Modeling Ideology and Predicting Policy Change with Social Media: Case of Same-Sex Marriage

Amy X. Zhang[1,2]
axz@mit.edu

[1]MIT CSAIL
Cambridge, MA 02139, USA

Scott Counts[2]
counts@microsoft.com

[2]Microsoft Research
Redmond, WA 98052, USA

ABSTRACT

Social media has emerged as a prominent platform where people can express their feelings about social and political issues of our time. We study the many voices discussing an issue within a constituency and how they reflect ideology and may signal the outcome of important policy decisions. Focusing on the issue of same-sex marriage legalization, we examine almost 2 million public Twitter posts related to same-sex marriage in the U.S. states over the course of 4 years starting from 2011. Among other findings, we find evidence of moral culture wars between ideologies and show that constituencies that express higher levels of emotion and have fewer actively engaged participants often precede legalization efforts that fail. From our measures, we build statistical models to predict the outcome of potential policy changes, with our best model achieving 87% accuracy. We also achieve accuracies of 70%, comparable to public opinion surveys, many months before a policy decision. We discuss how these analyses can augment traditional political science techniques as well as assist activists and policy analysts in understanding discussions on important issues at a population scale.

Author Keywords

political science; public policy; same-sex marriage; social media

ACM Classification Keywords

H.5.3. Group and Organization Interfaces: Asynchronous interaction; Web-based interaction

INTRODUCTION

From the years 2011 to 2014, over 50 pieces of legislation, court cases, and popular votes were contested in relation to same-sex marriage legalization in states across the U.S. In some states, radical changes in policy resulted, reversing decades-old legislation outlawing same-sex marriage, while in other states, policymakers halted any potential policy changes. What drove these different policy outcomes? One primary factor is the changing views and prevailing opinions of a policymaker's constituency [22].

CHI 2015, April 18 - 23 2015, Seoul, Republic of Korea
Copyright 2015 ACM 978-1-4503-3145-6/15/04 $15.00
http://dx.doi.org/10.1145/2702123.2702193

Within the past four years, national polls have shown a dramatic shift in public opinion so that same-sex marriage now has majority support. In terms of how these shifts translate to policy change, a central tenet of representative democracies is that elected officials will faithfully carry out the desires of their electorate. Historical evidence also demonstrates that politicians [14] and judges [26] respond by changing stances as their constituency's viewpoint changes. Given the intense policy battles within many states in recent years, the examination of state constituencies presents a unique opportunity to study the markers of population-scale ideological change and policy responsiveness on a contentious and timely issue.

In a broader historical context, society has always undergone significant shifts (e.g., women's suffrage, the civil rights movement). As the gay rights movement shifts today's society, we aim to demonstrate in this paper the multifaceted, nuanced, and real-time understanding of a constituency's ideology and its relation to policy change that can be found by examining millions of discussions unfolding on social media. This analysis can also augment opinion polling, a tool that has seen pervasive use in politics since the 1950s [12].

By virtue of the richness of social media data, we extract five categories of measures to characterize constituencies' opinions and feelings on the issue of same-sex marriage, including morality, personality, emotional expression, certainty, and engagement. We use these measures to cluster states into ideologically similar groups, and track the changes in states' ideologies over time. Further, we show how these measures can characterize different state populations leading up to important policy decisions. For example, we find that constituencies before policies that pass generally have higher levels of engagement with the issue, have lower levels of emotion, and morally frame the issue in terms of fairness.

We then use our measures to examine the link between prior constituency opinion and the outcome of potential policy changes. We find that we can predict with approximately 80% accuracy whether a potential policy change will pass given features taken from prior social media posts within the state, and that this method performs better than using polling data. Of our measures, we find that level of engagement, emotional expression, and moral framing are the most predictive of policy change. We also improve our results to 87% when incorporating the influence of other states' populations, with geographically closer states providing more sway.

