Speech Analysis to Predict Group Violence in Value-Based Groups

Donald E. Brown, brown@virginia.edu, Advisor
Mohammad al Boni, ma2sm@virginia.edu
Elaine Liu, yl9qr@virginia.edu
Gregory A. Wert, gaw8pa@virginia.edu
Benjamin Greenawald, bgh5dy@virginia.edu

Client and Sponsor: Army Research Laboratory (ARL)

Summary

In today's increasingly globalized society, many interactions occur between groups of different values. While many of these groups are peaceful, a concerning number of them have a tendency towards violence, with dire consequences. In 2014 alone, over 12000 deaths could be attributed to just two of these groups, Boko Haram and the Islamic State of Iraq and The Levant (ISIL) [2]. When our military or other humanitarian organizations interact with a new value-based group, it would be immensely helpful to know beforehand how inclined the given group is towards violence, but no system with the sufficient scale and accuracy required is in place. This project aims to provide that system, with language at its core. Given that language is at the center of all human interaction, it is natural to look towards language to provide hints towards the underlying capacity of the group for violence.

Analyzing language comes with an intrinsic set of problems. Human language is so complicated that traditional models are either ineffective at capturing its nuance, or require so much context-specific knowledge about the speech being analyzed that they do not generalize well. This project aims to tackle this issue by using deep learning models to analyze the highly dimensional data without making any assumptions about the underlying linguistic patterns. That is to say, the goal of this project is to develop a data pipeline implementing a neural net architecture that is able to perform the necessary classification in any language (given enough labeled data). Arabic will be the primary language used to show this concept since there are so many value groups for whom Arabic is a primary language. Urdu may also be tested should enough data be collected.

Linguistic Modeling Background

Previous work has been done on predictive linguistic modeling of religious texts. This work has taken various forms and investigated specific traits of language. One such form has been the analysis of performative features [4]. Work by Venuti et al. has sought to test the efficacy of these methods. Semantic and performative analyses were done on English language texts for the purposes of predicting linguistic rigidity (defined in the following paragraph) and tolerance levels [4]. Research in this area has, in general, found that while semantic analysis can show topic trends, it is a poor predictor; performative analysis, however, predicts accurately because it examines word usage for strong signals of intent [1]. Venuti et al.'s work supported this as the performative traits were more able to predict the linguistic rigidity than semantic ones.

Linguistic rigidity is a score assigned to text that aims to help quantify the role religious language plays in the given text. This scoring is crucial towards the understanding of religion since the flexibility of a language can explain religion's role in cultures and groups. This idea has been presented and argued by some religious scholars. They argue that the use of religious language, and how flexibly it is used, can encapsulate a religious identity and the understanding of this is crucial when attempting to build inter-group communication [1]. Our work may also have us using deep learning to predict linguistic flexibility scores. This builds on prior work which has worked to automate this process. Mechanically, this is a crucial task for a large-scale analysis of linguistic texts. This is because, traditionally, the rigidity scores are done by hand.

Assigning a score of 1 to 9 on the rigidity scale to a 500-750 word document takes approximately an hour when done manually [1] [4].

Further analysis has been done on the the exploration of performative traits for linguistic modeling. These efforts have attempted to use machine learning on the performative analysis. This process has involved applying various machine learning methods to religious texts for predictive analysis. Green et al.'s work on the matter showed the potential of this method, as their work was yielding accuracies above 90% [5]. The success of these efforts has warranted further exploration into the field. Careful consideration of the limitations of this work must also, however, be taken into account. The past research has used machine learning methods that are hinged on parameters, and parameter optimization [1]. Thus, the prior techniques cannot be applied towards other languages without having to recompute and discover parameters within those languages. Our hope is to automate these methods through the use of deep learning. Deep learning can provide a language-agnostic prediction and eliminate the problems of feature selection encountered in prior models.

Definitions and Terminology

A project such as this that deals with humans and human interaction requires a precise set of definition in order to approach the subjective subject matter in the most objective way possible. Below we lay out a definition of what constitutes a "value-based group", a further, what qualifications a group must fit in order to be considered violent.

