Latent Knowledge in Online Political Communities

Xavier Gutierrez

September 12, 2016

Abstract

The quantification of political ideology has traditionally been constrained to a one-dimensional continuum ranging from liberal to conservative, with network analyses accordingly restricted to the view that nodes exhibit one of two states. Using the Vector Space Model (VSM) as well as Latent Dirichlet Allocation (LDA) topic modeling on the Stanford Encyclopedia of Philosophy (SEP), I begin the inquiry of localizing users of Reddit.com in a high-dimensional space defined by concepts. By measuring the similarity between topic distributions of documents, one may establish a second VSM whereby documents are mapped onto a space implicitly defined by the SEP. I use Jurgen Habermas's model of deliberative democracy to outline future applications of this methodological framework.

1 Introduction

The concept of ideology has been approached by many different academic fields, including philosophers, psychologists, and political scientists. Interest in the subject being originated from a plurality of perspectives, there are many formal methodologies for studying it. Each methodology reflects a specific perspective on either the origin or phenomenology of ideology. These perspectives concern primarily questions of representation. Some researchers use non-spatial approaches which emphasize "thick descriptions of ideological content" as well as "symbolic/conceptual maps [4]. Others use spatial dimensions along which ideologies may be located to indicate emphasis on certain values, mirrored in the model's number of dimensions. While the traditional Left-Right (Liberal-Conservative) spectrum is the most widely used in political science, psychologists are able to use multidimensional scaling to discern the axes which organize the responses to survey questions. However, these statistically-generated axes may or may not correspond with psychological reality. A popular two-dimensional representation of ideology posits social tolerance, on the one hand, and attitudes to economic oversight on the other. Further permutations of this spatial model are not common. Another perspective to take highlights individual versus contextual accounts of ideology. When seen as being generated by a "spectrum of psychological needs basic to individuals", researchers focus on how social cognition and morality play into one's personality. Alternatively, social accounts of ideologies focus on historical and institutional factors in order

to "focus on the interactive dynamics by which ideologies are created, altered, and propagated". These accounts posit that ideology is made manifest primarily by language. The methodologies which these various perspectives generate can be classified by three categories: conceptual, discursive, and quantitative. No one academic field holds monopoly over any methodology, indicating that a full account of ideology requires a unification of these three methods. It is the unification of these methodologies which this paper addresses.

2 Methods

2.1 Vector Space Model (VSM)

Research in the field of information retrieval and language processing concerns methods of storing, organizing, and searching through bodies of documents. The fundamental issue is to identify an informatic representation of documents, e.g. a set of indices T_j. These indices can be thought of as the axes of this space. I wish to highlight the fact that the choice of indices ultimately dictates the document space's resultant configuration, for this will be important in the context of Topic Modeling (Section 2.2) as well as establishing the nature of my object of study (Section 3).

For example, one can imagine representing each document by a set of n specific words. On the one hand, by choosing vague words which won't significantly distinguish separate documents – such as "the", "and", "she", "they", etc. – as a result of each document using every index term, these documents' representations in a VSM will be grouped in one cluster. Alternatively, one can choose index terms pertinent to semi-distinct subsets of documents such that their representation in the VSM will also be (at least) semi-distinct. The words "genetics", "evolution", "disease", for instance, would distinguish articles pertaining to different sub-fields of biology in a set of scientific articles [1]

Furthermore, this model allows for various choices of vector-based similarity/distance measures, which provide information about the spatial distribution of documents. Documents with a high similarity measure will be represented by vectors that are close spatially, and vice versa [5]. In Section 3.3, I will discuss the Jensen-Shannon Distance (JSD), which evaluates the distance between two probability distributions. In Section 2.2, I will discuss how documents come to be represented by a set of vectors D_i where D_i is a discrete probability distribution, as well as how the JSD may be used to run a hierarchical clustering algorithm on a pairwise distance matrix calculated from D_i.

2.2 Topic Modeling

It is useful to clarify the assumptions of topic modeling, a broad class of machine learning algorithms, before describing the details of the specific model I used in my study of ideology.

First, topic modeling relies on a "bag-of-words" representation of documents, meaning that word order can be ignored. In the language of VSM, each document is represented as a vector $D_i = t_1, t_2, \ldots, t_n$ where (1) n is the total number of unique words in the set of documents, and (2) t_j is the probability of observing the jth word, i.e. (number of occurences of word j)/(total number

of words in document i). Second, topic models use "latent variables which aim to capture abstract notions such as topics" [2]. This means that T_j, the VSM index, will not refer to words; rather, it represents the set of topics produced by the model. I remind the reader that the set of documents on which the topic model is trained establishes a document space, and the VSM index – topics latent in these documents – define the configuration of this document space.

