Twitter Relations in an Election Year

Serim Park
Department of Electrical
& Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
serimp@andrew.cmu.edu

Nicolas Huynh Thien
Department of Electrical
& Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
n@cmu.edu

Laura Tipton
Department of Computational
& Systems Biology
University of Pittsburgh
Pittsburgh, PA 15213
lat60@pitt.edu

Abstract

Social media has recently come to play a large role in American politics, particularly in the 2012 presidential race. Yet, to date, no elections have been successfully predicted based on Twitter data in a reproducible and sound manner [1]. Few previous papers have taken into account the network aspect of the social networking site. Therefore we propose to cluster Twitter users based on their demographic data, keyword tweets, and follower/followee relationships. Once the best method is chosen, it is applied to a set of independent voters in order to predict their voting behavior to see if this is a viable method to predict the presidential election results.

1 Introduction

The notion that an individual is defined by his or her actions is translated into what one updates his/her status to in the world of social media. On Twitter, a strong measure of popularity and affinity is who one follows and is followed by. In the context of the American presidential election landscape, users that follow Barack Obama, Joe Biden, and the Democratic Party are likely going to vote for President Obama while those who follow Mitt Romney, Paul Ryan, and the Republican Party are likely to vote for Governor Romney. Yet this ignores a large portion of the follower/followee structure of Twitter. Additionally, much of the current literature on predicting elections based on Twitter data does not take this structure into account [1]. We therefore propose a method to cluster Twitter users based on a variety of features, including a bag-of-words model of their tweets and who they follow, classify new users based on their relationship and tweet similarities, and determine their political affiliation in the context of the 2012 presidential elections.

2. Dataset

2.1. Data Collection

Using Ruby scripts [3], information about users and their tweets is scraped directly from Twitter feeds through the Twitter API in Ruby. For the training users, only English-speaking users who follow either Obama or Romney were included. To ensure that we are collecting at American Democrats we require that followers of Obama also follow two or more of the following political figures: Joseph Biden, Jon Stewart, Stephen Colbert, Al Gore, or the Democratic Party. Similarly we required that followers of Romney also follow one or more of Paul Ryan, Rush Limbaugh, Bill O'Reilly, Glenn Beck, or the Republican Party. This requirement ensures a certain confidence in applying their political labels.

The training dataset consists of 5614 users, labeled either 0 for Obama (2694 instances) or 1 for Romney (2920 instances). This is larger than the average sample size used by the polling industry (IPSOS uses a sample size of 3,805 voters). The additional sampling effort was made to correct for the use of simple random sampling and the unrepresentative nature of Twitter users compared to the American population at large.

2.1 Data Features

We begin our clustering of Twitter users in much the same way as many others have before us - with demographic data and keyword tweets. Demographic data available on most Twitter users includes gender, US residency, age, and physical location. Additional data specific to the Twitter population contains relationship counts and preferences. Relationships between users makes up a separate section of features for our algorithm that will be discussed later but number of such relationships (followees, followers, and friends) is included the first section of the features. Additionally the number of tweets a user has made, his/her level of activity is included. Preferential choices considered in this section consist of the background color and text color of the account page. Together the demographics, relationship counts, and preferences make up the first section of the features to be used to cluster the initial group of training users. The second section of features to be used in our clustering algorithm includes a histogram of 221 predetermined keywords regarding current political issues. While some words were hand-selected for their inherent controversial aspect, many "buzzwords" were provided by the Global Language Monitor and the New York Times National Conventions word counts [7][8]. To account for the poor spelling and grammar associated with the limited character account of tweets, all keywords are in regular expression (regex) form [2]. By using regex, concerns over capitalization, pluralization, and part of speech are alleviated. Keywords for the 2012 presidential election include "Obamacare", "gun control", "Syria", "occupy movement", and so on. For a complete list of keywords, see appendix A. A word-count histogram is constructed for each user to indicate how many times they have used each keyword in their recent tweets. For one user this histogram looks like the image seen in Figure 1.

.

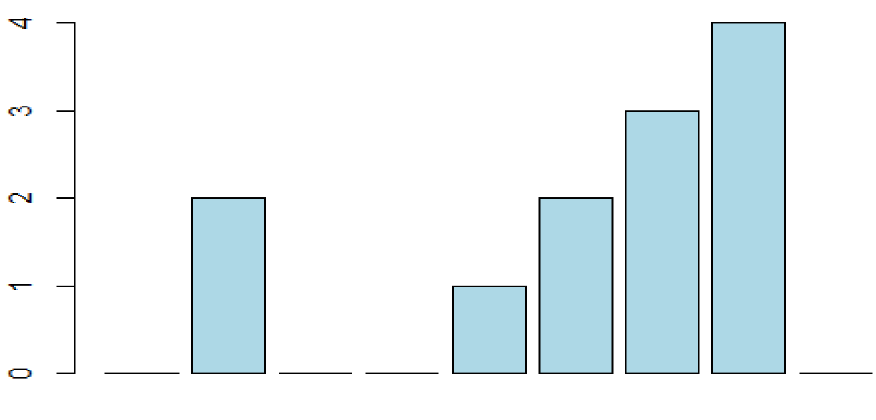

Figure 1 Histogram of some of the keywords in one users tweets. Bain Capital, Black, Iraq, Medicare, Middle Class, Obama, Romney, and Vote are representative of the keywords analyzed but are shown here for illustrative purposes only.