(i) In the context of this project, a "value-based group" should satisfy the conditions outlined below:

- A "value-based group" *must*
  - Operate under a common name
  - Have a publicly available statement/set of values. These values generally reflect a worldview and historical narrative
- Further, a "value-based group" must do at least one of the following:
  - Have an identifiable primary motive outside of making outside of making profit (making profit in no way disqualifies a group from being value-based, but there must be a motive on top of profit.) Under this definition pure media organizations, for example, are not value-based groups, but media organizations associated with governments or political parties would be
  - Recommend/seek out changes and actions, likely political in nature, that further their set of values
  - Make evaluations/judgments of outer groups

(ii) For the definition of a "violent" group, we turn to the World Health Organization who define violence as "... the intentional use of physical force or power, threatened or actual, against oneself, another person, or against a group or community, which either results in or has a high likelihood of resulting in injury, [or] death..” [5]. Using this definition, we can define a violent group in the following way:

- A violent group *must* satisfy both of the following:
  - Have a person or persons identifying with the group commit an act which falls under the above definition of violence.
  - The group must claim responsibility for the act of violence.

During the initial stage of the project, classifying is done at a group level. This means that if at any point a group commits an act that classifies them as violent, all of their documents will be labeled violent. Similarly, a nonviolent group will have all of their documents labeled as such.

(iii) These definitions suffice for the initial stage of the project, but a desired extension of this project is to classify documents at a document level, as opposed to a group level. The idea behind this is that it is inaccurate to assume that the behavior of a group is static over time. In reality, peaceful groups turn violent and violent groups become peaceful, and we would like to label their writings accordingly. Regrettably, there is no simple way to identify violence over time like this, thus any definition we present will be somewhat arbitrary. That being said, we present the following definition for violence at a document level.

- A document is classified as violent if it satisfies at least one of the following:
  - The group that published the document satisfied the criterion of a violent group at any time in the two years prior to the documents publishing
  - The group that published the document satisfied the criterion of a violent group at any time in the six months after the documents publishing

The selection of the two-year prior time frame comes from the Department of State which reevaluates the groups it lists as terrorist organizations every two years [3].

Objectives

We aim to achieve the following objectives:

1. Deliver a data pipeline including a neural net architecture that can be used to automate the classification of the documents in any language to assess whether certain religious groups are violent or non-violent with high accuracy. The proof of concept language for this project is Arabic, and analysis may be done in Urdu, should Arabic prove successful.
2. Present sufficient evidence showing our methodology outperforms the traditional machine learning methodologies for text analysis in terms of predicted accuracy
3. Gain deep understanding of deep learning methodologies and how the methodologies may be applied to text prediction in a novel way
4. (Desired) Show that this methodology is also effective in predicting the linguistic rigidity score for a set of documents in any language, with Arabic and Urdu being the proof of concept languages
5. (Desired) Show that this methodology can capture changes in groups that change from violent to non-violent (or vice versa) over time

Data Source

The data for this work will consist of Arabic and Urdu work transcripts of writings and speeches by value-based groups. The texts will come from both violent and nonviolent organizations. Nonviolent Arabic texts will be drawn from a wide variety of sources including Al Jazeera, other Arabic news sources, Friday sermons of various origins, and others. Arabic texts associated with violence will be drawn from various groups classified by the state department as terrorist groups. This includes Hamas, Hezbollah, ISIL, and others. Furthermore, primary source databases such as azsecure-data.org and jihadology.net will be used in the further acquiring of Arabic texts associated with violence. Urdu texts will be acquired through the UVA Center for Religion, Politics, and Conflict (RPC) of the religious studies department.

Software

The principal language used for the development of this project is Python 3, specifically the Anaconda distribution. We will utilize the virtual environment capabilities of Anaconda Python to ensure that our code is modular and that all of our dependencies are effectively managed. The *newspaper* package for Python will be used as the primary tool for web-scraping large websites, such as those used by news organizations. For text analysis and machine learning tasks, we will use the standard collection of data analysis packages in Python including *numpy, pandas,* and *sklearn.* The heftiest portion of this project involved deep learning, for which we will use the *keras* Python package, which is a wrapper package on top of *TensorFlow* or *Theano.* Visualization of results will primarily use the *matplotlib* package for Python, though it is possible that the *ggplot2* library in R will be used for some visualization tasks.

