UC Santa Cruz

UC Santa Cruz Electronic Theses and Dissertations

Title

Assessing Bias in Think Tanks

Permalink

https://escholarship.org/uc/item/3hr1w8t7

Author

Joshi, Mitalee

Publication Date

2020

Peer reviewed|Thesis/dissertation

eScholarship.org

Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

SANTA CRUZ

ASSESSING BIAS IN THINK TANKS

A thesis submitted in partial satisfaction of the requirements for the degree of

MASTER OF SCIENCE

in

APPLIED ECONOMICS AND FINANCE

by

Mitalee Joshi

December 2020

The thesis of Mitalee Joshi is approved:

Professor Ajay Shenoy

Professor Laura Giuliano

Professor Alan C. Spearot

Quentin Williams

Acting Vice Provost and Dean of Graduate Studies

(This page is intentionally left blank.)

TABLE OF CONTENTS

ABSTRACT vii

ACKNOWLEDGEMENT ix

INTRODUCTION 1

THINK TANKS DATA 3

MODEL 7

Clustering articles based on issue areas 7

Classifying articles based on the authoring think tank 10

ESTIMATION 11

Maximum-a-Posteriori estimator 12

Penalized Maximum Likelihood estimator 14

Regularized non-linear estimator (Support Vector Classifier) 16

Alternative methods – Decision Trees 19

VALIDATION 19

Hyperparameter optimization using cross validation 20

Model training and validation 22

RESULTS 26

Analysis 1. One-vs-all the other think tanks in each issue area 27

Analysis 2. Analyzing feature sets by political orientation in each issue area 34

CONCLUSION 39

FUTURE WORK 40

APPENDIX 41

REFERENCES 43

LIST OF FIGURES

Fig. 1 Distribution of articles across the nine issue areas 23

Fig. 2 Comparison of training and out-of-sample validation score 24

Fig. 3 Overlap of phrases of one-vs-all think tanks for ‘immigration’ 27

Fig. 4 Overlap of phrases of one-vs-all think tanks for ‘national security’ 28

Fig. 5 Overlap of phrases of one-vs-all think tanks for ‘judiciary’... 29

Fig. 6 Overlap of phrases of one-vs-all think tanks for ‘politics’ 30

Fig. 7 Overlap of phrases of one-vs-all think tanks for 'social policy’ 31

Fig. 8 Overlap of phrases of one-vs-all think tanks for 'monetary policy' 32

Fig. 9 Overlap of phrases of one-vs-all think tanks for 'trade policy' 33

Fig. 10 Overlap between phrases based on political orientation of think tanks using Naive Bayes classifier 36

Fig. 11 Overlap between phrases based on political orientation of think tanks using penalized multinomial logit.. 37

Fig. 12 Overlap between phrases based on political orientation of think tanks using linear support vector classifier 39

LIST OF TABLES

Table 1 Top five distinct phrases used by each think tank for ‘immigration’ 27

Table 2 Top five distinct phrases used by each think tank for ‘national security’.................. 28

Table 3 Top five distinct phrases used by each think tank for ‘judiciary’ 29

Table 4 Top five distinct phrases used by each think tank for ‘politics’ 30

Table 5 Top five distinct phrases used by each think tank for ‘social policy’........................ 31

Table 6 Top five distinct phrases used by each think tank for ‘monetary policy’ 32

Table 7 Top five distinct phrases used by each think tank for ‘trade policy’ 34

Table 8 Prominent feature sets by ideology for Multinomial Naive-Bayes classifier 41

Table 9 Prominent feature sets by ideology for Linear Support Vector Classifier................. 41

Table 10 Prominent feature sets by ideology for Multinomial Logistic Regression 42

ABSTRACT

Assessing Bias in Think Tanks

By

Mitalee Joshi

Think tanks are constructed on the premise of research-based advocacy of a subject. Yet, it has been observed that the published opinions of think tanks are often devoid of unbiased objectivity. In my Masters' thesis, I intend to establish that there exists an underlying bias in the articles published by a think tank, which is inferred from the choice of vocabulary used. For this, I used articles published between 2005 and 2019 by six prominent US-based think tanks to create a pool of about 100,000 articles. The think tanks are chosen randomly while ensuring that this selection represents the diversity of American political orientation. Since, the intrinsic nature of textual data poses the challenge of high-dimensional feature space, I chose to pre-emptively cluster the articles into prominent issue areas using word embeddings. This procedure of mapping articles to the maximum probable issue area is in part analogous to a probabilistic multi-label classification. Next, for each of the resulting subsets, I performed supervised multi-class classification and compared the output of various classifiers employed. The estimators deduced from these models help to construct a set of phrases that is characteristic of each think tank in each issue area. From this, I observed that for a classification rate in the range of 75 – 83%, these algorithms render similar feature sets for each think tank in a given issue area. These feature sets are further used to analyze the degree of overlap in the vocabulary used in the articles published by different think tanks. I observe that the data reflects a minimum overlap of approximately 20% between ideologically different think tanks, which increases to a maximum of 55% for certain ideologically aligned think tanks, thereby indicating that there exists an inherent bias in the analyses performed by think tanks, which can be primarily attributed to their founding ideology.

ACKNOWLEDGEMENT

I would like to take this opportunity to personally express my gratitude to my advisor Prof. Ajay Shenoy for his knowledge, patience, and motivation. He has been a constant source of inspiration while encouraging a sense of curiosity about the finer details entailed in the process. I would also like to thank Prof. Laura Giuliano, who has always been supportive of my aspiration to opt for a thesis additional to the coursework and went beyond the norm to enable it.

I thank the undergraduates at the Income Dynamics Lab for working with me to efficiently extract the colossal datasets involved in this analysis. In addition, I would like to sincerely thank Sandra Reebie in the Economics department at the University of California, Santa Cruz who ensured that the university processes remain hassle-free during my stay as a student at UCSC.

It goes without saying that none of this would have been possible without the incessant support of my family who have been there for me, in my highs and lows, always enabling me for the next step forward.

INTRODUCTION

Think tanks in today’s globally connected world form a significant basis of opinion-building and highly influence the policy-making process. In philosophy, these institutes intend to deliver purely research-based advocacy that is agnostic to any sort of alignments. However, it is often observed that in their published opinions, there exists an intrinsic bias which is strongly aligned with their founding principles. This comes across so frequently that it cannot be claimed to be a mere coincidence. In this paper, I intend to establish the existence of a biased undertone in the analyses published by these think tank through the words used in their text. Essentially, the objective here is to draw a set of words that is characteristic of a think tank, and to identify the extent with which it conforms to the founding ideology of the think tank.

I referred to the statistical predictability model of speech to delineate the differences between groups (Gentzkow, Shapiro, & Taddy, 2019) as my guiding framework. For this problem statement, first, I analyzed a pool of articles published by selected think tanks in a defined timespan. Next, I extended the design by pre-emptively classifying articles into their respective issue areas. Finally, I trained the transformed dataset using an array of linear and non-linear estimators to identify the prominent phrases that are instrumental in distinguishing a think tank from another.

The above procedure is carried out on a high-dimensional data structure, which often grows exponentially with the volume of distinct words used in the pool of articles. This poses a challenge of ever-increasing dimensionality with increasing number of articles. Also, inherent to the semantic qualities of textual data, single words, unrelated to their position in a sentence, are less informative than a sequence of words, which is often indicative of the context of a text. Hence, observing a sequence of words (also referred to as ‘phrases’), as the unit of analysis, reduces the dependence on the relative position of words while interpreting a piece of text. As a result, the key trade-off is between the size of a phrase used to train the model and the computational cost of training. Similar trade-off also exists in classification. In this work, I propose to represent the corpus using a sparse structure along with restricting the dimensionality of the feature set by weighing rare words more than extremely common words. This two-pronged strategy transforms the pool of articles into a sparse array, which can be decomposed to reduce the computational expense while improving classification by alleviating the noise entering the model training procedure.

Since the data exhibited a lack of perfect separation in the language used by the think tanks (selected for analysis), I enriched the dataset with additional information regarding the context of articles. This segmented the master-dataset into contextually similar subsets. This classification relies on latent variables, for which I used word embeddings of the text to map articles to their primary issue areas.

Following this, I trained the model to classify the articles based on the observable value of the authoring think tank using an array of estimators. The ultimate objective of this classification is to compare the weights given to the phrases used, thereby drawing the more informative phrases to identify a think tank. Linear estimators like maximum-a-posterior and maximum likelihood estimator and non-linear estimators like support vector classifiers and decision trees are employed for classification. In accordance with

(Gentzkow, Shapiro, & Taddy, 2019), I observed that classification performance improves after imposing a penalty to the objective as it helps to reduce the dimensionality of the dataset. However, given the low volume of articles available and the diversity of context in scope, the penalized maximum likelihood estimator failed to converge. Instead, support vector classifiers were a good choice in this case even with a linear boundary function. Furthermore, random forests proved to be a strong classification method, rendering eminent utilization of the abundant information available in the articles. However, due to the complexity of its multiple set of rules it is not feasible to use random forests to identify the characteristic phrases of a think tank.

Based on this experiment, the support vector classifier with a linear boundary function proved to be an optimal choice. By comparing the resulting set of phrases, I can conclude that the interpretation of facts by a think tank exhibits a definite tenor that is strongly propagated through its choice of vocabulary and conforms to its ideology.