hile this particular instance used a couple of keywords several times, there are many more keywords not used at all, leading to high sparsity in the data. This sparsity is addressed in each model applied to the data. In addition to covering hot button issues, the keywords themselves can be classified into different topics including health, economics, environmental, and political. These categories can either be used to aggregate keywords, creating a higher counts in the histograms and less sparsity in the data, or for segregating the keywords and using only select categories that create the most separation in the initial clustering.
The third, and most interesting, section of features to be used in the clustering is the social network that surrounds the user. This takes the form of a list of all the users they follow (followees), all the users who follow them (followers), and those who are both followees and followers (friends). Some of the followees, particularly key political figures, pundits, and "talking heads" will have a strong correlation with voting preference: users who follow any combination of two or more of Barack Obama, Joe Biden, John Stewart, Stephen Colbert, Al Gore, and the Democratic Party are most likely to vote for Obama and users who follow any combination of two or more of Mitt Romney, Paul Ryan, Rush Limbaugh, Bill O'Reilly, Glenn Beck, and the Republican Party are most likely to vote for Romney. But if you are defined by who you follow on Twitter then the followees of the followers of political leaders should also cluster into Democrats and Republicans. Friendship, defined as a mutual follower/followee relationship, is a very good indicator of similarity between two users.

2. Algorithms

2.1. Clustering

Once all the features are collected and the data is checked for contradictory relationships (for example a political correspondent who follows both Obama and Romney would be removed), we would like to build a classifier on the training data to determine how to split the users into Obama followers and Romney followers.

Before building a binary classifier, the training users are clustered using all data. Using a clustering algorithm (K-means [4] and Spectral Clustering [9]), the process is twofold: 1)extract a meaningful number of clusters by varying K as a parameter such that inter-class variance is greatest, and 2)assign a label to each cluster through a majority vote. This ensures that the process handles a number of clusters greater than 2. Each cluster is confirmed to have been given the correct label, either Obama or Romney, based on who the majority of cluster members follow.

Because K-means performed poorly on this high dimensional data, we moved on to implementing Spectral Clustering [9] on the dataset, with the hope that the affinity matrix could capture more complex data configurations which would have otherwise defeated K-means. Using a difference of squares metric with variance sigma as parameter, the affinity matrix obtained features two distinguishable blocks as shown in Fig 2.

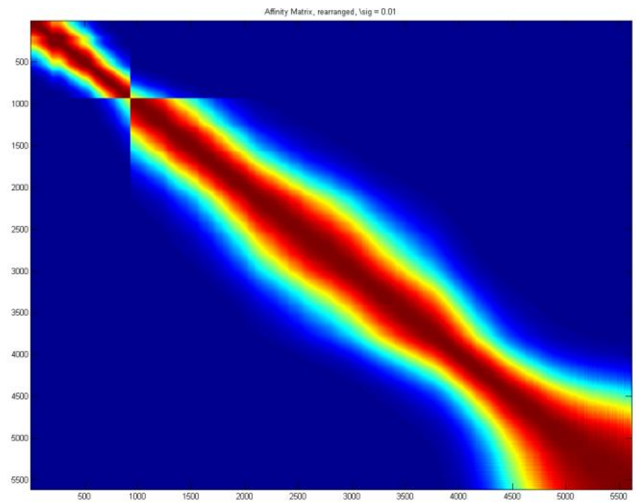

Figure 2. Reordered Affinity matrix with sigma = 0.04

The corresponding clustering is given by the following index vectors in Figure 3:

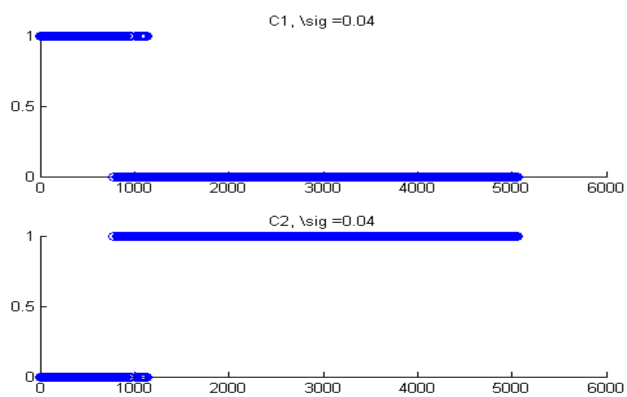

Figure 3. Cluster indexes using Spectral Clustering, K=2.