Data Pre-Processing

The texts will be brought in through a mixture of web scraping and file reading. Web scrapers packages in Python such as *newspaper* and *BeautifulSoup* will be used in this process. Along with the primary text, the date of publication should also be obtained if it is available. This is so that document level labeling can potentially be performed. From there, the raw texts will be split into roughly 1000 word documents. The reason for imposing a rough bound on the number of words in a single document is that documents with an extremely large amount of words can clog up the neural net, so applying a bound decreases computational time. The average word length of documents in roughly 1000, which is why this number was chosen as an approximate bound. An obvious issue surfaces as to what happens if a document is split in such a way that the meaning of the document is corrupted. To mitigate this problem, the documents will be scanned until a 1000 word chunk is identified at which point the document will continue to be scanned until a suitable breaking point is found (end of sentence/paragraph). The documents will be cleaned by removing features such as stop words. These documents will then be labeled with whether they are associated with violence.

Procedure and Analysis

After preprocessing, we proceed with the actual generation of models. The scope of this project will largely be restricted to different kinds of deep learning architectures. The primary reason of using deep learning architectures is to ensure that we are able to generate accurate predictive models while remaining agnostic to the language that the model is being trained on. A disadvantage of traditional machine learning methodologies for text analysis is that features must be extracted from the text in advance in order for the models to be used. Extracting features from text requires specific knowledge of the structure and syntax of the language. This restricts these models to whatever language the models were trained on.

Our desire to create a language-agnostic model for text classification necessitates the use of a model that can do automatic feature selection. Thus the model we will focus most heavily on is the Convolutional Neural Network (CNN). CNNs gained prominence in the field of image recognition for their ability to self-extract features from images. Similar approaches can be used to extract features from text. The approach is to simply slide a window of size *h* over the text and those become your features. For example, if *h* = 2, then every consecutive pair of words would become features. This operation of sliding a window over the text is the “convolution” in the “convolutional neural network. This architecture supports using multiple different window sizes simultaneously, and a pooling layer at the end extracts out only the most important features from each window size [8].

A similar architecture which will be explored is the Recurrent Convolutional Neural Net (RCNN), as proposed by Lai et al [6]. A major limitation of the CNN described above is that it depends on the window size or sizes selected. Choosing window sizes too large may neglect the contribution of individual or small groups of words. Further, large window sizes can lead to a large parameter space. On the other hand, selecting window sizes too small will miss critical information. Often, the main idea of a piece of text can span sentences or even paragraphs, and small window sizes miss that completely. They combine the CNN described in the previous paragraph with a recurrent neural network (RNN). RNNs analyze text word by word, storing all information about prior words in the hidden layers, meaning they are quite good at understanding the semantics of a piece of text. However, by themselves, RNNs are very biased since later words are favored more heavily. To address this, Lai et al propose the RCNN. First, a bidirectional recurrent structure is applied. This is extremely important for our use case, as some languages are read right to left, and others left to right. The bidirectional recurrent structure should accommodate both of these cases. Then, like in the CNN, a max pooling layer will be applied to consolidate the large feature space into a set of only the most important features. The RCNN, which uses elements from both the RNN and CNN, while addressing the weaknesses of both, could prove highly effective for our use case [6]. In both cases, the models will not output labels, but instead will output probabilities which will then be converted to labels based on thresholds which we will be able to tweak.

Our methodological approach from here follows the standard approach for machine learning classification tasks. The available data will be split into a stratified training and validation set. Because neural networks require a large amount of training data, the train/test split will favor a larger training set. An 80/20 split for train/test is a standard split when it is imperative to have the largest train set possible. cross-validation will be explored as an option, but because of the large computational load of training these networks, cross-validation may prove to be unviable, especially leave-one-out cross-validation.