THINK TANKS DATA

The think tanks data has been primarily collected from the current and archived repositories of six think tanks, namely, Brookings Institution (BI), Heritage Foundation (HF), Center for American Progress (CAP), Center for Immigration Studies (CIS), Cato Institute and Urban Institute (UI). I have restricted the choice of think tanks to those headquartered in the USA and founded before 2000 to assert substantial uniformity in the acquired data. The time span of observation ranges from the year 2005 to 2019 creating a database of approximately 100,000 articles with an average length of 150 words each. In collaboration with the Income Dynamics Lab, I created the dataset by scraping the internet archive. Then, I collated all the articles after converting from various downloaded file formats like PDF, CSV to a tabular format in a single CSV file where each record d_i in the database D contains the author's name, date of publishing the article, content of the article and name of the think tank. To standardize, I created two identifiers: (i) thinktank ID which is the think tank's initials, e.g., 'HF' for Heritage Foundation, (ii) article ID which identifies every article by concatenating the thinktank ID and an integer generated during the conversion of each article, uniquely for each think tank, e.g., 'BI_282' refers to the 282nd article picked from the Brookings Institution.

The primary input for training the model is a dataset A obtained after pre-processing the set D. Here, each record in A represents a published article. A record comprises four attributes, (i) article ID, (ii) name of author(s), (iii) think tank ID, (iv) processed text of the article (henceforth, referred to as the article represented as a_i). The pre-processing function is designed to clean the raw text and convert it into a set of bigrams, i.e., sequence of two adjacent words in the text, hereafter being mentioned as 'phrases'. The choice of two adjacent words to make a phrase justifies the trade-off of performance over computational cost. This is due to the exponential change in the dimensionality of the dataset as the size of a phrase increases from two to three adjacent words, i.e., each increment in the size 's' of phrases selected exponentially increases the dimensionality with the order s of the phrases considered (Gentzkow, Kelly, & Taddy, 2019). The preprocessing comprises of the following steps: (i) replacing punctuation marks and special characters with user-defined delimiter, (ii) removing ‘stopwords’[1] defined for English language, i.e., the extremely common words used in English sentence construction, (iii) transform words to their corresponding stem form based on the Porter2 stemming algorithm (Porter, 2006). This step converts every processed article to a corresponding set of word stems a_i. Here, I choose to remove articles with less than 10 elements in their processed set a_i, which results in the working dataset of 94,250 articles with six distinct think tanks.

This set of articles becomes the primary source of data for further analysis. I used this data to build a corpus that comprises all the meaningful phrases in the form of numeric vectors. Here, a meaningful phrase refers to any bigram phrases which either (i) is not extremely common, by virtue of semantics, i.e., not used in more than 75% of the articles, or (ii) is not extremely rare, i.e., not used in less than 0.01% of the articles. The vectorizer function is used in conjunction with a lasso penalty to reduce the dimensionality of the model. This results in corpus W that is a set of 1024 unique phrases, where a phrase can be represented by w_i, for j referring to the index of a phrase in the corpus.

The resulting corpus when overlaid on the articles results in a sparse matrix V in which every record V_i is a vector representation of the corresponding article a_i in A.

$$V_{ij} = count(w_j), \qquad w_j \in a_i$$

Henceforth, vector V_i derived from A, becomes the unit of analysis.

[1] Stopwords are taken from the NLTK stopwords corpus (*https://www.nltk.org/book/ch02.html*)

The choice of language of an article is often affected by various factors like its author, subject, time of publishing, etc. In our observation time-period, I have observed insignificant instances of any author moving from one think tank to another. Hence, it is safe to assume that the bias specific to an author cannot be significantly distinguished from the bias of the think tank under observation. This assumption is further strengthened due to the massive size of the writing teams housed in each think tank and the fact that most of these articles are authored by more than one authors, which makes the individual effect of an author on the phrases used in an article insignificant.

However, the subject of an article plays a decisive role in choosing the words used in an article. Since this information was not available in the data, I identified twenty issue areas based on the subjective knowledge of the issues that think tanks talk about. After iteratively refining the issue areas, nine issue areas (e.g., immigration, social policy, judiciary, etc.) were finalized based on the critical mass of similar articles created around each issue area. Here, similar implies syntactic as well as semantic similarity between articles. The term syntactic similarity refers to the phrases with the same etymological roots, i.e., grammatical transformations of the same word. Semantic similarity refers to two different phrases which imply similar connotation given the same context. For example, in the context of national security, 'military' and 'pentagon' will have similar connotation, whereas without context pentagon can be The Pentagon, a geometric shape or a credit union.

Hence, with 94250 articles published by either of the six think tanks and a corpus of 1024 unique phrases, I performed the analysis at the level of an article for a given issue area.

MODEL

From the problem statement, it is evident that there exists specificity in the vocabulary used by a think tank. However, in cognizance of the fact that every think tank is involved in research belonging to multiple issue areas and the context of research makes the language highly concentrated within an issue area. This essentially emphasizes partitioning the dataset based on issue areas due to a high degree of separation between the vocabulary used in within-area articles relative to that in across-area articles. Hence, prior to modelling the problem, I intend to define issue areas for each article in A.

Clustering articles based on issue areas

I begin with identifying broad categories of topics on which the articles have been written. This is primarily based on subjective observation. The objective is to further refine it into defined issue areas which partition the database into balanced subsets. The fundamental idea is to create clusters based on syntactic and semantic similarity between the phrases used in the articles. Syntactic similarity can be measured in the commonality of occurrence of a phrase between two articles of comparison, relative to the entire pool of articles. Semantic similarity, on the other hand, tries to establish the similarity of context in the usage of two different phrases. For example, for a phrase ‘pentagon looks’, any two articles with this phrase reflect syntactic similarity, however if one article has this phrase surrounded by phrases like ‘state marines’ and the second one has surrounding phrases like ‘geometric figure’ then these articles reflect semantic dissimilarity. This is because from the surrounding words it can be inferred that the first article talks about national security while the second one might be talking about something probably related to architecture.

I propose to establish this notion of similarity by using a reference corpus[2] and the continuous bag of words (CBOW) approach (Mikolov, Chen, Corrado, & Dean, 2013) to build a matrix J where each record is representative of an issue area or category. The matrix is structured as a rule set to define the typical phrases pertaining to a given issue area. It comprises three columns, (i) primary phrase which identifies with the respective ‘category’, syntactically and semantically, (ii) set of phrases which are semantically similar to the primary phrase, (iii) set of phrases which semantically negate the primary phrase for the given context or in other words, are definitely dissimilar to the primary phrase. Hence, J becomes an exhaustive representation of W which maps all the relevant phrases to a category of topic, referred to as ‘issue area’. This approach of using word embeddings instead of a numeric representation of each phrase allows to reasonably assume conditional independence of the phrases. The assumption of conditional independence implies that the posterior probability is indifferent of the position of a given phrase. When using word embeddings, the idea is to project the corpus in a high-dimensional space such that semantically similar phrases are located

[2] Corpus taken from gensim in Python, a corpus developed over the Wikipedia database

close to each other. The dimensionality of this vector space is defined by the attributes used to define the rule sets.

Further, to factor for the less frequent category of articles, I balanced it using term-frequency vectors (t_{ij}) corresponding to the articles. This, essentially, is formulated by the matrix multiplication of V with a vector idf representing the relative importance of all the phrases in W. The resulting matrix will represent each unique phrase w_j in an article a_i as a number which is the frequency of occurrence of that phrase in a_i normalized by the `frequency of occurrence of w_j in all the articles belonging to A.

$$idf(w_j, A) = log\left(\frac{n}{|\{a_i \in A | w_j \in a_i\}|}\right)$$

Hence, I arrive at the metric of similarity by taking a dot product of J and the product of V and idf.

$$\hat{k} = argmax_k\ (J_k(a_i(f(d_i(content)))) \cdot ((V_i)(idf))$$

$$q(a_i) = J_{\hat{k}}(category)$$

Here, q refers to the issue area to which article a_i belongs. This is employed as a recursive procedure starting from 20 categories to refining them down to 9 issue areas.

Now the data set A is updated with a fifth attribute, i.e., the issue area of an article, represented by q. For the ease of computation, I have modeled a one-to-one assignment between the articles and issue area. On validation, this works well as there always is a dominating theme of every article even if it talks about multiple issues.

Classifying articles based on the authoring think tank

After mapping each article to an issue area, it can be stated that the vocabulary specific to a think tank for a given issue area can be given as:

$$V_i \sim Multinomial\left(V_{it}, y_{qt}\right)$$

where $V_{i,t} = \Sigma_{j\in[1,n]} V_{ij,t}$ refers to all the phrases used by a think tank and y_{qt} refers to the vector of probability of a phrase being used in an article authored by think tank t and issue area q. This probability can be derived in frequentist terms given by:

$$y_{qt} = [p(w_j \in a_i)|(issue\ area(a_i) = q)\ \&\ (think\ tank(a_i) = t]$$

In this way, I can say that the phrases that an article from a think tank t is likely to use for a specific issue area q can be defined by the set of all phrases used by think tank t (i.e., V_{it}) and the conditional probability vector y_{qt}.

In practical sense, a published article always has the name of the authoring think tank as credible information, therefore this can be used to drive multi-class supervised classification. This implies that for any article a_i mapped to issue area q, conditional on the phrases w_j, it is predicted to be authored by think tank $\hat{t}$. Here, the idea is to estimate the probability of each think tank t_k being the author of a given article, conditional on the phrases it contains and the issue area it is talking about. Then the think tank which has the maximum conditional probability is predicted to be the authoring think tank $\hat{t}$.

Based on this classification, I obtained a set of coefficients corresponding to the phrases in the corpus which indicate the relative importance of the phrases to identify the authoring think tank. This result set is then used to extract the set of phrases which are informative to identify the authoring think tank specific to the mentioned issue area.