2.2. Topic Model

When spectral clustering failed to produce high accuracy with a training error of 47% , a topic model was considered [12]. Knowing that a topic model would ignore the social relationships within the data, we hoped to better model the keywords by including a layer of two latent "topics". This was done using a variational expectation-maximization (VEM) algorithm. Upon implementation the topic model algorithm found that the most important word to determine a Romney voter is "oil", while for an Obama voter is is "hipster". This may say more about the skewed demographics of Twitter users than any commentary on the Democratic Party.

2.3 Complete Dataset

In order to use all available data, more diverse methods that are specifically designed to handle sparse data are required. We started with sparse generalized partial least squares regression (SGPLS) for classification [13]. This method applies partial least squares, an alternative to least squares regression meant for highly collinear data, to generalized linear regression. It has been successfully for simultaneous classification and variable selection on highly correlated and collinear datasets such as gene expression arrays. The parameters involved, a mixture parameter eta and k number of groups, were optimized using five-fold cross validation on the training data, as shown in figure 4. The optimal

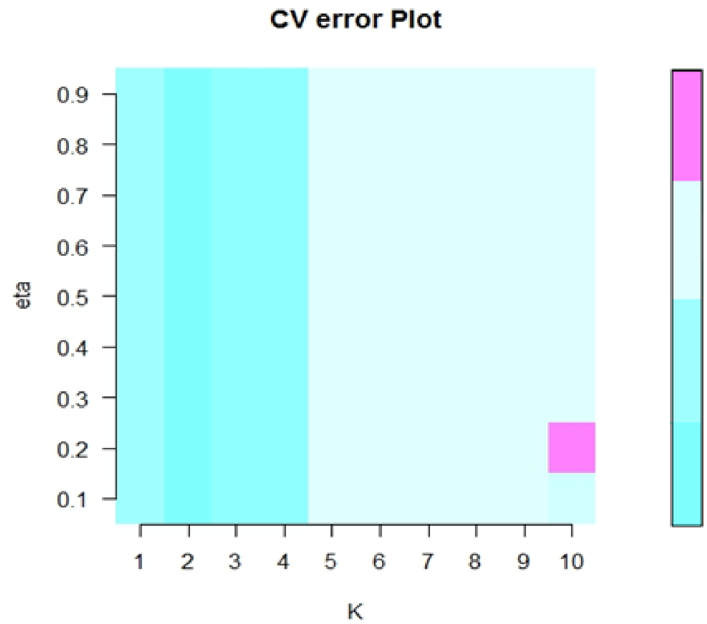

Figure 4. Five-fold cross validation error plot for SGPLS used to find an optimal eta and k.

eta was found to be 0.9 and the optimal k was again found to be 2. After regressing the training data only three variables were selected to have non-zero coefficients: indicators of following two people, and the word count of "terrorism" or "terrorist". The final model predicting a vote for Romney is:

Pr(R) = $0.053 + 0.441X_1 + 0.216X_2 + 0.210T$ where X_1, and X_2 are the indicators of following the first and second influential users, and T is the count of the word terrorism.

Other sparse classifiers applied to the complete data were sparse linear discriminant analysis and sparse mixed discriminant analysis, both with a VEM algorithmic implementation (SLDA - VEM and SMDA) [14]. These models represent adaptations of traditional linear discriminant analysis to deal with the not semi-definite nature of a sparse covariance matrix, with the addition of a Gaussian (rather than linear) boundary for SMDA.

2.4 Dimensionality Reduction

Given the very sparse and high dimensional nature of the data (1783 features per instance), principal component analysis (PCA), penalized linear discriminant analysis (PLDA), and sparse linear discriminant analysis (SLDA) [11] were explored for dimensionality reduction before training the data as well as histogram transformation. These methods have all been used on high-dimensional sparse data [5] with varying results. Using PCA, we reduced dimension to various size to achieve best correct rate of prediction, using PLDA, we reduced the dimension to 50, and using Sparse Linear Discriminant Analysis, we reduced the dimension to 1. After reducing the dimension of our dataset, we built simple nearest neighbor classifier to predict the label on the test set.

SLDA reduces the dimension of the data to 1 dimension, and then applies Fisher criterion to achieve highest between class separability, and uses linear discriminant analysis to predict the data set. Ten-fold cross validation yields 63.63% accuracy. Figure 5 shows the classification result based on SLDA. The red graph shows the ground truth label of the test data, and blue graph shows the

	Obama	Romney
Obama	2694	2042
Romney	0	878

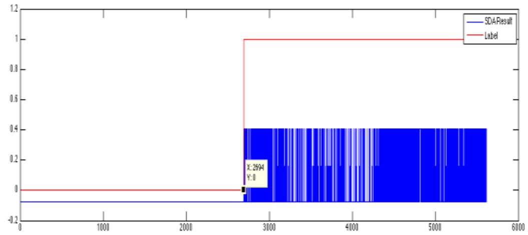

Figure 5. SLDA prediction results

predicted label. Obama followers were labeled with 0 and Romney followers were labeled with 1. As can be seen from the graph, Obama followers are 100% correctly classified, while there are some misses in detecting Romney followers.

