other parameters were kept as default.

IV. *Gradient Boosted Trees (GBT)*

The xgboost implementation was used with 1000 trees, a max depth of 10 for each tree, a learning rate of 0.5, and a binary logistic objective [15]. All other parameters were kept as default.

V. Average Ensemble Model (AE)

A new model we introduced, ensemble models take the predictions from other models and combine them into a new model. In this first implementation, we took the probabilities from the two top performing models, logistic regression and random forest, and averaged them into a single probability estimate, where 0.5 was used as the classification cutoff.

VI. Weighted Ensemble Model (WE)

This model also takes the probabilities from the same 3 models, but weights them according to model performance. In other words, if a given model had much higher performance than the others, its prediction will be given more weight than the other models. These weights were computed as the normalized accuracies for each model. For example, the weight for logistic regression would be as in (3) where *acc* is the accuracy for the given model.

$$w(LR) = \frac{acc(LR)}{acc(LR) + acc(RF)} \tag{3}$$

RESULTS

As described in previous sections, we performed empirical evaluations using LOGO-CV for the proposed language-dependent and language-independent models. We calculated the classification accuracy and two F1-measures (positive and negative) for all the models at the document level. Table 2 shows the LOGO-CV macro-classification performance.

TABLE 2
LOGO-CV RESULTS ON THE DIFFERENT MODEL CLASSES

Model	Model Name	Accuracy (threshold of 0.5)	Positive F1	Negative F1
Language Dependent	LR	78.97%	0.8511	0.8680
	RF	83.17%	0.8584	0.9018
	SVMs	71.02%	0.7623	0.7943
	AE	83.64%	0.8582	0.9088
	WE	83.56%	0.8600	0.9070
Language Independent	LR	81.81%	0.8409	0.8936
	RF	79.21%	0.9127	0.8136
	SVMs	75.99%	0.7786	0.8541
	GBT	79.92%	0.8196	0.8778
	AE	81.52%	0.8533	0.8947
	WE	81.57%	0.8346	0.8958

For language-dependent models, the average ensemble model achieved the best performance across all evaluation metrics. For language-independent models, LR and average ensemble models had comparable performance with relatively high accuracy and negative F1 score; RF outperformed other models in positive F1 score. The overall high performance of both language-dependent models and language-independent models demonstrate the viability of using text to predict the tendency towards violence of a value-based group. It is worth noting that language-dependent models which incorporated NLP tools yielded a slightly better performance.

Further, for the models with high F1 scores, LR AND RF, we compared their performance at a group level and showed groups with less than or equal to 50% accuracy. Table 3 shows that LR and RF both classified 5 groups with less than 50% accuracy. However, RF performed better on violent groups with only one group showing accuracy slightly less than 50%.

TABLE 3
GROUPS WITH LOW PERFORMANCE ON SELECTED MODELS

Model	Group Name	Is Violent	Accuracy	F1
LR	KKK	Yes	45.16%	0.6222
	Sea Shepherds	Yes	32.89%	0.4950
	Unabomber	Yes	37.04%	0.5405
	ACLU	No	17.95%	0.3043
	Malcolm X	No	20.00%	0.3333
RF	White Supremacist	Yes	46.58%	0.6356
	ACLU	No	0.00%	0.0000
	AEU	No	20.00%	0.3333
	Bahai	No	34.25%	0.5102
	Schizophrenia	No	17.14%	0.2927

CONCLUSIONS AND FUTURE WORK

Value-based groups have a vast impact on the political and social order of the world. However, cultural and language barriers often impede active engagements with those groups. The ability to quickly identify those group’s tendency towards violence can facilitate an effective strategy for organizations to interact with them. In this paper, we presented a solution to identify violent and non-violent value-based groups using various techniques of computational linguistics. We implemented both language-dependent models and language-independent models on a large corpus of English text associated with various value-based groups. We tested our models using a leave one group out cross-validation (LOGO-CV) and evaluated each model using accuracy score and F1 metrics. Overall, we were able to demonstrate that the text can be predictive of violent groups. Both language-independent and dependent approach are viable solutions. However, language-dependent models which incorporate language-specific natural processing tools (NLP) performed better than language-independent models with a slight improvement of 2% in accuracy.