Analogous to the assumption of independence of irrelevant alternatives of the multinomial logit model, this specification relies on the assumption that there is no significant difference between selecting all or some of the phrases w_j from the feature set g_{qt}. This implies that even if few phrases are excluded from g_{qt} that would not have any impact on the relative importance of the remaining phrases. Even though this limits the design by overriding the contextual preference of one phrase over the other, yet as stated in (Gentzkow, Shapiro, & Taddy, 2019) this can be used as a potential benchmarking tool while adhering to the simplicity of computation of a high-dimensional feature set.

In this context, it is evident that while every article can only be authored by one think tank, there exists a high probability that a single article addresses multiple issues. However, here, I proceed with the assumption that each article can be mapped to any one issue area that dominates the content of that article. By observation of the distribution of probability of an article to be talking about each of the nine issue areas, it is found that the dominating issue area has much higher probability than the others. Hence, the above assumption is safe and aids to computationally simplify the analysis.

ESTIMATION

The classification problem here can be modeled as a multi-class, supervised learning classification problem. A straightforward approach to solve this is to identify the probability of a think tank being author to a given article and label articles based on the most probable think tank. Hence, the proposal is to formulate a maximizing objective based on a probabilistic function.

Maximum-a-Posteriori estimator

A straightforward approach would have been the maximum likelihood estimator which is based on the concept of conditional density of the dependent variable on the independent features. However, if the system can be enriched with the knowledge of the probability of distribution of the independent feature vector, it is possible to align it to the idea of Bayesian conditional probability. Based on this information, I use the maximum a posterior (MAP) estimator for classification of articles to their authoring think tank. The MAP estimator employs a maximization objective of the conditional probability based on a prior information over the quantity to be estimated.

This can be described as:

$$\hat{t}_{MAP}(a_i) = argmax_{t_k} \, f(w|t_k)g(t_k)$$

where $f(w|t_k)$ refers to the probability of occurrence of a set of phrases w (which is a subset of the corpus W) when the article has been published by think tank tk. $g(t_k)$ refers to the prior density of the think tank classes.

The most natural choice, in this context, is to incorporate the MAP estimator through the Naïve-Bayes classifier (Manning, Schütze, & Raghavan, 2008). The classifier learns through the data itself. The problem statement is looking for the posterior probability to evaluate the probability of think tank t_k authoring an article ai which contains the phrase w_j. This can be expressed as:

$$\hat{P}(t_k|a_i) = \hat{P}(t_k)\prod_{j=1}^{n}\hat{P}\left(w_j\middle|t_k\right)$$

This can be transformed to logarithmic function to utilize the idea of maximizing the log-likelihood of $P(w_j|t_k)$ as:

$$log\,\hat{P}(t_k|a_i) = log\,\hat{P}(t_k) + \sum_{j=1}^{n} log\,\hat{P}\left(w_j\middle|t_k\right)$$

Here, $P(t_k)$ refers to the empirical frequency of an article to be authored by think tank t_k given by $\mathrm{P}(\mathrm{t_k}) = \frac{\mathrm{N(thinktankID(a_i)=t_k)}}{\mathrm{N}}$. The likelihood here is given by the ratio of empirical frequencies of the phrase w_j in all articles authored by think tank t_k to that in the entire corpus W, i.e., $\mathrm{P}\left(\mathrm{w_j}\middle|\mathrm{t_k}\right) = \frac{\mathrm{n_{w_j,t_k}}}{\mathrm{n_{w_j}}}$.

Henceforth, the target class can be predicted through a maximum a posteriori (MAP) objective, i.e.,

$$\widehat{t_{MAP}} = argmax_{t_k}\hat{P}(t_k|a_i)$$

The Naïve-Bayes classifier is based on the assumption of independence which implies that the conditional probability of a phrase is independent of the position it takes in the article. This is a strong assumption to make, as this will cause the classifier to learn a phrase in the same way in all its occurrences irrespective of its contextual positioning. However, this has been taken care of to a large extent during topic classification where syntactically similar phrases with different semantic interpretations have been tagged to different issue areas. This ensures semantic exclusivity of phrases to implement this approach.

However, this approach relies heavily on the information of the prior. It is possible that the relative contribution of think tanks to this pool changes dramatically which would cause the model to be re-calibrated.

Penalized Maximum Likelihood estimator

With the MAP estimator, the model relies on learning the posterior distribution based on the knowledge of prior distribution. If the prior is uniformly distributed, the MAP estimator is the same as the maximum likelihood estimator (MLE). In practical terms, the distribution of the articles from different think tanks does not follow uniform distribution. However, if a penalty (Zou & Hastie, 2005) can be imposed on the estimator to shrink the feature sets, then pertaining to that limited pool of phrases the think tanks can be assumed to similarly contribute to the pool of articles over a given timespan.

Let t be the vector of possible target classes with nine elements, each representing a think tank t_k. Then the log-likelihood that a phrase w_j is used by a think tank t_k in any of its articles $a_i^{(k)}$ is given by:

$$L(\boldsymbol{\beta}) = \sum_{j=1}^{n} log\, P\left(w_j \middle| t_k\right) = f_{t|W}(t|W;\ \beta)$$

Here, $\widehat{\beta}$ refers to the estimated parameters. The objective here can be written as a maximization problem, i.e.

$$\widehat{\boldsymbol{\beta}} = argmax_{\beta} \sum_{k=1}^{9} L(\beta) - \lambda \sum_{j=1}^{n} \left|\left|\beta_j\right|\right|_1$$

This represents the value of parameters β obtained after maximizing the log-likelihood while penalizing the features. The likelihood function depicts the probability of an article being authored by think tank t_k conditional on the phrases present in the article from our corpus. This results in a vector $\widehat{\boldsymbol{\theta}}$ where each entry θ_{kj} represents the additional probability of the t_k being the authoring think tank if it comprises of the phrase $w_{j.}$

The second term of the objective function aims to shrink the estimators so as to reduce dimensionality, i.e., for phrases with minimal relevance, it will try to reduce the estimator to zero. Here, relevance is defined based on the frequency of occurrence of a phrase in the subset of the corpus derived from the pool of articles authored by a think tank t_k. While penalizing, it is helpful to opt for L1 penalty as it helps to reduce the dimensionality of the problem case. A blend of 80% L1 penalty along with the remainder L2 penalty, has shown to perform marginally better, however the computational overhead makes L1 as the obvious choice.

This can be incorporated using logistic regression with an L1 penalty and choosing a suitable optimization algorithm. However, with data volume not being sufficient to scope in the multi-faceted nature of content available, the convergence of error is challenging. It suffers from creation of local minima of the least square error term. However, in different runs I observe that even though the estimator differs in magnitude, the relative order of variables/features remains the same. Hence, it becomes sufficiently useful to conclude the most informative phrases for each think tank given an issue area.

Inherent to the approach of finding group differences, these approaches suffer from a bias as the correlation between the error term and the choice of words is not perfectly independent. Often, there exist exogenous reasons to use certain phrases more than others, which can dominate the model during training.

Further, for a very large population the MAP estimator tends to asymptotically converge to the maximum likelihood estimator in distribution. So, in cases where the prior starts to shift from its original state, the maximum likelihood estimator will begin to perform better given that it does not take account of the prior information of the target classes. Despite converging in distribution asymptotically, the penalized MLE is not rigidly tied to the assumption of conditional independence of features, hence it adjusts its parameters accordingly to maximize the conditional likelihood of the data (Ng & Jordan, 2002). Also, by virtue of a sparse corpus, the frequency term of the less-frequently phrases shrinks to zero.

Regularized non-linear estimator (Support Vector Classifier)

An alternative approach to solve this problem employs a regularized m-estimator. Also, known as the Support Vector Machine (SVM), this entails projecting the pool of articles in a space and then finding a hyperplane which separates them into different groups. In theory, for a perfectly separable dataset there exist infinite hyperplanes. In context of the problem statement, I have implemented a ‘one-vs-rest’ strategy to formulate the multi-class classification as multiple binary classification problems.

For each think tank t_k, the target class can be defined as:

$$\widehat{y_i} = +1 \; if \; x_i^T \boldsymbol{\beta} > 0$$

$$\widehat{y_i} = -1 \; if \; x_i^T \boldsymbol{\beta} < 0$$

where x_i represents the vector of the intercept term and the n phrases forming the corpus W. This implies that for each think tank t_k, this problem will be modeled as a binary decision problem, i.e., to create a separation between articles authored by think tank t_k with a threshold at unity. Then this can be defined as a convex optimization problem:

$$\widehat{\boldsymbol{\beta}} = argmin_{\boldsymbol{\beta}} \left(\frac{1}{2}\left|\left|\boldsymbol{\beta}\right|\right|^2\right), subject\ to\ y_i(x_i^T \boldsymbol{\beta}) \geq 1$$

Here, the norm of the parameter vector is the Euclidean distance of a vector x_i, representing an article a_i, from the maximal margin. The maximal margin is the separating hyperplane for which the perpendicular distance of the nearest node is the maximum. The objective depicts the requirement to find a separating hyperplane which has the minimum distance to the farthest articles in the hyperspace projection, i.e., minimizes the distance of an article from the separating hyperplane for any article in the dataset. The constraint asserts the need of the assigned articles to be in the right side of the boundary in the hyperspace projection of articles. In other words, for an article a_i represented by ($x_i^T\boldsymbol{\beta}$) authored by think tank t_k, the estimators should be such that $x_i^T\boldsymbol{\beta}$ lies above the margin and the corresponding estimate of y_i for the margin associated with the kth think tank should be at least 1. Based on this, the estimator is used to predict the confidence score of each article that represents the signed Euclidean distance of that article to the classifying hyperplane. Finally, the predicted confidence scores are used to predict the authoring think tank of an article.

This estimator gives the flexibility of choosing a non-linear boundary function over a linear boundary function, in case of an imperfectly separable dataset. For this experiment, I have employed a linear boundary function as shown in the equation above given the computational complexity of non-linear boundary function.