We believe this research can be leveraged in a variety of ways in the areas of political science and public policy. First, by

being able to capture the underlying moral values and other characteristics of a population, we can build applications better suited for provoking discussion or improving mutual understanding and civility. Second, the ability to predict policy change and pinpoint specific ideological groups provides actionable information to the public, policymakers, and activists to tailor and direct their resources and messages. More broadly, we highlight how social media analysis can be a powerful tool to understand the interplay between public policies and the people they affect.

BACKGROUND AND RELATED WORK

In terms of computational approaches to political science questions, much research focuses on classification of political positions from textual data such as speeches [21, 31]. Other research makes use of social data, such as social annotation [34] or social network features [7]. Prior work on analyzing social media text gained insights similar to public opinion surveys, but primarily used measures such as sentiment [29] and volume [25]. However, most computational research in this area focuses on predicting election outcomes [32, 11] or political orientation [6, 7], while we specifically address characterizing and predicting policy change. There are facets to this problem that make it different from existing problems explored. For instance, while many elections are decided by popular vote, the link between policy change and constituents is less direct and has many factors. Some factors include time to the next election, the national party line, the personal ideology of the policymaker, and interest group influence [23]. A major factor that we focus on is the ideology of the population that a policymaker represents [14, 22]. Most political science research demonstrates this connection through public opinion surveys, qualitative interviews, or proxies for opinion such as demographic data. We build on this work by using social media expressions to characterize a constituency's ideology.

With respect to understanding constituency ideology, political psychology studies have found evidence that people on different sides of the ideological spectrum have different preferences for a host of values [17]. For instance, ideology has been linked to different personality traits. Of the Big Five personality traits from psychology research, *openness to experience* has been found to be higher for liberals while *conscientiousness* is higher for conservatives [4]. A right or left leaning ideology has also been correlated with different moral frames, such as loyalty and respect or fairness and compassion, respectively [20]. To extract these measures from social media, we leverage commonly used lexicons, including the Linguistic Inquiry and Word Count (LIWC) software [30]. Many categories on LIWC have been scientifically validated as performing well on Internet language and short text such as Twitter [8] to understand large populations. To then organize these measures, we draw on prior work that applies framing analysis to gain insight into controversial issues [5, 9].

Few research exists that attempt to capture nuanced features of ideology, such as morality, in social media text. Related work on capturing political orientation uses techniques such as examining the follow graph of politicians [7], measuring volume, sentiment, or mood [3], or looking for explicit for/against statements regarding an issue [15]. We believe our approach allows us to capture greater nuance in text and to characterize a larger volume of data, improving accuracy over other approaches. Also by focusing on a set of validated measures as opposed to using bag-of-words or topic modeling, we can more easily interpret our findings and potentially generalize to different issues.

Term	Count	Term	Count
marriage+gay	588644	married+gay	92319
marriage+equal	181677	#noh8	81585
marriage+state	145929	marriage+man+woman	61020
marriage+same+sex	138667	marry+gay	57175
marriage+right	96483	doma	50480

Table 1. Top search terms for same-sex marriage

DATA

We begin by discussing our method for gathering constituency discussions about same-sex marriage, focusing at the state level. We chose to work with Twitter data because it is public, provides free-text personal and emotional expression, and also contains important metadata such as time and location. Instead of looking for explicit pro/con declarations about same-sex marriage, which would be quite sparse, we chose to collect all messages related to same-sex marriage and then study the implicit framing used.

Twitter Dataset

From a qualitative examination of Twitter posts, community wikis, news, and other discussion about same-sex marriage, we manually built a set of key terms, phrases, and hashtags related to same-sex marriage. The most popular ones from our dataset are shown in Table 1. We took care to include search terms that would capture rhetoric on opposite sides of the discussion by consulting Twitter accounts and websites that were both for and against same-sex marriage.