2.4.1 Kernel(RBF) SVM

After the 10-fold cross validation to determine for best sigma, the average training correction rate for 10 fold is 99.95% and the average test correction rate results in 53.28%, where SVM on sparse high dimensional dataset gives wrongly biased classifier. As sigma for RBF kernel increases, the correct rate improves in small amount.

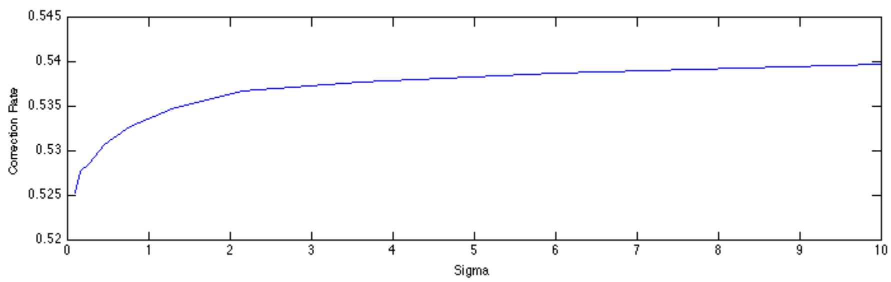

Figure 6. Correct Rate using Kernel SVM classifier vs varied sigma for kernel

2.4.2 Kernel(RBF) SVM + PCA

In order to reduce the large variance in classifier, dimensionality reduction using PCA was explored. By far the most popular dimensionality reduction method, PCA is simple to use and provides fairly good approximation of the original data. Figure 7 shows how the correct rate on training and testing change along the number of principal directions for training the kernel SVM classifier.

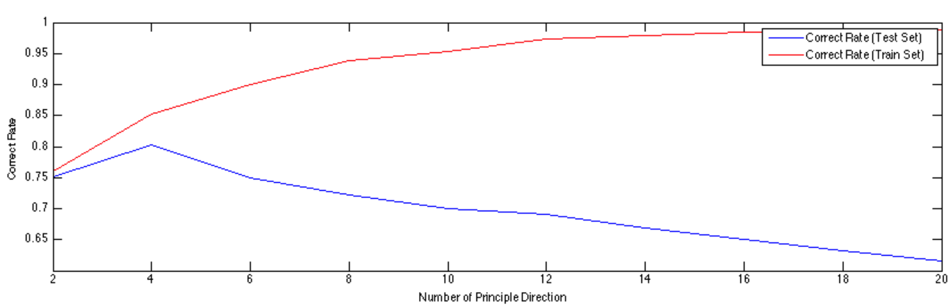

Figure 7. Correct Rate as number of principal components changes using kernel SVM. The red curve is training correct rate and the blue curve shows test correct rate.

Sigma was set to 0.16681 for RBF kernel, which is determined in previous k-fold. Highest test correct rate is achieved at 80.32% with 4 PCs, and it is interesting to note that as the testing correct rate goes down, the training correction rate increases as PCA direction size increases.

2.4.3 PLDA

Penalized LDA originally arose for problems associated with small sample size. Since our data set has such high dimension, it can be judged that we have a small sample size compared to the dimension of the features in our data. Data is reduced with 50 main PLDA directions (which gives highest class separability) and alpha = 0.0029471 is determined with 10-fold cross validation. Figure 8 shows the projection of data into 2 main plda direction for easy visualization. Cleary, two sets are discriminant

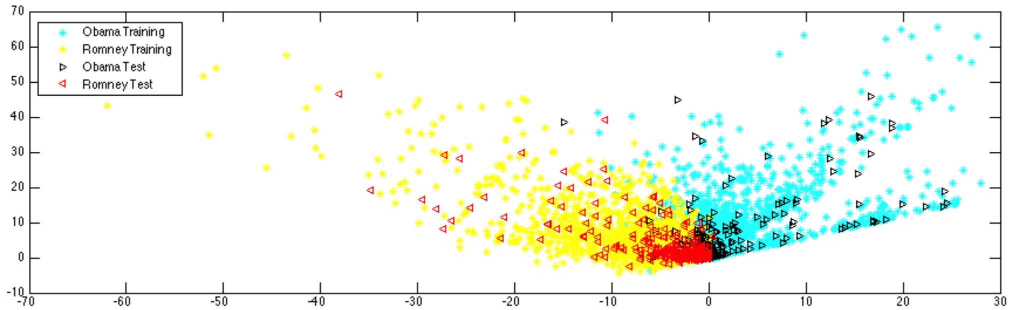

Figure 8. Projection of dataset into 2 main plda direction. Yellow and red dots represents each training and testing set for Romney followers, and cyan and black dots represents each training and testing set for Obama followers.

although there is some overlaps. Average prediction correct rate is 90.15% with optimum alpha.