Support vector machines are empirically known to perform better in high-dimensional spaces. However, I observed that the estimator is able to define the maximal margin more clearly while comparing think tanks of different ideology as the contrast is starker in those cases. For think tanks with similar ideology, the separation is not well-defined due to overlap in the features. However, a limitation of the support vector classifier lies in its design. As it works by putting data points, above and below the classifying hyperplane there is no straightforward probabilistic explanation for the classification, as with the other two estimators.

In implementation, our definition of penalized MLE is quite similar to the SVM. The boundary function of the SVM corresponds to the log-likelihood function of the penalized MLE and the SVM is penalized with L2-norm while MLE is penalized with L1-norm. However, SVM is more efficient in memory allocation causing it to be computationally faster than penalized multinomial logistic regression. The multinomial Naïve-Bayes becomes helpful in classifying think tanks for categories for which relatively fewer articles have been published, for example, issue areas ‘trade policy’ and ‘politics’.

Alternative methods – Decision Trees

Decision-tree based classifiers also render classification formulation suitable for high-dimensional spaces. The random forest classifier is an adept choice where a non-linear classifier learns through decision trees. The decision trees are sampled recursively to reduce variance and are hence known to perform a stronger classification. The estimate of the conditional average treatment effect is the average over the estimates computed over B subsampled trees.

$$\hat{\beta}_{RF} = \left(\frac{1}{B}\right)\sum_{b=1}^{B} \hat{\beta}^{*}{}_{m,b}$$

The value of B ranges from 80-120, with distinct knee at 100 estimators and performance is observed to plateau beyond that. However, the complexity of the numerous estimators created makes it difficult to interpret the relative importance of the features in order to create a set of highly informative phrases.

VALIDATION

Based on the estimators described in the previous section, I chose to compare the performance of the Naïve-Bayes classifier, multinomial logistic regression with L1 penalty and support vector classifier with linear kernel to classify the articles into target think tanks.

A straightforward implementation would be to create a feature vector comprising of the corpus of phrases and the nine issue areas as dummy variables. In this case the coefficients to each issue area would reflect the fixed effect of that issue area. Alternatively, I chose to train the model to classify articles from each issue area separately. This ensures strong classification for primarily as it resolves the following possible concerns: (i) there exist overlapping phrases between articles from different issue areas due to contextual references, (ii) there exist syntactically similar phrases which are semantically different and can indicate different context. Such ambiguity is intrinsic to textual data. However, in cognizance of the fact that the issue areas have been assigned based on word embeddings of the text, partitioning the database based on these issue areas ensures the clustering of articles with syntactic as well as semantic similarity together. Hence, the primary difference amongst the articles within each subset remains of their authoring think tank which, in this case, is an observed variable.

Hyperparameter optimization using cross validation

Hyperparameters are the parameters whose value is used to control the learning process while training a model. These are different from the usual 'parameters' as their values cannot be derived via training and need to be defined while initiating the model. The hyperparameters required are specific to an algorithm or classifier, as in this case. The hyperparameters for the classifiers used for this problem statement are:

(i) Naïve-Bayes
'alpha' – defined as the additive smoothing parameter, refers to the amount of smoothing the estimator undergoes

Parameters	**Possible Values**
Alpha	0.0, 1e-5, 1e-6, 1e-7, 1e-8, 1e-9, 1e-10, 1.0

(ii) Multinomial logit
'penalty' – specifies the norm used in penalization, and whether to penalize
'C' – defined as the inverse of regularization strength

Parameters	**Possible Values**
Penalty	L1, L2
Class weight	Balanced
Solver	SAG, SAGA
Multi class	Multinomial
Warm start	True, False
Max iterations	100, 500, 1000, 5000

(iii) Linear Support Vector Classifier
‘penalty’ – specifies the norm used in penalization
‘loss’ – specifies the choice of loss function
‘C’ – defined as the inverse of regularization strength
‘multi_class’ – Specifies the choice multi-class strategy that may be one-vs-rest or optimizing over a joint objective over all classes

Parameters	**Possible Values**
Penalty	L1, L2
Loss	Hinge, Squared hinge
Class weight	None, Balanced
Dual	True, False
Multi-class	Ovr, Crammer singer
C	1.0, 1.25, 1.5
Max iterations	1000, 5000

These hyperparameters need to be selected to serve two purposes. One, to optimize the performance of the model, and second, for model selection or selecting underlying algorithms that best suit the given scenario. To decide the various hyperparameters available, I have employed the Grid Search approach. Grid Search is an exhaustive searching technique for hyperparameter optimization. This technique exhaustively searches in a pre-defined subset of the hyperparameter space of the learning algorithm and compares the relative performance measured by cross validation based on refit time and a scoring function. In this case, I have chosen 'accuracy score'[3] as the objective metric.

After exhaustive searching amongst these parameter values, the performance after cross-validation of various permutations of hyperparameters is compared to find the best performing permutation for each issue-area-based partition of the dataset.

Model training and validation

The procedure comprises twenty-seven training routines, with three classifiers for each issue area. As mentioned earlier, the first step is to classify articles based on the issue area they are talking about. This classification is based on word-embeddings based on the continuous bag-of-words approach. The best performing classifier for each issue area is chosen based on cross-validated model training, hence tested on out-of-sample data.

[3] Accuracy score (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html) refers to the fraction of correctly classified samples.

A primary concern, as stated earlier, is the unbalanced nature of the dataset, especially in terms of the number of articles available to train the classifier within each issue area. This can be seen in the data distribution plot below.

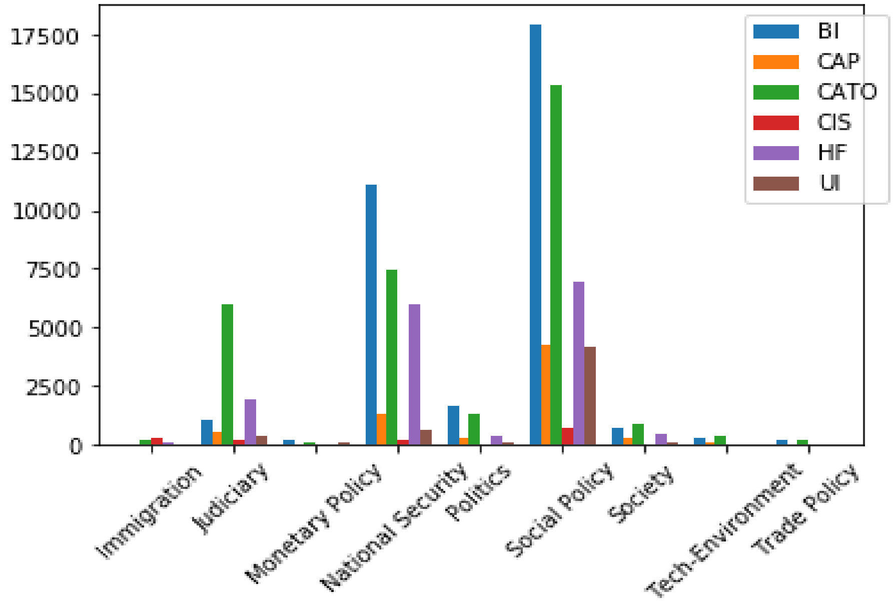


Fig. 1 Distribution of articles across the nine issue areas

In Fig. 1, it can be observed that the number of articles for the category 'society' and 'technology & environment' collectively constitute only 3% of all articles. Also, the diversity of sub-topics under these headers presents a challenge to train a classifier with such a small dataset. Therefore, I decided to drop the articles mapped to these two categories from our analysis. The categories 'trade policy' and 'politics' also have small datasets but the limited topics these articles talk about give sufficient pool of phrases to classify it using a probabilistic classifier, still being inadequate for the support vector classifier.

A comparison of the performance of the three classifiers for each issue area is given in figure 2. Here, 'Naïve-Bayes' refers to multinomial Naïve-Bayes, 'Multi-logit' refers to multinomial logistic regression with L1 penalty and 'Linear SVC' refers to support vector classifier with linear kernel.

