Generating a Dataset of Speeches tagged with Persuasion Techniques

Ankit Mathur, University of California, Berkeley
Ana-Maria Popescu, University of Washington
David Bamman, University of Berkeley, California

September 22, 2017

1 Introduction

Politics has always been a game of persuasion. Throughout the course of time, politics has been about connecting with people about issues that matter to them to get their support. Political speeches have been a big part of this, developing into a subtle art of wordplay, leveraging techniques that often play on elements of the human psyche. Certain arguments and stories provide the subtext of ethos that a politician wants to convey to sway voters to them.

This is often considered an imprecise issue - what might be something that evokes emotion to one person might be completely neutral for another. We argue that political speeches are actual quite consistent in their usage of specific persuasion techniques. In this work, we categorize several types of persuasion techniques, and we claim that reasonable rates of agreement can be reached among independent annotators of political speeches, given these categories and their definitions. Building such a dataset is the first step to conceiving of a machine learning based solution to teaching computers to understand second order concepts in language.

We first address the question of how to categorize political persuasion techniques, leaning on former work throughout the political science community. Then, we address the process by which we acquired and annotated the data, focusing on the universality of these techniques across political speeches through time. We discuss the method we use to evaluate the similarity of annotations in this context, where annotations that may be semantically different but still be touching at the same "area" of the sentence should be marked as similar. To properly evaluate this, we need a definition of similarity that is more relaxed. Finally, we present the results for the similarity of our annotations across multiple annotators and categories.

2 Related Work

Work on analysis of political speeches is quite expansive, and, with political analysis being well-discussed topics, several publications have offered frameworks for understanding the kinds of techniques that are used to influence opinion.

Fundamentally, one class of these techniques are appealing to the basic emotions of people. These kinds of persuasion techniques are often categorized as motivational appeals [6, 8]. Using fear as a motivational appeal is mentioned consistently throughout the literature in the context of political ads and speeches [6, 3]. Playing on the secular emotion of patriotism helps to leverage the nationalistic tendencies imposed by countries and political systems [6]. On the other hand, humor is a motivational appeal reflective of charisma and relatability [6]. In addition, several works hit at the concept of warmth conveyed through story, often categorized as "emotional" appeals. Emotional anecdotes are an umbrella term used to describe the kinds of stories that invoke emotions by playing on these emotions [6, 3].

Order within the text of the speech is also common in an attempt to persuade. The usage of repetition can be used to drive home important points [6]. This usage of order extends to incredibly complex techniques as well, using the placement of stronger and weaker arguments to influence overall argument construction.

Another level of techniques involves usage of words and phrases that associate feelings of a certain kind with a concept. Literature on propoganda discusses techniques used to associate positive or negative feeling through the usage

of phrases and words. Name calling and stereotyping are simple examples used for this purpose. Using virtue words (referenced as glittering generalities) to associate general words that are broadly positive with concepts is highlighted by several works as a common persuasion strategy [10, 5]. Associations go beyond emotions, as associating to rural roots is referenced as using a plain folks approach, and associating with well-known public officials as testimonial [10, 5]. The most critical association for a politician is often associating policy action with themselves, categorized by many as credit claiming [9, 4].

Attempts to tag speeches with specific persuasion techniques has been somewhat rare. Disagreement agrees across works in terms of what techniques are consistent perusasion categories to tag speeches with. In a related vein, work has been done to tag political speeches with audience reactions - with the assertion that audience reactions correspond to moments in the speech where there has been a hot spot of persuasive communication [7]. Further attempts have been made to tag political speeches in specific persuasion contexts - in particular, in the context of branding techniques, tags have been made for value-based techniques in the context of political speeches [2]. In total, no dataset exists that tags political speeches with an agreed upon set of persuasion techniques.

Categories

In order to generate a dataset with such annotations, we need a codebook with consistent definitions that are proven to be reproducible. We propose the following codebook of definitions for persuasion techniques:
In general, when classifying, highlight the entire sentence.

- Name Calling - Giving a bad name to "individuals, groups, nations, races, practices, beliefs and ideals he would have us condemn" to "make us form a judgment without examining the evidence". When looking for this, just look for cases where candidates call the candidates from opposing parties (or same party) out by name and use a negative adjective to describe them.

- Glittering generalities - Identifying one's message with virtue by appealing to good emotions. Virtue words include "truth", "freedom", "honor", "liberty", "public service", "the American way", etc. These words that are listed are good tags to look for, and, in addition to that, anything that talks in glittering language about the American promise or dream also qualifies as a glittering generality. These terms should be identified at a sentence level (if a sentence contains glittering generalities, tag the whole sentence).