3. Results

The table below presents the error rates on the training and testing sets (with a 90|10% split) for each model applied to the data.

Algorithm	Training Error	Testing Error
Spectral Clustering	47%	46%
Topic Model	45%	45%
SGPLS	27%	28%
SLDA - VEM	11%	15%
SMDA	52%	45%
SLDA		36%
PLDA		10%

The next table shows the resulting voter distribution on a pool of 1889 unlabeled users following the *LA Times*. For comparison purposes, the national results for the 2012 US Presidential Elections are included in bold.

	Obama Voters	Romney Voter
Actual Popular Vote	**50.9%**	**47.3%**
SLDA - VEM	81.1%	17.9%
PLDA	53.0%	47.1%

We can see that PLDA approximates the national results most closely. Some of the discrepancy (particularly for Obama) stems from the fact that this framework assumes a binary classification. Consequently, voter abstention (which accounts for 1.8% of the US population) is not reflected.

Additional inflation of the Obama voters may be due to the younger population of Twitter compared to the US population.

4. Conclusions

Both the high dimensionality and sparsity of the data represented major issues in our classification framework. Reducing the feature dimension led to better classification. While dimensionality reduction produced acceptable results reflecting the outcome of the 2012 presidential elections, a few limitations of our approach must be put forward. The dataset of 5414 Twitter users was collected at one point in time, and as such does not reflect the dynamics of the social network landscape which characterizes an presidential election period. Hence, any event - a gaffe, a scandal - which may result in the sudden shift of political affinities could not be captured in the above framework. It is possible to remedy this drawback by introducing incidence matrices [10], which can successfully capture edge insertions/deletions in a social graph; adding time series could potentially improve our model.

Additionally our data treated the relationships between users as individual data points rather than a unified graph. New research in the area of combining topic models with graphical models in the form of MMS models would be an ideal next step in the evaluation of this type of data. As this field of research is maturing, it seems poised to be used to analyze Twitter data to predict the 2016 election and beyond.

References

[1] Gayo-Avello, D. (2012) “I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper” -- A Balanced Survey on Election Prediction using Twitter Data. arXiv:1204.6441 [cs.CY].

[2] Goyvaerts, Jan. Regular-Expressions.info: The Premier Website about Regular Expressions. Regular-Expressions.info. N.p., 2003. Web. 04 Nov. 2012. <http://www.regular-expressions.info/>.

[4] MacQueen, J. (1967) Some Methods for Classification and Analysis of Multivariate Observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol 1, pp 281-297.

[5] Xiong, T. (2007) Classification Methods for High-Dimensional Sparse Data. University of Minnesota <http://dl.acm.org/citation.cfm?id=1292936>.

[6] Cortes, C., Vapnik, V. (1995) Support Vector Networks. Machine Learning, vol 20, num 3, pp 273-297.

[7] Political BuzzWord, The Global Language Monitor, <http://www.languagemonitor.com/?cat=169>

[8] At The National Conventions, The Words They Used, The New York Times, <http://www.nytimes.com/interactive/2012/09/06/us/politics/convention-word-counts.html>

[9] Jianbo Shi and Jitendra Malik, "Normalized Cuts and Image Segmentation", IEEE Transactions on PAMI, Vol. 22, No. 8, Aug 2000.

[10] Ning, Xu, Chi, Gong, Huang (2007). Incremental Spectral Clustering With Application to Monitoring of Evolving Blog Communities. Proceedings of the Society for Industrial and Applied Mathematics.

[11] Improving Implementation of Linear Discriminant Analysis for the High Dimension/Small Sample Size Problem", tebbens, Schlesinger, published in Computational Statistics and Data Analysis, vol. 52, no.1, pp. 423--437, 2007

[12] Grun B, Hornik K (2011). “topicmodels: An R Package for Fitting Topic Models." Journal of Statistical Software, 40(13), 1-30.

[13] Chun, H., & Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(1), 3-25.

[14] Clemmensen, L., Hastie, T., Witten, D., Ersbøll, B., (2011). *Sparse discriminant analysis*, Technometrics, 53(4): 406-413.

Other Resources

Bishop, C. (2006) Pattern Recognition and Machine Learning. Springer.

Ruby Programming Language. N.p., n.d. Web. 04 Nov. 2012. <http://www.ruby-lang.org/en/>.

JSON. <http://www.json.org/>.

Mathworks. Matlab vR2012a.

R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Twitter API. <https://dev.twitter.com/>.