Deep learning methods could provide a more advanced way to classify text in a language agnostic way. Advances in deep learning, especially in the area of text analysis, make these feasible options for such a task.

ACKNOWLEDGEMENT

We would like to thank Dr. Jonathan Teubner and Zain Moulvi of the University of Virginia’s Religious Studies department for their expertise and critical insights into the role of value-based groups in society. The work for this paper is partially supported by a grant from the U.S. Army Research Laboratory.

REFERENCES

[1] N. Venuti, B. Sachtjen, and H. McIntyre 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

[2] S. Green, M. Stiles, K. Harton, and S. Garofalo et al, “Computational analysis of religious and ideological linguistic behavior,” *2017 Systems and Information Engineering Design Symposium (SIEDS ‘17)*, pp. 359–364, Charlottesville, VA, 2016.

[3] M. Kröll and M. Strohmaier, “Analyzing human intentions in natural language text,” *Proceedings of the fifth international conference on Knowledge capture - K-CAP 09*, Sep. 2009.

[4] A. Boussidan and S. Ploux, “Using topic salience and connotational drifts to detect candidates to semantic change.,” *Proceedings of the Ninth International Conference on Computational Semantics*, pp. 315–319, Jan. 2011.

[5] K. L. Hacker, D. Boje, and V. L. Nisbett et al, A. Abdelali, and N. Henry, “Interpreting Iranian leaders’ conflict framing by combining latent semantic analysis and pragmatist storytelling theory,” *Political Communication Division of the National Communication Association annual conference*, Nov. 2013.

[6] N. E. Landrum, C. Tomaka, and J. Mccarthy, “Analyzing the Religious War of Words over Climate Change,” *Journal of Macromarketing,* vol. 36, no. 4, pp. 471–482, Dec. 2016.

[7] T. Cohen, B. Blatter, and V. Patel, “Exploring dangerous neighborhoods: latent semantic analysis and computing beyond the bounds of the familiar,” *AMIA Annual Symposium Proceedings*, vol. 2005, p. 151, 2005.

[8] M. Yang, S. C. P. Wong, and J. Coid, “The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools.,” Psychological Bulletin, vol. 136, no. 5, pp. 740–767, 2010.

[9] K. Glasgow and R. Schouten, “Assessing Violence Risk in Threatening Communications,” *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology*: *From Linguistic Signal to Clinical Reality*, pp. 38–45, 2014.

[10] A. Kutuzov, E. Velldal, and L. Øvrelid, “Temporal dynamics of semantic relations in word embeddings: an application to predicting armed conflict participants,” *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Jul. 2017.

[11] N. Venuti, H. Mcintyre, and P. J. Grazaitis et al., “Hyperparameter Optim*ization for Predicting the Tolerance Level of Religious Discourse,” Social, Cultural, and Behavioral Modeling Lecture Notes in Computer Science*, pp. 335–341, Jul. 2017.

[12] Y. Yang and J. Pedersen. “A comparative study on feature selection in text categorization”. In *International Conference on Machine Learning (ICML)*, 1997. [Online] Avaiable: http://courses.ischool.berkeley.edu/i256/f06/papers/yang97comparative.pdf

[13] C. Manning, P. Raghavan, and H. Schutze. “Introduction to Information Retrieval”, Cambridge University Press. 2008. [Online] Available: https://nlp.stanford.edu/IR-book/

[14] F. Pedregosa, G. Varoquaux, and A. Gramfort et al., “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research,* vol. 12, p. 2825-2830, 2011. [Online] Available: http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

[15] T. Chen and C. Guestrin, "XGBoost: A Scalabe Tree Boosting System," *CoRR,* vol. 1603.02754, 2016. [Online] Available: http://arxiv.org/abs/1603.02754.

AUTHOR INFORMATION

Ben Greenawald, M.S. Student, Data Science Institute, University of Virginia.

Greg Wert, M.S. Student, Data Science Institute, University of Virginia.

Elaine Liu, M.S. Student, Data Science Institute, University of Virginia.

Mohammad al Boni, Research Associate, System and Information Engineering, University of Virginia.

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