Latent Dirichlet Allocation (LDA) is a generative model that uses a three-tiered representation: document-topics-words. The inference and parameter estimation algorithms, whose statistical details extend beyond the interests of this paper, seek to find an optimal representation of documents as a distribution over topics, as well as of topics as a distribution over words. This model is generative in the sense that the topics "discovered" by the algorithm are assumed to have been sampled in the creation of the documents, and can be used to recreate a document sufficiently approximate to the original "bag-of-words" that the model was trained on. For this reason, I will hereon refer to the set of documents (or corpus) which the LDA model was trained on as the reference corpus.

The utility of such statistically-generated topics is determined by its interpretability. While a generative model finds the "best set of latent variables that can explain the observed data... there is no notion of mutual exclusivity" [6].

This model's ultimate purpose is to make a query with a set of unseen documents. The output of this operation is a representation of the unseen documents in the space established by the reference corpus, the basis of this representation being the set of topics T_j. For this reason, I will hereon refer to the set of unseen documents (or corpus) as the query corpus.

3 Social and Conceptual Systems

In this section, I will describe the systems that produced this paper's reference and query corpora and whose relationship is my object of study. In Section 5 I propose that, by applying these methods to two sets of political discourse, one may track the "deliberative legitimation processes" implemented by a "division of labor" which organizes the flow of political actors' published opinions through the public sphere [3].

3.1 Stanford Encyclopedia of Philosophy (SEP)

I have used a set of 1,098 articles originating from this online encyclopedia as my reference corpus. By training my LDA model on this corpus and using the resultant topics T_j as a VSM index, my document space is configured such that it optimally represents these documents. One can then see that the reference and query corpora have an asymmetric relationship, as the fitting algorithm of this model produced topics meant to distinguish the (semi-)distinct sub-sets of documents in the reference corpus but not the query corpus. This is not to say that these topics are irrelevant to, or that we get an inaccurate representation of, the query corpus. Rather, we are to understand the representation of the query corpus *in the context of the reference corpus*. For instance, by identifying the potential political aims of this encyclopedia (perhaps by an analysis of the resultant topics), we can then discern the effective political influence of this

source on the system which generated the query corpus. One can then imagine using the textual output of various political institutions as reference corpora, on which the application of this methodological framework could then indicate (1) the political aims of these institutions, as well as (2) the influence of these institutions on another "specialized deliberative arena" [3] of the public sphere.

As described above, the topics are produced to satisfy statistical, not semantic, requirements. Ideally, however, these topics will have a clear analogue to some well-established philosophical concept from the encyclopedia.

3.2 r/politics

The founders of Reddit.com identify the function of their social media site as follows: to "bridge[] communities and individuals with ideas, the latest digital trends, and breaking news". Users of this site choose which communities (individual forums known as subreddits and identified with the prefix r/) they want to subscribe to and may participate freely in any particular conversation (or post) by attaching their comments, to which other users my respond. Each subreddit is monitored by a set of particular users, known as moderators, who either were among the users who created the subreddit or applied and were accepted to be a moderator. Moderators have the ability to censor posts or comments that do not comply with the participation rules and submission guidelines of the subreddit, and eventually may ban users who consistently break them.

The r/politics subreddit is meant for discussion of U.S. political news, and as of September 2016 has 3.1 million subscribers. My query corpus is composed of the individual comments left by about 161,000 distinct users during the year of 2014.

3.3 Spatial Model of Reference Corpus' Discourse

As mentioned in Section 2.1, the JSD can be used as a measure of distance between documents in this VSM. It is calculated by the equation JSD(a,b) = H_m - $(H_a + H_b)/2$, where H is the entropy function, a and b are discrete probability distributions over topics T_k, and m = (a+b)/2. The JSD is a number between 0 and 1.

By calculating JSD($Article_i$,$User_j$) for all i, one has now established a nested VSM, VSM_2, in which documents are expressed in terms of the Jensen-Shannon distance between a given $User_j$ (found in the query corpus) and each article in (which comprise the query corpus). It is necessary to note that establishing VSM_2, with index $Article_i$, requires VSM_1, with index $Topics_k$, as an intermediate step.

In place of representations in this space being located by the index T_k, where T_k are the set of topics generated by LDA , users of r/politics (or, generally, documents from any query corpus) are located by statistical similarity to the SEP articles (or documents from any reference corpus). Simply put, VSM_2 is a space in which documents in the query corpus are represented in terms of the entire reference corpus. While not wholly eschewing the uncertain coherence of LDA topics, the choice to represent a document in VSM_2 by calculating JSD(R_i, Q_j) for all i (where R_i and Q_j are the reference and query corpora, respectively), is meant to support the interpretation that placements in VSM_2 reflect the query corpus' engagement with each document of the reference corpus.