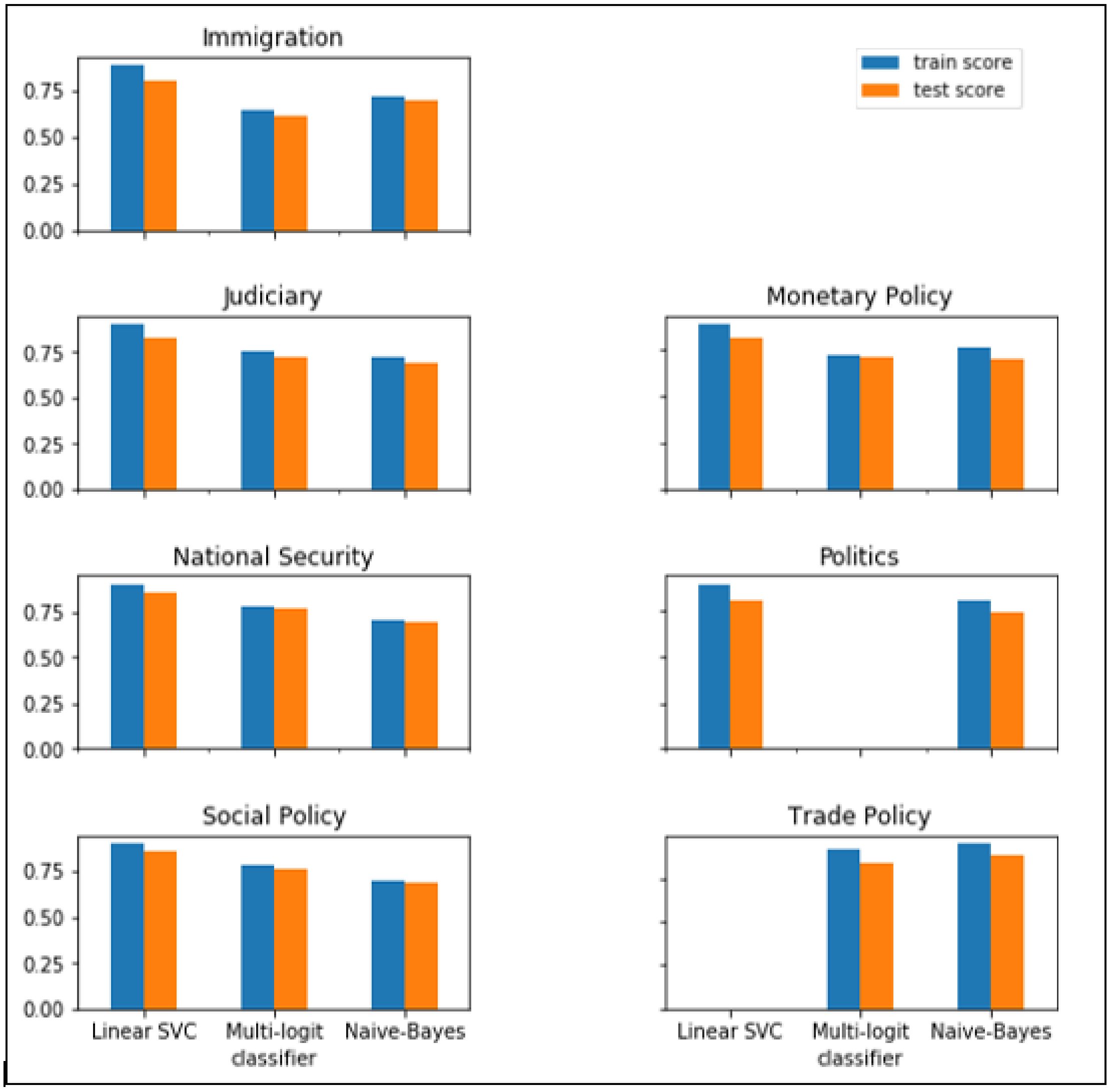


Fig. 2 Comparison of training and out-of-sample validation score

It can be observed that the linear SVC shows a higher training score for each issue area. However, observing the predicted think tanks for articles in categories like 'trade policy' and 'politics', it is found to be ineffective. This is primarily due to relatively fewer articles in these categories compared to the diverse phrases being used. This makes their projection in the hyperspace not perfectly separable by a linear boundary function. Choice of a radial boundary function seems computationally expensive when compared to the classification achieved by multinomial naïve-bayes classifier.

For the issue area ‘politics’, the penalized multinomial logit is not able to converge due to the presence of multiple local minima. This is attributed to the low volume of articles available for training while the dataset is unbalanced towards a few think tanks. It was observed that multinomial logit classified in close approximation with linear SVC when the datasets were synthetically balanced. This, however, posed two challenges: (i) the resulting feature sets were crowded with similar phrases due to disproportionate imbalance between articles of different think tanks, (ii) the overall process of training a synthetically balanced dataset proved to computationally expensive.

The multinomial naïve-Bayes is known to suffer from finite sample bias in categories where the number of distinct phrases is much higher relative to the number of training articles. It also imposes two strong requirements, (i) the prior probability distribution of articles remains unchanged, (ii) the data holds the assumption of conditional independence. However, for this problem statement, the pre-emptive partitioning of the dataset resolves the assumptions, hence, making the classifier sufficiently able to extract the relative importance of phrases to identify an authoring think tank from another. In contrast to this, linear SVC is able to perform better than multinomial naïve-Bayes without the constraint of prior probability distribution or the assumption of conditional independence of phrases. However, linear SVC is constrained as it requires the articles authored by different think tanks to be sufficiently distant from each other.

For categories where the different think tanks are less polarized, then a linear boundary function becomes inadequate to create a perfectly separating hyperplane. Multinomial logit is intuitively expected to perform similar to linear SVC for the given problem statement. However, lack of perfectly separable data, primarily due to a lot more phrases compare to the volume of training articles, results in a non-converging model.

Further, linear SVC depicts much lower refit time in cross-validation as compared to multinomial logit for the given partitioned datasets. Hence, linear SVC becomes a feasible choice as long as the size of corpus, indicative of the dimensionality of the model, is smaller relative to the volume of articles to be trained.

RESULTS

The results can be analyzed in two ways. The first approach is to compare the feature sets of one think tank vis-à-vis all other think tanks, and the second approach is to cluster think tanks based on their political orientation and then compare the feature sets between the think tanks of different political orientation. The former analysis aims to distinguish between the phrases one think tank tends to use distinctly for a certain issue area compared to the vocabulary all the other think tanks use. This tells about the differences amongst think tanks in their outlook of the same subject. The latter analysis, on the other hand, tries to depict similarity of expression within think tanks of similar ideology and their distinction from the think tanks of another ideology.

Analysis 1. One-vs-all the other think tanks in each issue area

Here, for each issue area, I select the classifier based on the best performance in out-of-sample validation with minimal observed overfit. Based on this selection criteria, I conclude the following:

(i) **Immigration**:

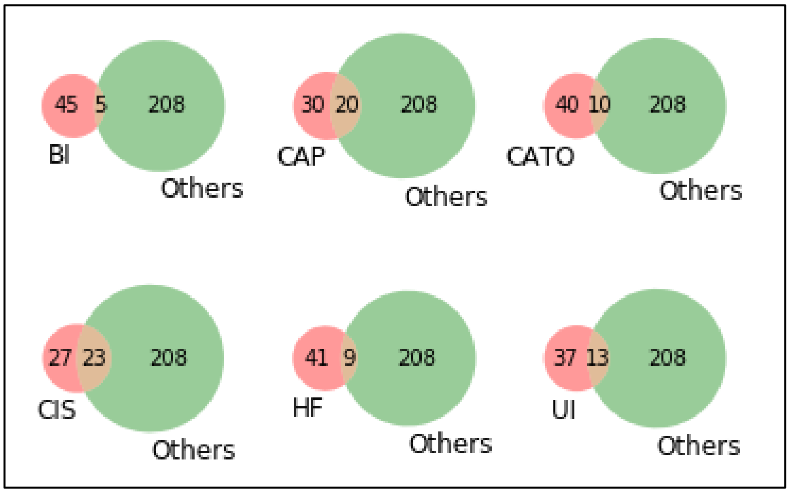


Fig. 3 Overlap of phrases of one-vs-all think tanks for ‘immigration’

For articles talking about ‘immigration’, I chose the linear support vector classifier as the number of articles in the pool is larger than the corpus of phrases used for the concerned articles. By its very nature, the topic of immigration has a two-dimensional treatment of analysis by the various research bodies.

Table 1 Top five distinct phrases used by each think tank for ‘immigration’

Think Tank	**Prominent Phrases**
BI	civil society, human capit, hillari clinton, educ system, africa growth
CAP	american people, econom polici, elect office, bush administer
CATO	govern would, central bank, cost billion, high cost
CIS	cut spend, crimin justice, civil society, civil liberti
HF	american militari, american people, econom reform, china sea

UI	develop country, crimin justice, insur program, feder budget

(ii) **National security**:

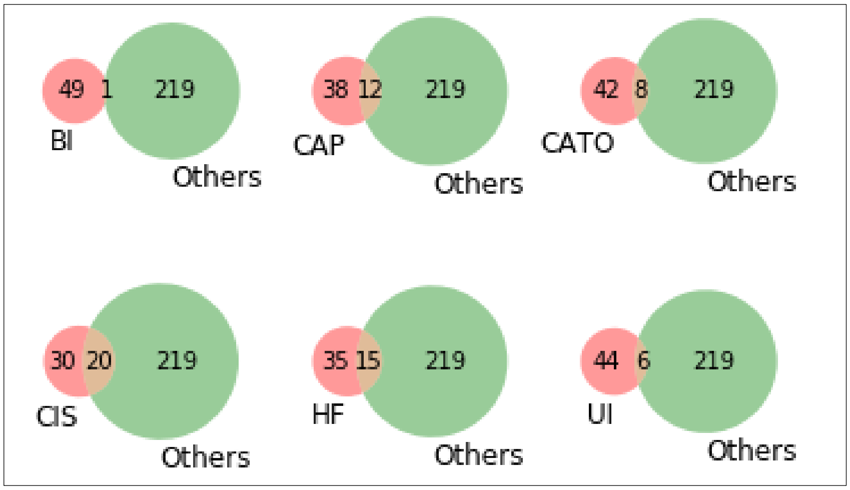


Fig. 4 Overlap of phrases of one-vs-all think tanks for ‘national security’

For articles about national security, by knowledge the articles mostly deal about illegal entities entering the border or the international relationships related to border security or international security measures. This indicates multi-dimensional treatment of analysis by various research bodies. However, the numerous articles published on national security ensured the sufficiency of dataset to train the model.