We then searched for occurrences of these items within posts from the Twitter Firehose, a dataset of all public posts from Twitter made available to us through an agreement with Twitter, between January 1, 2011 and June 30, 2014. We focused on this time frame as many state-level same-sex marriage policies were decided during this time. We eliminated any retweets from our dataset as these posts were originally posted by another Twitter account, and we were concerned about over-representing particular terminology. We also eliminated any posts containing hyperlinks, as we were interested in expressions of opinions and feelings, and many of these posts were simply reporting events or quoting news. While this strategy may have removed some relevant Twitter posts, we were primarily concerned with maintaining a high level of precision. Finally, we manually went over a random subset of the posts to find common misclassifications (e.g., posts containing "child marriage" with "right" or "state") and purged the dataset of them. We found 8 phrases of this kind in total.

We evaluated our dataset using crowd workers recruited through Amazon's Mechanical Turk. We gathered a random sample of 1000 posts from across our entire dataset and showed each post to 3 separate Master Workers who had a minimum 95% approval rating, English language proficiency,

Name	State	Date	Outcome	Short Description
Donaldson v. State of Montana	Montana	12-17-2012	Fail	Supreme Court rules 4-3 that a same-sex marriage ban was not unconstitutional
Senate Bill 172	Colorado	05-15-2012	Fail	House Committee kills a bill 6-5 legalizing civil unions after public hearing
Hawaii Marriage Equality Act	Hawaii	11-13-2013	Pass	Governor signs bill legalizing same-sex marriage the day after passing Senate
Griego v. Oliver	New Mexico	12-19-2013	Pass	Supreme Court unanimously rules in favor of legalizing same-sex marriage

Table 2. Examples of final policy decisions related to same-sex marriage at the state level.

and familiarity with Twitter. Workers categorized whether the post they saw was related to same-sex marriage. In the end, 87.8% of posts were categorized as relating to same-sex marriage when using the majority category out of the 3 votes, with a different worker providing each vote. Of these posts related to same-sex marriage, 12.6% had one dissenting opinion, while the rest had unanimous agreement. Many of the tweets coded as unrelated on inspection were due to lack of knowledge of specific terminology, sometimes Twitter-specific, related to same-sex marriage. Others were difficult to interpret due to ambiguous language.

Geographically tagging posts at the state level

Next, we geographically tagged the posts to a particular U.S. state. Prior research has found it is possible to tag posts to the state or city level using manually constructed dictionaries and matching them to a Twitter user's profile location field [28]. This method yields far more geographically-tagged posts and may be less biased overall than using posts that have an associated latitude and longitude [19]. The dictionaries we constructed for each state consisted of the state name, the state postal code preceded by a comma, the names of the top 5 cities within each state, and the capital of the state. We also found the top 100 cities in the U.S. by population and added them if they were not included already. For cities with duplicate names, we associated a city to a particular state if its metropolitan area was greater than two times the population of the other state's metropolitan area. If there was not one city that was much larger than the other, we removed both cities from our location dictionary. We also removed cities with duplicate names outside the U.S. that were in the top 200 most populous cities in the world. Finally, we manually added common nicknames of states informed by the most frequent location field values from our dataset that were not tagged. Comparing our post volume tagged to each state from 2011 to 2014 and population counts from each state from the 2013 U.S. Census, we found a strong correlation (ρ=0.904, $p<.0001$). In total, we had 1.84 million posts related to same-sex marriage and tagged to U.S. states.

Policy Event Dataset

We built a dataset of legislative and judicial events related to same-sex marriage legalization that occurred at the state level between 2011 and mid-2014. Using different news articles and data from state proceedings, we first manually compiled a list of legislative documents and judicial court cases related to same-sex marriage policy for each state. This included items about same-sex marriage, civil unions, domestic partnerships, or any other policy that dealt with the legal representation of same-sex couples. We then determined the date of the event that produced a final decision, *pass* or *fail*, for that policy. We consider a passing legislative policy as one in which a bill gets voted into law, while a passing judicial policy is one in which the court rules in favor of the prosecution. In total, we had 46 events separated into 28 policies that passed and 18 policies that failed; we show a sample in Table 2.