Appendix A: List of Buzzwords

<table>
<tr><th>Buzzword</th><th>Category</th></tr>
<tr><td>/doom(ed)?/i,</td><td>War</td></tr>
<tr><td>/strikes?/i,</td><td>War</td></tr>
<tr><td>/veterans?/i,</td><td>War</td></tr>
<tr><td>/drones?/i,</td><td>War</td></tr>
<tr><td>/troops?/i,</td><td>War</td></tr>
<tr><td>/pull(|-)out/i,</td><td>War</td></tr>
<tr><td>/casualt(y|ies)/,</td><td>War</td></tr>
<tr><td>/benghazi/i,</td><td>War</td></tr>
<tr><td>/guns?/i,</td><td>War</td></tr>
<tr><td>/\bNRA\b/i,</td><td>War</td></tr>
<tr><td>/politics?/i,</td><td>Politics</td></tr>
<tr><td>/red state/i,</td><td>Politics</td></tr>
<tr><td>/second(|)term/i,</td><td>Politics</td></tr>
<tr><td>/partisans?/i,</td><td>Politics</td></tr>
<tr><td>/buffet(|)rule/i,</td><td>Politics</td></tr>
<tr><td>/liberals?/i,</td><td>Politics</td></tr>
<tr><td>/romney/i,</td><td>Politics</td></tr>
<tr><td>/senator/i,</td><td>Politics</td></tr>
<tr><td>/keystone/i,</td><td>Politics</td></tr>
<tr><td>/illegal(|)aliens?/i,</td><td>Politics</td></tr>
<tr><td>/immigrants?/i,</td><td>Politics</td></tr>
<tr><td>/politics of fears?/i,</td><td>Politics</td></tr>
<tr><td>/unions?/i,</td><td>Politics</td></tr>
<tr><td>/obamacare/i,</td><td>Politics</td></tr>
<tr><td>/american(|)dream/i,</td><td>Politics</td></tr>
<tr><td>/obama(|)administration/i,</td><td>Politics</td></tr>
<tr><td>/opponents?/i,</td><td>Politics</td></tr>
<tr><td>/transparency/i,</td><td>Politics</td></tr>
<tr><td>/elections?/i,</td><td>Politics</td></tr>
<tr><td>/romney(|)wealth/i,</td><td>Politics</td></tr>
<tr><td>/obamamania/i,</td><td>Politics</td></tr>
<tr><td>/bush/i,</td><td>Politics</td></tr>
<tr><td>/progressive(|)politics?/i,</td><td>Politics</td></tr>
<tr><td>/stalement/i,</td><td>Politics</td></tr>
</table>

/obstructionist(s)?/i,	Politics
/congress/i,	Politics
/bipartisans?/i,	Politics
/tea(|)party/i,	Politics
/obstructionists?/i,	Politics
/congress/i,	Politics
/bipartisans?/i,	Politics
/tea(|)party/i,	Politics
/gun(|)control/i,	Politics
/tcot\b/i,	Politics
/ocra\b/i,	Politics
/tlot\b/i,	Politics
/sgp\b/i,	Politics
/hhrs\b/i,	Politics
/gop\b/i,	Politics
/conservatives?/i,	Politics
/republicans?/i,	Politics
/democrats?/i,	Politics
/super(|)pacs?/i,	Politics
/economy/i,	Economy
/billionaires?/i,	Economy
/social(|)security/i,	Economy
/medicare/i,	Economy
/medicaid/i,	Economy
/middle(|)class/i,	Economy
/balanced(|)budget/i,	Economy
/jobs?/i,	Economy
/deficits?/i,	Economy
/uS(|)economy/i,	Economy
/the(|)(1%|one percent)/i,	Economy
/healthcare/i,	Economy
/manufacturing/i,	Economy
/auto industry/i,	Economy
/tax cuts/i,	Economy
/auto(|)industry/i,	Economy
/outsourcing/i,	Economy