Table 2 Top five distinct phrases used by each think tank for ‘national security’

Think Tank	**Prominent Phrases**
BI	intern commun, human capit, global develop, intern monetary, civil war
CAP	american militari, american peopl, econom polici, bush administr, african american
CATO	econom polici, econom reform, civil liberti, bank monetari, gener govern
CIS	crimin justic, depart homeland, estim million, court appeal, african american

HF	american people, intern organ, civil society, intern trade, bin laden
UI	elect office, crimin justic, insur coverag, feder budget, increas econom

(iii) **Judiciary**:

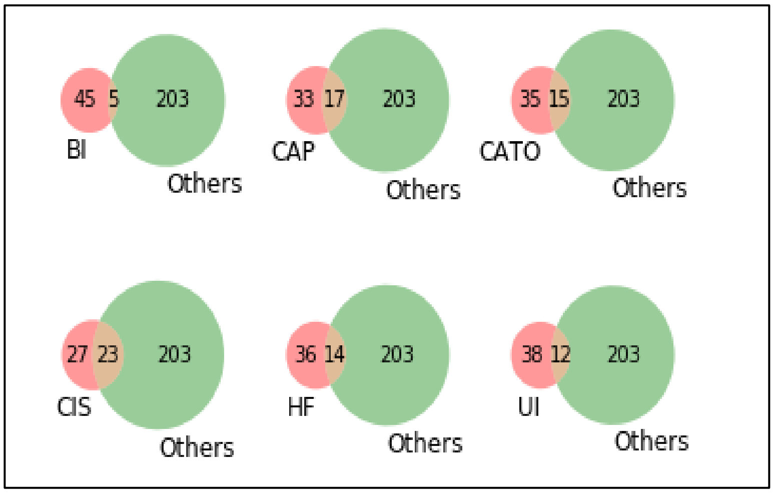


Fig. 5 Overlap of phrases of one-vs-all think tanks for 'judiciary'

For judiciary, both the linear SVC and multinomial Naïve-Bayes, despite their drastically different training scores, depict feature sets of similar relative feature importance. However, the dependency on the assumption of conditional independence is relaxed in the linear SVC and hence, I choose to use linear SVC for articles from the 'judiciary' issue area.

Table 3 Top five distinct phrases used by each think tank for 'judiciary'

Think Tank	**Prominent Phrases**
BI	updat econom, global develop, foreign polici, world bank, health care
CAP	health care, african american, climat chang, public opinion, feder govern
CATO	suprem court, govern polit, health care, constitut law, law court
CIS	illeg alien, immigr reform, law enforc, immigr law, censu bureau

HF	health care, suprem court, nation secur, presid obama, obama administr
UI	use data, health care, child care, social secur, incom tax

(iv) **Politics**:

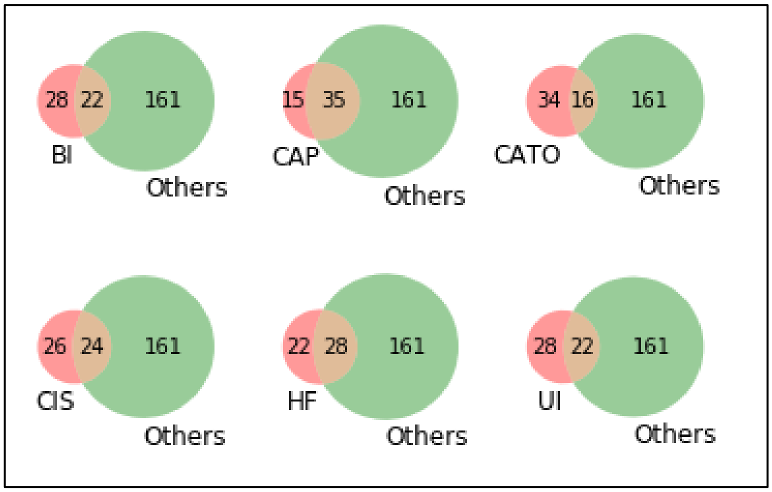


Fig. 6 Overlap of phrases of one-vs-all think tanks for 'politics'

The articles from 'politics' issue area show a restriction for the penalized multinomial logistic classifier to converge due to the fewer articles available relative to the multi-faceted nature of subjects in scope. However, the prior distribution here is strong as almost every think tank contributes a defined share of their publications to political subjects. In this case, the multinomial Naïve-Bayes classifier performs better than the others.

Table 4 Top five distinct phrases used by each think tank for 'politics'

Think Tank	**Prominent Phrases**
BI	foreign polici, global develop, econom growth, health care, labor market
CAP	health care, climat chang, barack obama, afford care, econom benefit
CATO	govern polit, health care, law court, feder govern, foreign polici
CIS	illeg alien, immigr polici, immigr law, econom benefit, law enforc

HF	health care, feder govern, econom freedom, homeland secur, civil society
UI	child care, afford care, great recess, hous market, econom recoveri

(v) **Social policy**:

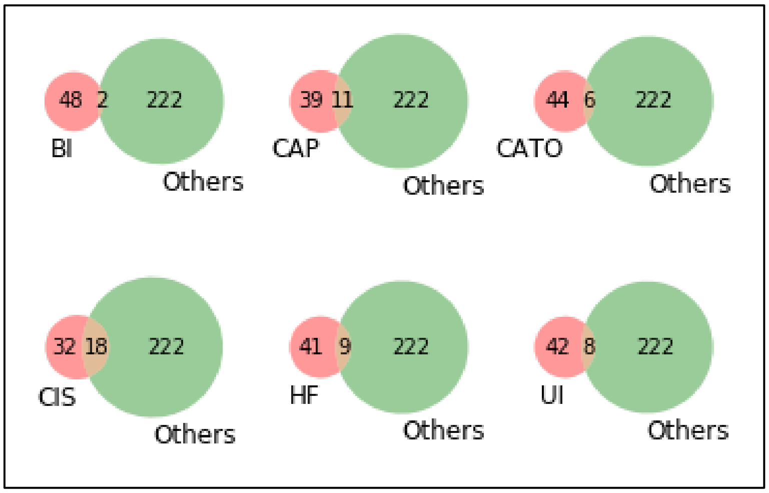


Fig. 7 Overlap of phrases of one-vs-all think tanks for 'social policy’

The issue area ‘social policy’ encompasses multiple subjects and each think tank, irrespective of its primary founding principles, contributes to this category of articles. This makes this subset of articles rich in information such that linear SVC is able to perfectly create maximal margins to differentiate between articles published by different think tanks.

Table 5 Top five distinct phrases used by each think tank for ‘social policy’

Think Tank	**Prominent Phrases**
BI	center health, global develop, africa growth, futur develop, educ system
CAP	environment protect, gross domest, bush administr, econom secur, creat job
CATO	intern licens, educ child, bank monetari, child polici, constitut law
CIS	illeg alien, immigr polici, legal immigr, american worker, labor forc

HF	appear nation, institut econom, econom freedom, law require, arm forc
UI	care act, hous repres, assist program, estim percent, incom tax

(vi) **Monetary policy**:

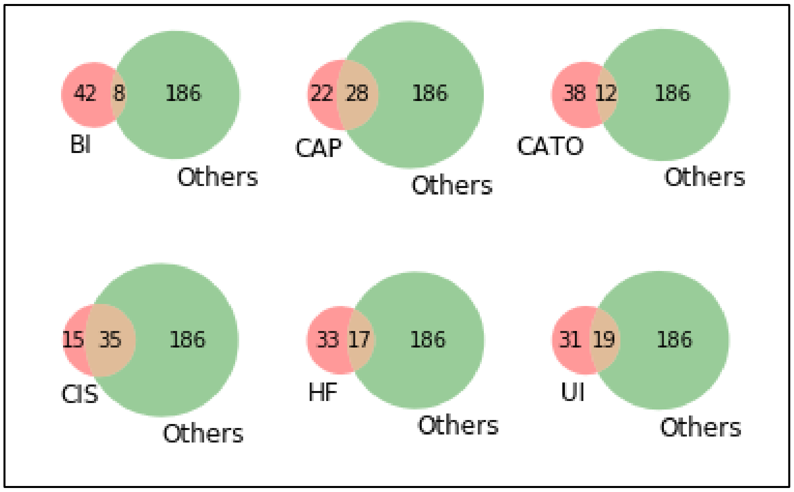


Fig. 8 Overlap of phrases of one-vs-all think tanks for 'monetary policy'

The issue area ‘monetary policy’ depicts similarity with the issue area ‘immigration’ in terms of its distribution of articles and distinct phrases due to fewer subjects in scope and uniformity of the subjects they relate to. hence, extending the reasoning linear SVC gives good classification model for the articles written about the country’s monetary policies.

Table 6 Top five distinct phrases used by each think tank for ‘monetary policy’

Think Tank	Prominent Phrases
BI	futur develop, center health, develop goal, data collect, long run
CAP	job growth, econom secur, american family, bush administer, immigr reform
CATO	develop immigr, govern polit, creativ common, financ bank, limit govern

CIS	illeg alien, illeg immigr, censu bureau, homeland secur, lower cost
HF	appear washington, assist secretari, defens budget, econom freedom, entitl program
UI	care act, earn income, increas number, improv quality, care provid

(vii) **Trade policy**:

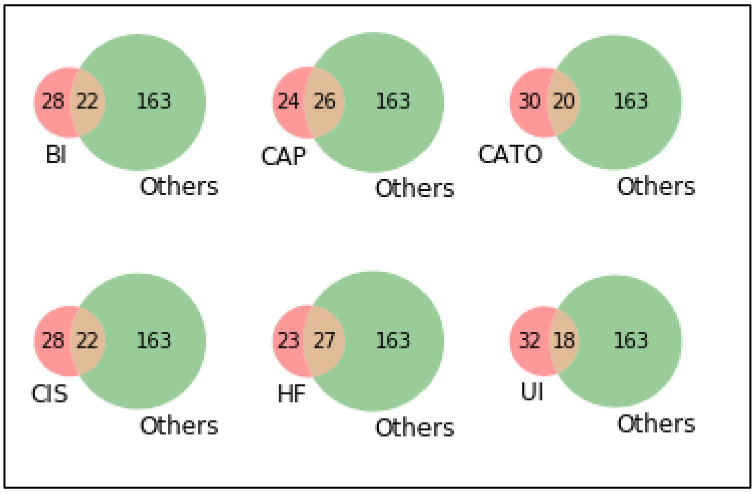


Fig. 9 Overlap of phrases of one-vs-all think tanks for 'trade policy'

In the ‘trade policy’ category of articles, there exist fewer articles with much more overlap between the articles from different think tanks. This category comprises of articles which primarily talk about the trade relations of the USA with international entities. A defining pattern emerges in the names of countries that the various think tanks prefer to focus upon. This distinction, as I explore, is much guided by the ideological leanings of each think tank. Hence, multinomial Naïve-Bayes shows a better trained classification model. Penalized multinomial logistic regression shows similar feature sets in all its iterations, however the presence of local minima causes it to not converge within computationally feasible limits.