Another interesting type of validation presents itself when we consider how we anticipate this model to be used. The actual way we intend this model to be applied is, given documents from some unseen group, can the model accurately predict whether or not that group is violent. Therefore, another potential cross-validation approach that will be tested (computational load permitting) is a modification on leave-one-out cross-validation, where instead of omitting a single document, we omit a single group (leave-one-group-out cross-validation). We will take all the documents for a given group and leave them out of training, and validate the model using them. Ideally, this is repeated for each group. This sort of validation could work very well for our problem because it more realistically mimics the way this model is intended to be applied.

We need now explicitly define the metrics that will be used to evaluate our models. While raw accuracy will be used (total number of correct guesses/total number of guesses), it can be misleading by itself, especially if there is an imbalance in the distribution of the response. For example, if 90% of the train data was violent texts and only 10% was nonviolent texts, then a model that only guessed violent text would have a 90% accuracy rate. To augment the information provided by the raw accuracy, we will also use the F1 measure, a relative standard in classification problems. The F1 measures is a metric for accuracy that considers both the recall (total number of true predicted positives/total number of predicted positives) and precision (total number of predicted positives/total number of positives). The F1 measure also works for analyzing the negative class. For this, the measure becomes an average of the total number of true predicted negatives/total number of predicted negatives and the total number of predicted negatives/total number of negatives.

On top of the metrics defined above, we need to show that our methods are superior to the standards in text classification. For the standard text classification approach, we will extract the bag of words or n-gram model to represent the texts and will use the TF-IDF weights for our feature space. From there, we will use a logistic regression algorithm as our baseline. We may also explore feeding these features into an artificial neural network and see how our models compare with another neural network approach. Like our models, these models will output probabilities which will be mapped to labels based on thresholds that we can tweak.

A desirable bit of analysis would be to see if our model is also able to predict the linguistic rigidity of documents. The setup would be exactly the same as previously described, except now the outcome would be multilevel instead of just binomial. This is contingent on getting enough hand-labeled data since neural nets require a great deal of data. On the same vein, we would like to explore labeling the documents at the document level. This would require labeling the documents as violent or nonviolent based on whether the group was violent around the time the document was published (see **"Definitions and Terminology"** section (iii) for a formal procedure for labeling documents in this way). This would be interesting because if our model is able to track the changes in violent tendencies of a single group over time, then we have a strong case that our model is effective. It is one thing to parse out violence versus nonviolence in different groups, but it takes an extremely good model to detect changes within a single group. This analysis is also contingent on getting enough documents who have a publication date available.

Project Deliverables

1. **The minimum viable product (MVP) for predicting religious group violence.** The primary deliverable will be an MVP that utilizes neural net architecture to predict whether certain value-based groups are violent or not with acceptable accuracy.

2. **Repository of texts in Arabic.** The second deliverable will be a repository of pre-processed texts in Arabic from both nonviolent and violent groups. These texts will be labeled as either violent or nonviolent. If available, these texts will also have a date of publication. For each group, there will be at least 1,500 documents. In total, the repository will include at least 3,000 documents.

3. **Model for predicting linguistic rigidity (Desired).** This would entail using the model generative in objective (1) to predict the linguistic rigidity of a set of documents. This deliverable is contingent on getting enough documents, both violent and nonviolent, labeled with linguistic rigidity scores. This analysis could be done in Urdu or Arabic, but Urdu is preferred.

4. **Repository of texts in Urdu (Desired).** Similar to objective (2), this repository would contain roughly 3,000 labeled documents, roughly 1,500 from violent groups, and 1,500 from nonviolent groups. The documents will also contain the date of publication if available. The primary obstacle to this deliverable is finding enough scrapable online sources in Urdu.

5. **Document-level analysis (Desired).** This would entail labeling the documents on a per document basis, as opposed to assigning all documents by a group with the same label. The labeling at the document level will be carried out using the procedure defined in part (iii) of the **"Definitions and Terminology"** section. The deliverable is dependent on finding enough documents that contain their publication date.