Generally, there are many events that happen in succession for a single law or case, such as a house vote followed by a senate vote. Only the final, pivotal event counted as an event that we considered, as this event determines whether a policy change will occur or not. In cases when a final decision has not been determined as of the time of this work, we separate those events out as undecided. For instance, a judicial ruling followed by a stay has not reached a final decision nor has it affected policy yet. We found 12 events of this nature and do not include them. We also did not include policies that were *against* same-sex marriage legalization, such as bills seeking to amend the constitution to ban same-sex marriage. There were few of these in our time period, and it was unclear how our measures would need to be recalibrated to properly reflect opinions on pro versus anti same-sex marriage policies; we consider this for future work. For each policy, we recorded what date the final event occurred, the state that the event impacted, and the outcome.

MEASUREMENTS

Our goal is to paint a multifaceted picture about what is happening within a constituency leading up to a potential policy change. We calculate the following measures for each state in the U.S. from our Twitter Dataset. Because many events related to same-sex marriage touched the nation, such as the repeal of the Defense of Marriage Act, we normalize our data to isolate what is happening within a state by subtracting without-state (all states other than the target state) measures from within-state measures. Thus each state's measures are normalized against the national average.

First, we are interested in understanding the *ideological* makeup of a population and how that changes over time using the following measure categories:

Morality: As discussed earlier, research has shown that people of different ideologies often employ different moral judgements. To measure this, we collect the occurrence of terms related to the five major categories of *harm*, *fairness*, *purity*, *ingroup*, and *authority* using the supplemental LIWC dictionaries developed by Graham et al. [17]. Table 3 lists examples of posts that demonstrate each of the five categories. Generally, *harm* and *fairness* has been found to be emphasized more by liberals while the remaining three are more emphasized by conservatives. Given that same-sex marriage is one among many issues that are religiously charged, we also measure the prevalence of *religion* terms using LIWC [30].

Personality: Research has also found that the Big 5 Personality Traits of *openness* and *conscientiousness* correlate with

Moral Foundation	Twitter Post
Harm	*If you're #LGBT & **hurting** because of **cruelty** & bigotry please know SO MANY of us **FIGHT** for your rights & love you #NoH8*
Fairness	*#LegalizeGayMarriage because everyone deserves to be treated **equally** and nobody should be **discriminated** by their sex*
Purity	*I believe a marriage is meant to be a **sacred** unit between man and woman. #judgeme*
Ingroup	*"Twitter Me This",why would Obama say: "Gay marriage doesn't weaken **families**, it strengthens **families**".It has done the opposite in **family's**!*
Authority	*well then I guess the gays need to establish some **tradition** of their own bc marriage isn't something that is going to change.*

Table 3. Posts expressing opinions related to each of the 5 Moral Foundation categories with dictionary terms bolded.

ideology [4]. We use research that finds correlations between LIWC categories and personality traits [33] to build a measure for each trait. The measure takes the frequency of each LIWC category weighted by their correlation with the trait and combines them linearly. We obtained the best results when we set a cutoff of greater than 0.2 correlation, either positive or negative. As *conscientiousness* has no LIWC categories that correlate above 0.2, we do not include it and only measure *openness*.

We also collect the following measures that further contextualize the changes that may be happening within a constituency on this issue.

Emotionality and Sentiment: We are interested in the emotions people use in conjunction with expressions on same-sex marriage. Previous research has shown that measuring sentiment on Twitter using lexicons correlate with public opinion polls reasonably well [29]. To capture this, we use LIWC to collect a basic sentiment measure of *positive* and *negative* affect, as well as the prevalence of the emotions *anger* and *anxiety*, and the prevalence of *swear words*.