- Testimonial - Using testimonials from trusted figures (celebrities, experts, people similar to those in audience) in support of one's message. Generally, one sees this in parts of the speech where the person is thanking the local candidate from the area, in an effort to tie the candidate's image to something that people from that area are used to/have heard of before. There are other instances, such as when candidates cite specific endorsements. This includes candidates talking about endorsements received as well as speakers endorsing other candidates.

- Plain Folks - Trying to win confidence by appearing to be like an average person (discussing "common things of life" - family, shared passtimes, experiences, etc.). In general, this falls under attempts to be folksy. This is when candidates describe how they grew up in very simplistic households/middle class working families to create a relatable feel to their image. All sentences from a story that reminds you of the plain folks tag featuring the specific plain-folks style language should show up (i.e. all *relevant* sentences should be tagged). This can show in many different contexts: speakers can be referencing other individuals' stories or they could be making calls to stories that reference plain folks ideals in the abstract.

- Credit Claiming - claiming responsibility for getting a specific outcome (e.g. getting a bill passed, etc.). The scope of this can be large - the speech can be claiming credit for a policy or it can be claiming credit for some kind of rejection of the other party's platform. The point of this is the candidate is connecting some *result* to their actions. This content can be referencing a different candidate performing some action and assigning credit to that person in a positive way. Promises of action should not be included.

- Stereotyping - Conventional notion of an individual, group of people, country, etc. as held by a number of people. Stereotypes can be negative or positive. Stereotypical content can be explicit or implicit (in the form of "cues" - coded language, evocative visuals, etc.) Stereotypes can be positive as well. Examples of what might

be misclassified as stereotyping:

- Patriotism - reference to patriotic appeal to appeal to the crowd ("flag waving") that involves mentions of patriotic emblems (flag, the veterans, eagle, etc.), how great the nation is, etc. In general, this also includes references to historical greatness, shared heritage, as well as references to national improvement / restoring greatness. This can sometimes be confused with glittering generalities. Feel free to mark both labels if it is both.
- Repetition - repeating arguments with the same message. Select the repeated phrase, *along with its repitition*. For example, if the repeated phrase is yes we can. Highlight "yes we can. yes we can.", not just the first "yes we can". Examples that are not exact repetition but are still using the same technique more loosely should be tagged as repetition.
- Fear - usage of fear to motivate audience to support either a political strategy or a candidate in general. When categorizing this, look for threats, discussion about current or future impacts (poverty, failure, unemployment). When annotating examples of fear, highlight the whole sentence (or full clause) - not just the fragment of the fear portion that you've identified.
- Emotional Anecdotes - telling stories while leveraging value-based language to appeal to the emotions of voters to convince them of a candidate's humanity. It is sometimes difficult to determine exactly how much to highlight here. Identify stories, and highlight up to 3 of the sentences most core to the story being told (or if it's a very long anecdote, break into multiple annotations). Note that emotional anecdotes do not have to be about a specific person - they can cite generic character tropes as well.

3 Data Acquisition and Annotation

The process of acquiring the data involved scraping several websites for political speeches. We wanted a corpus of speeches that would be generalizable to more than just the current era of political speaking. Using sources such as whatthefolly and AmericanRhetoric as sources, we downloaded top speeches from the past century. Naturally, record-keeping of older sources of political speeches are sparse, and so our collection of 41 speeches was carefully chosen to select for speeches that were likely to have persuasive content within them. These speeches range from modern political speeches by Donald Trump and Hillary Clinton and stretch back to speeches by FDR.

Certain documents that were scraped were naturally too small or too basic to be considered for annotation. These documents were removed from the corpus because annotating them would not add any value to a dataset that aimed to offer annotations of various persuasion techniques.

We began by defining and refining the categories that we described above and creating a stable policy for how to select each of the categories. We ran this definition by several sources in order to confirm that there was a stable definition for each category, We also aimed to include topical examples for each category.

The BRAT annotation was used to standardize the process of annotating speeches. We specified whether annotation should be at the sentence level or the phrase level to standardize the way annotation was done across categories.

We make two important claims here:

1. We contend the categories that we define are reproducible ways of annotating for persuasion techniques.
2. This can be proven by independent annotators annotating the same dataset and computing a similarity score based on this.

4 Evaluating Similarity

Evaluating similarity between annotations is a well-studied problem. Inter-annotator agreement rates in the context of linguistics were surveyed effectively in the past [1]. We will focus on the chance-corrected agreement rate coefficients.

Let A_O be the observed agreement. If we define A_e as the amount of agreement expected by chance, then $1 - A_e$ is how much agreement more than chance is possible.

$$S, \pi, \kappa = \frac{A_O - A_e}{1 - A_e}$$
$$A_e = \sum_{c \in C} P(c|a_1) \cdot P(c|a_2)$$

where C denotes all the categories and a_1 and a_2 are the annotators. All three metrics assume independence of the two annotators, which we also assume. Each of these leverages different assumptions about computing the probability a category is chosen by a coder. S assumes that if the annotators were operating by chance alone, we would get a uniform distribution. π assumes that we would get the same distribution for each annotator. κ makes no such assumptions about the distribution for each annotator - they may be different. κ also requires a set of observable priors for each of the categories.