<table>
<tr><td>/affordable(|)care(|)act/i,</td><td>Economy</td></tr>
<tr><td>/affordabe(|)healthcare/i,</td><td>Economy</td></tr>
<tr><td>/wall(|)(street|st)/i,</td><td>Economy</td></tr>
<tr><td>/occupy(|)(movement|mvts?)/i,</td><td>Economy</td></tr>
<tr><td>/bain(|)capital/i,</td><td>Economy</td></tr>
<tr><td>/tax/i,</td><td>Economy</td></tr>
<tr><td>/recessions?/i,</td><td>Economy</td></tr>
<tr><td>/stimulus/i,</td><td>Economy</td></tr>
<tr><td>/bailout/i,</td><td>Economy</td></tr>
<tr><td>/crisis/i,</td><td>Economy</td></tr>
<tr><td>/unemploy(ment|ed)/i,</td><td>Economy</td></tr>
<tr><td>/middle(|)class/i,</td><td>Economy</td></tr>
<tr><td>/spending/i,</td><td>Economy</td></tr>
<tr><td>/westboro/i,</td><td>Religion</td></tr>
<tr><td>/atheist(s)?/i,</td><td>Religion</td></tr>
<tr><td>/agnostic(s)?/i,</td><td>Religion</td></tr>
<tr><td>/humanist(s)?/i,</td><td>Religion</td></tr>
<tr><td>/secular?/i,</td><td>Religion</td></tr>
<tr><td>/choice/i,</td><td>Religion</td></tr>
<tr><td>/Creationis(m|t)/i,</td><td>Religion</td></tr>
<tr><td>/darwin/i,</td><td>Religion</td></tr>
<tr><td>/gay(|)marriage/i,</td><td>Religion</td></tr>
<tr><td>/tolerance/i,</td><td>Religion</td></tr>
<tr><td>/lgbt/i,</td><td>Religion</td></tr>
<tr><td>/angry whites?/i,</td><td>Religion</td></tr>
<tr><td>/pregnanc(y|ies)/i,</td><td>Religion</td></tr>
<tr><td>/mormon(s|ism)?/i,</td><td>Religion</td></tr>
<tr><td>/christians?/i,</td><td>Religion</td></tr>
<tr><td>/catholics?/i,</td><td>Religion</td></tr>
<tr><td>/evangelists?/i,</td><td>Religion</td></tr>
<tr><td>/methodists?/i,</td><td>Religion</td></tr>
<tr><td>/baptists?/i,</td><td>Religion</td></tr>
<tr><td>/prayers?/i,</td><td>Religion</td></tr>
<tr><td>/church(es)?/i,</td><td>Religion</td></tr>
<tr><td>/\bFSM\b/i,</td><td>Religion</td></tr>
</table>

<table>
<tr><td>/bible/i,</td><td>Religion</td></tr>
<tr><td>/gods?/i,</td><td>Religion</td></tr>
<tr><td>/christ/i,</td><td>Religion</td></tr>
<tr><td>/lord/i,</td><td>Religion</td></tr>
<tr><td>/\bsin(s)?\b/i,</td><td>Religion</td></tr>
<tr><td>/mosques?/i,</td><td>Religion</td></tr>
<tr><td>/ayatollah?/i,</td><td>Religion</td></tr>
<tr><td>/allah/i,</td><td>Religion</td></tr>
<tr><td>/qurans?/i,</td><td>Religion</td></tr>
<tr><td>/burqa/i,</td><td>Religion</td></tr>
<tr><td>/niqab/i,</td><td>Religion</td></tr>
<tr><td>/hijab/i,</td><td>Religion</td></tr>
<tr><td>/synagogues?/i,</td><td>Religion</td></tr>
<tr><td>/rabbis?/i,</td><td>Religion</td></tr>
<tr><td>/shabbat/i,</td><td>Religion</td></tr>
<tr><td>/(yarmulkes?)|(yamakas?)|(kippa(hs?)?)/i,</td><td>Religion</td></tr>
<tr><td>/contraceptives?/i,</td><td>Fem. Rights</td></tr>
<tr><td>/war(|)on(|)women/i,</td><td>Fem. Rights</td></tr>
<tr><td>/abortions?/i,</td><td>Fem. Rights</td></tr>
<tr><td>/planned(|)parenthood/i,</td><td>Fem. Rights</td></tr>
<tr><td>/gun(|)control/i,</td><td>Fem. Rights</td></tr>
<tr><td>/voter(|)fraud/i,</td><td>Conspiracies</td></tr>
<tr><td>/voter(|)suppression/i,</td><td>Conspiracies</td></tr>
<tr><td>/birthers?/i,</td><td>Conspiracies</td></tr>
<tr><td>/voter(s)?/i,</td><td>Conspiracies</td></tr>
<tr><td>/big(|)gov(ernment|t|)/i,</td><td>Conspiracies</td></tr>
<tr><td>/big(|)money/i,</td><td>Conspiracies</td></tr>
<tr><td>/fracking/i,</td><td>Environmental</td></tr>
<tr><td>/hydraulic(|)fracturing/i,</td><td>Environmental</td></tr>
<tr><td>/recycl(e|ing)/i,</td><td>Environmental</td></tr>
<tr><td>/climate/i,</td><td>Environmental</td></tr>
<tr><td>/global(|)warming/i,</td><td>Environmental</td></tr>
<tr><td>/wwf/i,</td><td>Environmental</td></tr>
<tr><td>/greenpeace/i,</td><td>Environmental</td></tr>
<tr><td>/conservation/i,</td><td>Environmental</td></tr>
</table>