6. **Summary of findings.** A brief summary of our findings on our model's performance on both religious group violence prediction and linguistic rigidity prediction. We will present the neural net architecture, cross-validation results, and any limitations of our approach. We will also present how our models perform under temporal events.

7. **Report.** The final deliverable will be a formal scientific documentation drafted in LaTeX. In this report, we will present the background and introduction to our research, literature review, data description, data pre-processing procedures, description of the neural network architecture and our approach, a summary of our results with graphical illustrations, and conclusions.

8. **Presentation.** The last deliverable would be a concise version of our report in a presentation format.

Schedule

We started the project in the week of September 18, 2017, and will end on April 23, 2018. We divided our project into 6 phases. The 6 phases are (overlapping) subdivisions of the project each focusing on a specific area.

- The first phase will start in the week of September 18, 2017, and end in the week of January 22, 2018, with the completion of the primary deliverable. This will be the minimum viable product, a model which predicts violent or nonviolent based on documents labeled at the group level.
- The second phase will start in the week of October 23, 2017, and end in the week of February 5th, 2018. This phase will focus on applying our model to documents labeled with linguistic rigidity, assuming enough labeled data is collected. This phase will also shift the language of focus to Urdu, also pending the collection of enough data. This phase will end with a minimum viable model which can predict linguistic rigidity based on documents in Urdu.
- The third phase will start in the week of February 5th, 2018, and end in the week of March 12, 2018. This phase will involve frequent discussions with the client to modifying and refine existing procedures.
- The fourth phase will start in the week of February 5th, 2018, and end in the week of April 2, 2018. This is a relatively open phase. Its purpose is to continue to explore with our models, potentially putting in more time to our desired deliverables, and ends with a summary of our findings to the client.
- The fifth phase will start in the week of April 2, 2018, and end in the week of April 16, 2018, with the completion of the research report.
- The sixth phase will start in the week of April 16, 2018, and end in the week of April 23, 2018, with the completion of the presentation.

Please refer to the Appendix A for a detailed project schedule.

Budget

SIEDS Registration Fees: $550. This includes the $75 registration fee for the four team members as well as symposium fees for the faculty advisor ($100 for IEEE member) and client representative ($150).
AWS: $5000 for infrastructure and resources needed to train and test the neural networks.

References

[1] D. E. Brown, H. Mcintyre, P. J. Grazaitis, R. M. Hazell, and N. Venuti, “Hyperparameter Optimization for Predicting the Tolerance Level of Religious Discourse,” *Social, Cultural, and Behavioral Modeling Lecture Notes in Computer Science,* pp. 335–341, 2017.

[2] D. Searcey and M. Santora, “Boko Haram Tops ISIS in Ranking Terror Groups,” *The New York Times*, para. 2, Nov. 14, 2015 [Online]. Available: https://www.nytimes.com/2015/11/19/world/africa/boko-haram-ranked-ahead-of-isis-for-deadliest-terror-group.html. [Accessed Sept. 24, 2017].

[3] “Foreign Terrorist Organizations,” U.S. Department of State. [Online]. Available: https://www.state.gov/j/ct/rls/other/des/123085.htm. [Accessed: 26-Sep-2017].

[4] N. Venuti, B. Sachtjen, H. Mcintyre, C. Mishra, M. Hays, and D. E. Brown, “Predicting the tolerance level of religious discourse through computational linguistics,” in *2016 IEEE Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, USA, 2016*, S. Donohue, 2016.

[5] S. Green, M. Stiles, K. Harton, S. Garofalo, and D. E. Brown, “Computational analysis of religious and ideological linguistic behavior,” in *2017 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, USA, 2017,* R. MacDonald, 2017.

[6] S. Lai et al, “Recurrent Convolutional Neural Networks for Text Classification,” in the *Proceedings of the Twenty-Ninth Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence,* Austin, TX, 2015, pages 2267-2273.

[7] “Violence,” WHO. [Online]. Available: http://www.who.int/topics/violence/en/. [Accessed: 26-Sep-2017].

[8] Y. Kim, “Convolutional Neural Networks for Sentence Classification,” in the *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, 2014, pages 1746–1751.