In the context of our problem, we will use κ. It is important that we be able to include a prior, since we do not necessarily have the same number of annotations for every class. In addition, annotators are expected to be slightly different, so expecting the same distribution is an unnecessary constraint. Proving that the categories we propose are stable should mean we make no assumptions about the real distributions.

We use the following definition for the expected agreement given the priors.

$$A_e = \sum_{c \in C} P(c|a_1) \cdot P(c|a_2) = \sum_{c \in C} \frac{n_{a_1 c}}{n_{ann}} \frac{n_{a_2 c}}{n_{ann}} == \frac{1}{n_{ann}^2} \sum_{c \in C} n_{a_1 c} n_{a_2 c}$$

To include a prior in the computation of our similarity score, we include the percentage of annotations that are of a certain category. This is a relevant prior because the percentages are nontrivially different (with some categories being rarer used than others). As such, to compute the probability of random selection, we must include probabilities derived from how common annotations of a certain category were as the prior.

However, these annotation measures are insufficient because they assume a space of annotations where there is a clear boundary. In the case of persuasion techniques, annotations are often of somewhat varying length, but that does not mean that those annotations are not tagging the same thing. For example, if the sentence being annotated were “America is a great country where greatness thrives.”. A valid patriotism annotation would be “America is a great country where greatness thrives”, but the annotation “America is a great country” has a strong overlap with the previous annotation. This is not the kind of disagreement that is relevant in identifying a stable definition for persuasion techniques, since the difference is purely semantic.

To resolve this problem, we used a window-based agreement check. We start by defining a reasonable constant k which is the minimum number of words that must overlap to mark two annotations as agreeing. Then, as we compute the score of similarity, rather than checking for string equality, we use this metric to compute whether the annotations should be included in the set of agreements.

5 Results

Of our two annotators, annotator 1 had 1409 annotations and annotator 2 had 1263 annotations. We had the following priors for the categories that were annotated in our shared dataset of speeches:

Name Calling	0.09505988023952096
Glittering Generalities	0.2638473053892216
Testimonial	0.0815868263473054
Plain Folks	0.053517964071856286
Credit Claiming	0.1534431137724551
Stereotyping	0.016092814371257484
Slogans	0.016467065868263474
Patriotism	0.08907185628742514
Repetition	0.020958083832335328
Fear	0.14895209580838323
Emotional Anecdotes	0.06100299401197605

The computed similarity scores of these annotations was $\kappa = 0.7345$.

This similarity score is high, and it demonstrates that the category definitions that were used for this annotation exercise were stable. This work constitutes a quantitative proof that our categories are a stable, learnable, and predictable definition for persuasion techniques. In addition to this, the dataset generated by this work has been shown to have labels that are reproducible and replicable.

6 References

References

[1] Ron Artstein and Massimo Poesio. "Inter-coder Agreement for Computational Linguistics". In: *Comput. Linguist.* 34.4 (Dec. 2008), pp. 555–596. ISSN: 0891-2017. DOI: 10.1162/coli.07-034-R2. URL: http://dx.doi.org/10.1162/coli.07-034-R2.

[2] Richard Barberio and Brian Lowe. "Branding: Presidential politics and crafted political communications". In: *Annual Meeting of the American Political Science Association* (2006).

[3] Ted Brader. "Striking a Responsive Chord: How Political Ads Motivate and Persuade Voters by Appealing to Emotions". In: *The American Journal of Political Science* (2005).

[4] Jamie L. Carson and Jeffrey A. Jenkins. "Examining the Electoral Connection across Time". In: *Annual Review of Political Science, Vol. 14, pp. 25-46, 2011* (2011).

[5] Mike Conway, Maria Elizabeth Grabe, and Kevin Grieves. "VILLAINS, VICTIMS AND THE VIRTUOUS IN BILL O'REILLY'S "NO-SPIN ZONE", Revisiting world war propaganda techniques". In: *Journalism Studies, Vol. 8, No 2* (2007).

[6] Robert H. Gass and John S Selter. "Persuasion: Social Influence and Compliance Gaining". In: (2007).

[7] Marco Guerini et al. "The New Release of CORPS: A Corpus of Political Speeches Annotatated with Audience Reactions". In: *LNAI* (2010).

[8] Lynda Lee Kaid. "Ethics and Political Advertising". In: *Political Communication Ethics: An Oxymoron?* (2000).

[9] David R. Mayhew. "Congress: The Electoral Connection". In: (1974).

[10] Clyde Miller. "How to Detect Propaganda". In: (1938).