One can then imagine using transcripts from the speeches of various politicians as the reference corpus to study the degree to which users are represented by these politicians' public opinions. Other potential applications will be discussed in Section 6.

4 Deliberative Democracy

Deliberative democracy can refer to a host of theories developed by thinkers ranging from antiquity to modernity. Its cornerstone is that decision-making must be preceded by deliberation in order to be legitimate. Jurgen Habermas referred to the "truth-tracking potential" of political deliberation as an "epistemic dimension" of democracy, and asserts that this paradigm has been verified empirically for small groups. Habermas' work, however, is preoccupied with the contrast between the existing "power structure of the public sphere and the dynamics of mass communication" in media society, on the one hand, and the "normative requirements of deliberative politics", on the other. In outlining specific variables which could "explain failures in the maintenance" of these requirements, Habermas has laid the groundwork for fruitful applications of the methodological framework outlined in this paper.

Habermas' communicative model of deliberative democracy requires a "self-regulating media system and... proper feedback between public sphere and civil society. First, by self-regulating, Habermas specifically refers to "effective counterframing". Second, concerning the communication between public sphere and civil society as the very process of deliberation, to be legitimate is to meet the following criteria: (1) gather "relevant issues... required information... and [specific] interpretations, (2) provide "proper arguments for and against" these contributions, and (3) produce "rationally motivated *yes* and *no* attitudes [which] determine... procedurally correct decisions". Both these normative requirements can be tested using the methods outlined in this paper.

5 Future Directions

Using this union of the Vector Space Model and topic modeling, the task of posing questions becomes, in large part, that of choosing query and reference corpora. With respect to the particular corpora used in this paper, where the encyclopedia is meant to span a broad range of subjects and the r/politics data reflects the online behavior of users, it would be useful to train an LDA on r/politics and map the SEP onto the VSM_2 implicitly defined by individual users. Instead of revealing the engagement of users with the various articles of the SEP, such an endeavor would prove the relevance of the SEP to what users are talking about in response to U.S. political news.

Furthermore, there are a multitude of politically-oriented subreddits engaged explicitly with specific ideologies, reflecting identifications along the traditional Left-Right spectrum as well as more ambiguously-located philosophies (Anarcho-Capitalism, for instance). There are subreddits centered on specific issues as well (TPP, 4th Amendment, etc.).

Of particular relevance to Habermas' communicative model of deliberative democracy would be to map individuals onto a VSM_2 defined by media sources.

First, by simply looking at the JSD between documents in the reference corpora, one would test the extent of counter-framing across media sources. To address the feedback between the public sphere and civil society, one may train multiple LDA models where the reference corpora are aligned temporally, then query these models with similarly organized user data. Consequently, trajectories in either VSM_1 or VSM_2 could reflect the relationship between mass communication and public opinion.

6 Acknowledgments

This work was done in collaboration with Professor Simon DeDeo and Dr. Marion Dumas at the Santa Fe Institute. Additionally, I'd like to thank Professor Colin Allen at Indiana University for providing the text files for the Stanford Encyclopedia of Philosophy; Jason Baumgartner for collecting through their API and publicly releasing the Reddit data; and Will Hamilton at Stanford University for providing this data in a clean, accessible format.

References

[1] David M. Blei. Probabilistic Topic Models. *Commun. ACM*, 55(4):77–84, April 2012. ISSN 0001-0782. doi: 10.1145/2133806.2133826. URL http://doi.acm.org/10.1145/2133806.2133826.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=944919.944937.

[3] Jürgen Habermas. Political Communication in Media Society: Does Democracy Still Enjoy an Epistemic Dimension? The Impact of Normative Theory on Empirical Research1. *Communication Theory*, 16(4):411–426, November 2006. ISSN 1468-2885. doi: 10.1111/j.1468-2885.2006.00280.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1468-2885.2006.00280.x/abstract.

[4] Thomas Homer-Dixon, Jonathan Leader Maynard, Matto Mildenberger, Manjana Milkoreit, Steven J. Mock, Stephen Quilley, Tobias Schröder, and Paul Thagard. A Complex Systems Approach to the Study of Ideology: Cognitive-Affective Structures and the Dynamics of Belief Systems. *Journal of Social and Political Psychology*, 1(1):337–363, December 2013. ISSN 2195-3325. doi: 10.5964/jspp.v1i1.36. URL http://jspp.psychopen.eu/article/view/36.

[5] G. Salton, A. Wong, and C. S. Yang. A Vector Space Model for Automatic Indexing. *Commun. ACM*, 18(11):613–620, November 1975. ISSN 0001-0782. doi: 10.1145/361219.361220. URL http://doi.acm.org/10.1145/361219.361220.

[6] Mark Steyvers and Tom Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440, 2007.