/eco/i,	Environmental
/natural/i,	Environmental
/oil/i,	Environmental
/coal/i,	Environmental
/47%/i,	Environmental
/clint/i,	Pundits
/bloomberg/i,	Pundits
/cheney/i,	Pundits
/chris christie/i,	Pundits
/powell/i,	Pundits
/warren(|)buffet/i,	Pundits
/bush/i,	Pundits
/al gore/i,	Pundits
/biden/i,	Pundits
/paul ryan/i,	Pundits
/clinton/i,	Pundits
/coulter/i,	Pundits
/palin/i,	Pundits
/glenn(|)beck/i,	Pundits
/limbaugh/i,	Pundits
/o('|)reilly/i,	Pundits
/fox/i,	Pundits
/assange/i,	Pundits
/anonymous/i,	Pundits
/white(s)?/i,	Slang
/beaners?/i,	Slang
/chinks?/i,	Slang
/camel jockeys?/i,	Slang
/fag(|s|gots?)/i,	Slang
/dollaz?/i,	Slang
/prez/i,	Slang
/douche(bag(s)?)?/i,	Slang
/cracker(s)?/i,	Slang
/\bho(es)?\b/i,	Slang
/\bvatos?\b/i,	Slang
/motherfuckin(g)?/i,	Slang

<table>
<tr><td>/niggas?/i,</td><td>Slang</td></tr>
<tr><td>/niggers?/i,</td><td>Slang</td></tr>
<tr><td>/weed/i,</td><td>Slang</td></tr>
<tr><td>/bitch(es)?/i,</td><td>Slang</td></tr>
<tr><td>/bitch(|)ass/i,</td><td>Slang</td></tr>
<tr><td>/puss(y|ies)/i,</td><td>Slang</td></tr>
<tr><td>/dawg(s)?/i,</td><td>Slang</td></tr>
<tr><td>/fuck(|s|ed|ing)?/i,</td><td>Slang</td></tr>
<tr><td>/minority/i,</td><td>Slang</td></tr>
<tr><td>/black(s)?/i,</td><td>Slang</td></tr>
<tr><td>/brotha(s)?/i,</td><td>Slang</td></tr>
<tr><td>/\b[Sh]it\b/i,</td><td>Slang</td></tr>
<tr><td>/ass(es)?/i,</td><td>Slang</td></tr>
<tr><td>/hipster(s)?/i,</td><td>Trends</td></tr>
<tr><td>/vote(s)?/i,</td><td>Trends</td></tr>
<tr><td>/get(|)out(|)the(|)vote/i,</td><td>Trends</td></tr>
<tr><td>/swag/i,</td><td>Trends</td></tr>
<tr><td>/yolo/i,</td><td>Trends</td></tr>
<tr><td>/fixie(s)?/,</td><td>Trends</td></tr>
<tr><td>/\bapple\b/i,</td><td>Trends</td></tr>
<tr><td>/ip(a|o)d(s)?/i,</td><td>Trends</td></tr>
<tr><td>/iphone(s)?/i,</td><td>Trends</td></tr>
<tr><td>/android(s)?/i,</td><td>Trends</td></tr>
<tr><td>/samsung(| nexus| galaxy)/i,</td><td>Trends</td></tr>
<tr><td>/degree(s)?/i,</td><td>Education</td></tr>
<tr><td>/college(s)?/i,</td><td>Education</td></tr>
<tr><td>/student(|)loans/i,</td><td>Education</td></tr>
<tr><td>/student(|)debt/i,</td><td>Education</td></tr>
<tr><td>/teachers?(|)unions?/i,</td><td>Education</td></tr>
<tr><td>/teachers?/i,</td><td>Education</td></tr>
<tr><td>/universit(y|ies)/i,</td><td>Education</td></tr>
<tr><td>/science(s)?/i,</td><td>Education</td></tr>
<tr><td>/\bNASA\b/i,</td><td>Education</td></tr>
<tr><td>/tuition/i,</td><td>Education</td></tr>
<tr><td>/drop(|-)out(s)?/i,</td><td>Education</td></tr>
<tr><td>/palestin(e|inians?)/i,</td><td>International</td></tr>
</table>

<table>
<tr><td>/america/i,</td><td>International</td></tr>
<tr><td>/wikileaks?/i,</td><td>International</td></tr>
<tr><td>/\bhamas\b/i,</td><td>International</td></tr>
<tr><td>/gaza/i,</td><td>International</td></tr>
<tr><td>/bombs?/i,</td><td>International</td></tr>
<tr><td>/suicide/i,</td><td>International</td></tr>
<tr><td>/israel(|is?)?/i,</td><td>International</td></tr>
<tr><td>/terroris(m|ts?)/i,</td><td>International</td></tr>
<tr><td>/nuclear(|)iran/i,</td><td>International</td></tr>
<tr><td>/euro(|)crisis/i,</td><td>International</td></tr>
<tr><td>/\biraq\b/i,</td><td>International</td></tr>
<tr><td>/afghanistan/i,</td><td>International</td></tr>
<tr><td>/rise of china/i,</td><td>International</td></tr>
<tr><td>/sandy/i</td><td>International</td></tr>
</table>