Table 7 Top five distinct phrases used by each think tank for ‘trade policy’

Think Tank	**Prominent Phrases**
BI	polici update, foreign polici, updat econom, metropolitan area, prime minist
CAP	health care, african american, climat chang, clean energi, interest rate
CATO	constitut law, govern polit, law court, foreign polici, tax budget
CIS	illeg alien, homeland secur, labor market, justic depart, immigr polici
HF	suprem court, health care, polici studi, feder govern, obama administer
UI	health care, crimin justice, afford care, social secur, great recess

Analysis 2. Analyzing feature sets by political orientation in each issue area

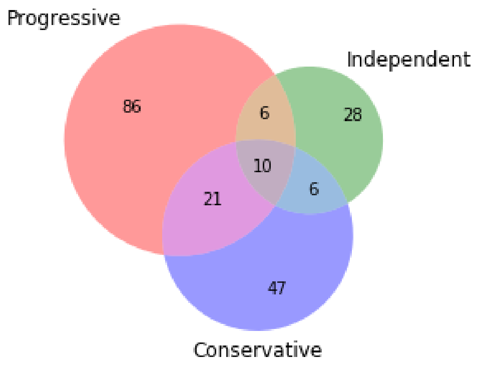
Once the feature sets of all think tanks are available, as a next step to this analysis, I assert that the differences in the vocabulary used is primarily guided by the founding ideology of a research body. This is often indicated through their documented political orientation. Based on this information, I grouped the six think tanks into three groups, i.e. (i) Conservative, (ii) Progressive, (iii) Independent & Liberal. For the think tanks in question, Heritage Foundation and CATO Institute are conservative-aligned think tanks, Center for Immigration Studies constitute the independent category of think tanks, and Brookings Institution, Center for American Progress and Urban Institute make up the progressive-aligned think tanks. Based on this classification, the prominent feature sets of the three classifiers are presented in the appendix.

(i) Naïve-Bayes Classifier

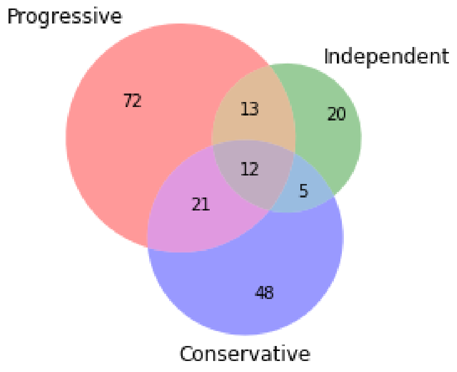
Based on the multinomial Naïve-Bayes classifier, I find a certain degree of overlap between the vocabulary used by the progressive-aligned think tanks and the independent think tanks as compared to their overlap with the conservative think tanks. The extent of overlap is found to vary for different issue areas which reflects the issue areas where the independent think tanks align with progressive ideology. This evidence corroborates the correlation between the founding ideology and the slant of articles published by a think tank.

The sets of prominent identifying phrases delineate the emphasizing tone of progressive think tanks on analysis of government policy, especially foreign policy, middle class, health care and job creation. On the other hand, supreme court, interpretations of constitutional law and child policy remain the focal talking points of the conservative think tanks across issue areas.

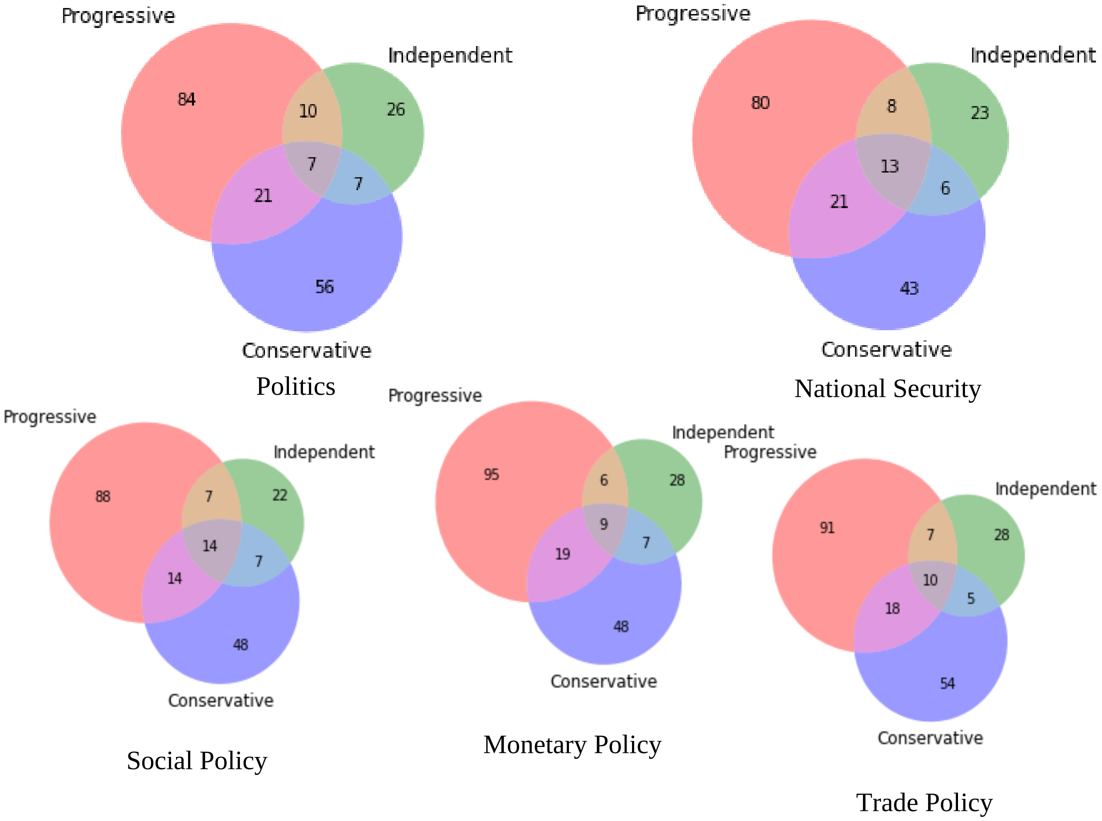
A representation of the overlap of phrases use between these ideological groups is shown below.



Immigration



Judiciary



Politics

National Security

Social Policy

Monetary Policy

Trade Policy

Fig. 10 Overlap between phrases based on political orientation of think tanks using Naive Bayes classifier

(ii) Multinomial logit

Even though the logistic regression could not converge within computationally feasible iterations, it is evident that the classifier is able to distinguish the most informative phrases that indicates the difference between the authoring think tanks. As depicted by the Venn representations below, the overlap is widest between the progressive and independent think tanks for articles related to judiciary and immigration. This reflects the alignment of these think tanks on judiciary and immigration-related issues.

Extending this analogy, a thin overlap between the progressive and conservative think tanks in articles on social policy indicate the polarized perspectives these wings adhere to in issues pertaining to social policies.

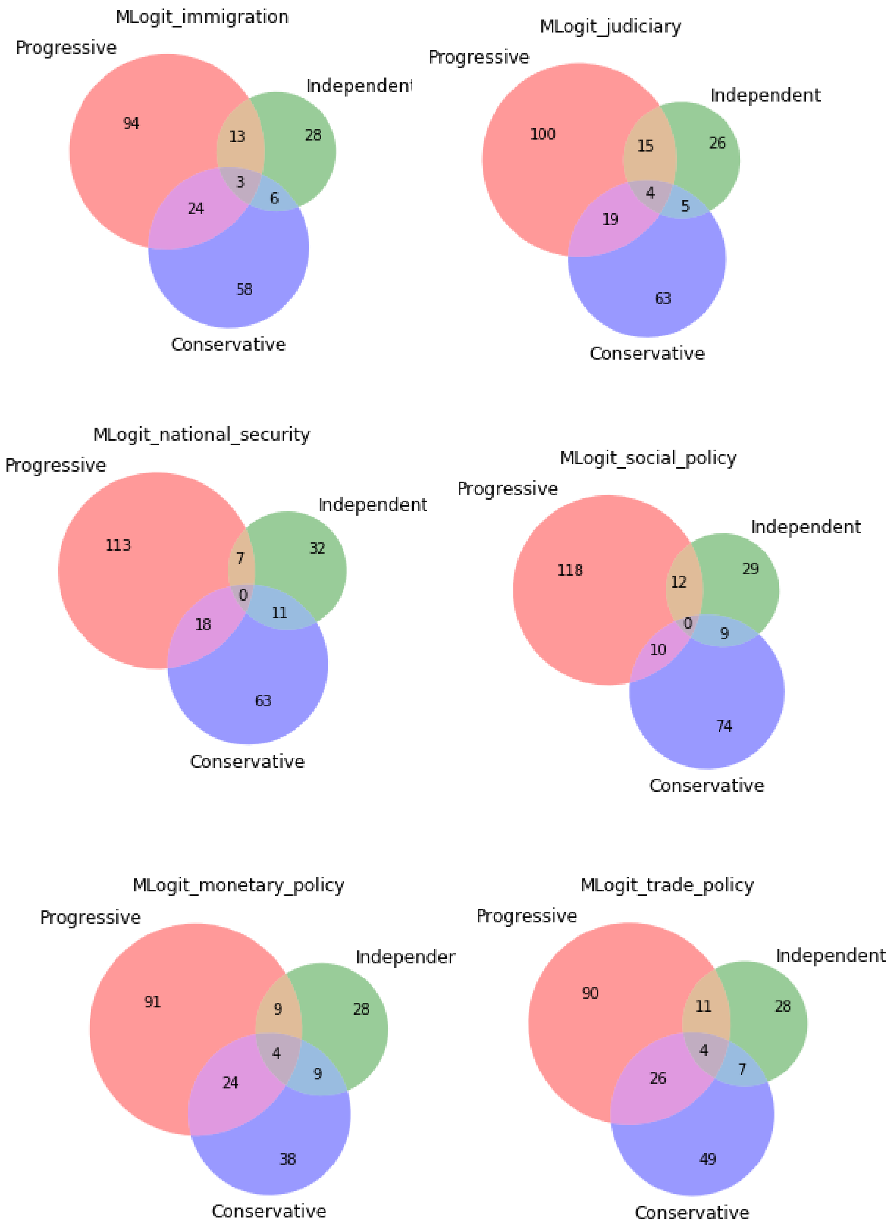
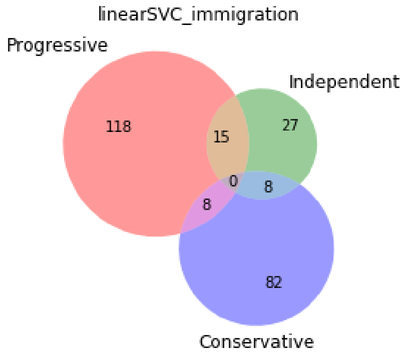
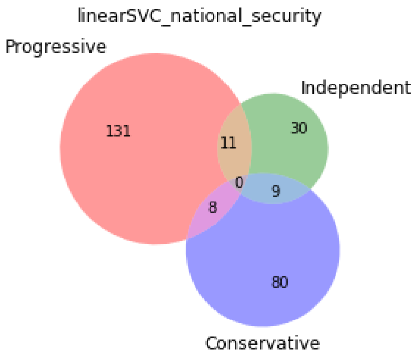


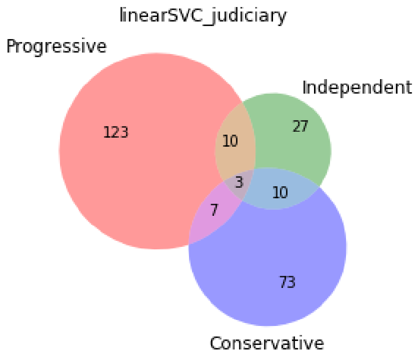
Fig. 11 Overlap between phrases based on political orientation of think tanks using penalized multinomial logit

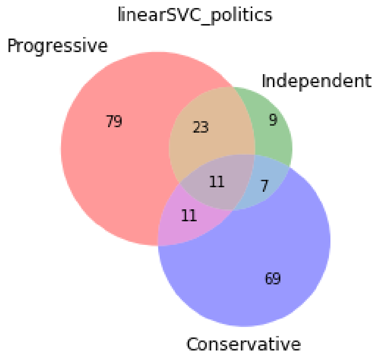
(iii) Linear Support Vector Classifier

As with the other two classifiers, the linear SVC depicts certain characteristic overlaps between the phrases used in think tanks of varying ideologies. Corroborating the previous results, social policy, national security and immigration form the issue areas with the maximum diverging vocabulary between different ideologies. It also conforms with our previous observations of relatively lesser polarized opinions in articles related to monetary policy and politics between the independent and progressive think tanks.









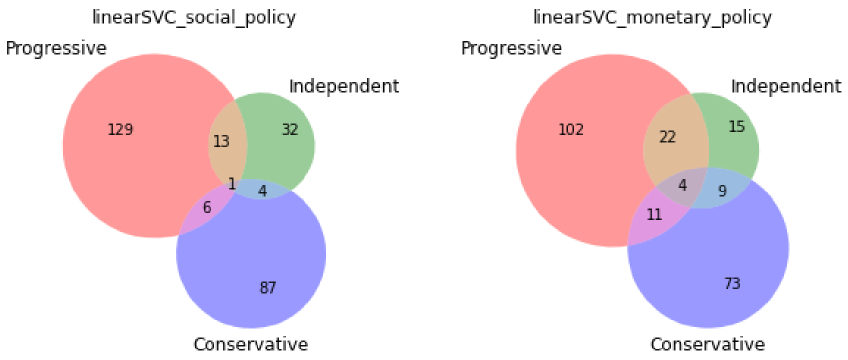


Fig. 12 Overlap between phrases based on political orientation of think tanks using linear support vector classifier

CONCLUSION

From these results, I can conclude that even though think tanks present research-based advocacy, it is not devoid of an inherent bias about the topics they talk about. As evident from the comparison of overlapping phrases based on political orientation, this subjectivity of a think tank's opinion can be often ascribed to their founding ideology. Through the course of this experiment, I have seen that while the discussions coming from progressive philosophy revolve around the middle class, explicitly or in undertones. To contrast, the conservative opinion often chose to use the words 'american' in various permutations while referring to the populace or used references to 'supreme court' in most of their discussions. An equally apparent overlap is also observable in the phrases used by the think tanks of similar political orientation. At the same time, even think tanks with a high-level of coherence in their ideology depict phrase sets with certain differences. For example, CIS depicts a clear heaviness of slanting towards the rights of immigrants while CAP has a leaning towards phrases related to creating employment opportunities with an emphasis on economic growth.

Hence, these feature sets provide a methodical proof of one-to-one correspondence with the stated ideology of a think tank. This establishes that the articles authored by a think tank suffer from subjectivity which is reflective of their founding principles.

FUTURE WORK

A potential follow-up to this work can be to design an estimate which evaluates the extent of bias prevalent in the articles at any given time. This can possibly be done by designing a metric to evaluate the distance between the vocabulary of a set of articles. Such a metric can be used as a tool to perform a time-series analysis with the objective to understand the evolving polarization of expression of opinion and the relative shift in the focal area of interest with time.

APPENDIX

Table 8 Prominent feature sets by ideology for Multinomial Naive-Bayes classifier

Issue Area	Progressive	Conservative
Immigration	foreign polici, updat econom, tax credit, middl class, islam state, bush administer	suprem court, govern polit, constitut law, health care, obama administer, tax rate
National Security	foreign polici, polici update, health insur, global develop, world bank	govern polit, suprem court, constitut law, obama administer, feder govern
Judiciary	updat econom, health care, african american, social secur, foreign polici	suprem court, govern polit, constitut law, nation secur, presid Obama
Politics	health care, afford care, african country, labor market, great recess	law court, feder govern, econom freedom, homeland secur, health insur
Social Policy	foreign polici, health care, social secur, climat chang, hous market	suprem court, health care, law court, nation secur, obama administr
Monetary Policy	updat econom, health care, foreign polici, hous market, tax polici	suprem court, govern polit, constitut law, nation secur, trade polici
Trade Policy	polici update, foreign polici, health care, climat chang, metropolitan area	suprem court, govern polit, law court, tax budget, feder govern

Table 9 Prominent feature sets by ideology for Linear Support Vector Classifier

Issue Area	Progressive	Conservative
Immigration	bush administer, african american, insur program, govern program, insur coverag	govern polit, develop immigr, institut econom, entitl program, child polici

National Security	bush administer, center health, insur coverag, creat job, africa growth	child polici, center foreign, institut econom, increas feder, entitl program
Judiciary	afford care, bank institute, african american, econom secur, bush administer	develop immigr, law court, child polici, american people, justic depart
Politics	econom secur, care provid, feder fund, iraq war, barack obama	institut econom, child polici, bank monetary, econom freedom, homeland secur
Social Policy	gross domest, care act, econom secur, creat job, african american	intern licens, bank monetary, child polici, center foreign, constitut law
Monetary Policy	care act, earn income, job growth, econom secur, improv quality	develop immigr, govern polit, constitut law, entitl program, defens budget

Table 10 Prominent feature sets by ideology for Multinomial Logistic Regression

Issue Area	**Progressive**	**Conservative**
Immigration	bush administer, polici analyst, care act, tax credit, updat econom	govern polit, constitut law, trade polici, feder spend, econom freedom
National Security	presid barack, presid donald, polici nation, bush administer, immigr polici	institut econom, tax budget, law court, illeg alien, develop immigr
Judiciary	polici update, presid barack, presid donald, incom tax, afford care	law court, tax budget, trade polici, econom freedom, topic foreign
Social Policy	african american, bush administer, care act, incom tax, econom secur	govern polit, constitut law, law court, econom freedom, center foreign
Monetary Policy	updat econom, tax reform, polici update, job growth, world bank	govern polit, constitut law, topic foreign, trade polici, budget polici
Trade Policy	war era, polici update, world bank, barack obama, great recess	budget polici, constitut law, law court, polici studi, econom freedom

REFERENCES

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as Data. *American Economic Association*.

Gentzkow, M., Shapiro, J. M., & Taddy, M. (2019). Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech. *Econometrica*.

Manning, C. M., Schütze, H., & Raghavan, P. (2008). *Introduction to Information Retrieval.*

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space.

Ng, A. Y., & Jordan, M. I. (2002). On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes.

Porter, M. (2006). The Porter Stemming Algorithm.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*.