Certainty: Not only are we interested in the viewpoints of a constituency on an issue, we also seek to understand their degree of conviction in the views they hold. To do this, we measure the frequency of both *certain* and *tentative* language once again using LIWC.

Engagement: Finally, we measure the amount of activity around the issue of same-sex marriage by collecting the total *post volume* from each state, normalized by the state's population taken from the U.S. Census. We also measure the number of people discussing the issue by calculating the number of *unique users* posting from each state, also normalized by population. Last, we wish to collect an understanding of the degree of engagement per user on the issue of same-sex marriage. We expect that users who are more passionate about an issue would post more often about that issue; thus, we collect the *average number of posts per user* from the set of users posting about same-sex marriage within the state.

CAPTURING IDEOLOGY ON SAME-SEX MARRIAGE

We compare our Twitter-based measures with statistics obtained from traditional, poll-based methods for contextualization and validation. First, we compare our ideological measures with Gallup statistics [10] on percentage of liberals versus conservatives within a state. While the Gallup data is not specifically related to same-sex marriage, we would still expect to see a correlation between some of our ideological measures and general population levels of conservatives versus liberals within a state. We compute a *Gallup ideology score* by subtracting the percentage of conservatives from the percentage of liberals for each state. Thus, a measure that has a positive correlation with the Gallup score means that it is positively correlated with a state's degree of liberalism.

Neg. Correlated	ρ	p	Pos. Correlated	ρ	p
Religion	-0.65	$<.0001$	Ingroup	0.09	0.485
Purity	-0.56	$<.0001$	Openness	0.37	$<.01$
Authority	-0.55	$<.0001$	Fairness	0.59	$<.0001$
Harm	-0.24	$<.1$			

Table 4. Correlation between Gallup conservative/liberal score and each of our ideological measures ordered by correlation score.

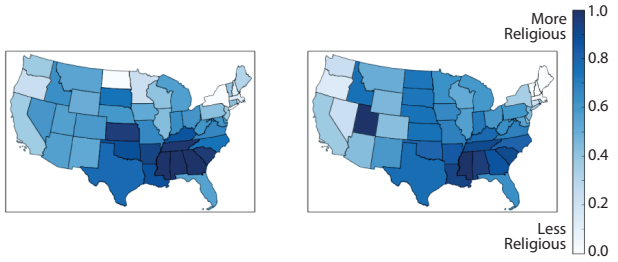

Figure 1. Degree of religious language on Twitter (left) and percentage of very religious people according to Gallup (right) ($\rho = 0.77$, $p<.0001$). Normalized to a 0-1 scale with a higher score meaning more religious.

According to previous research, we would expect our ideological measures of *harm*, *fairness*, and *openness* to positively correlate with liberalism and *purity*, *ingroup*, *authority*, and *religion* to positively correlate with conservatism [17]. As seen in Table 4, several measures are correlated using Spearman's rank correlation with the Gallup score. The exceptions are *harm* and *ingroup*, of which *harm* had a moderately strong inverse correlation. Examining the posts containing *harm* dictionary terms, we found many *harm*-related terms, such as 'protect', 'hurt', 'destroy', and 'defend' were being used not to describe people but the institution of marriage. In this context, *harm* actually weakly correlated with greater conservatism. Additionally, many *harm* terms were related to war and violence, and many *ingroup* terms were related to nationalism. These issues may not be as relevant in the discussion around same-sex marriage, but could be more relevant for a different issue such as immigration or gun rights.

Figure 1 additionally illustrates alignment of our measures with Gallup data [10], showing that our *religion* measure correlates strongly with the percent of highly religious people within a state ($\rho = 0.77$, $p<.0001$). Some differences we see could be due to the fact that the Twitter data deals exclusively with the issue of same-sex marriage while the Gallup scores are general. Also we have little Twitter data from some predominantly rural states such as North Dakota, possibly leading to more noise or bias from those states.

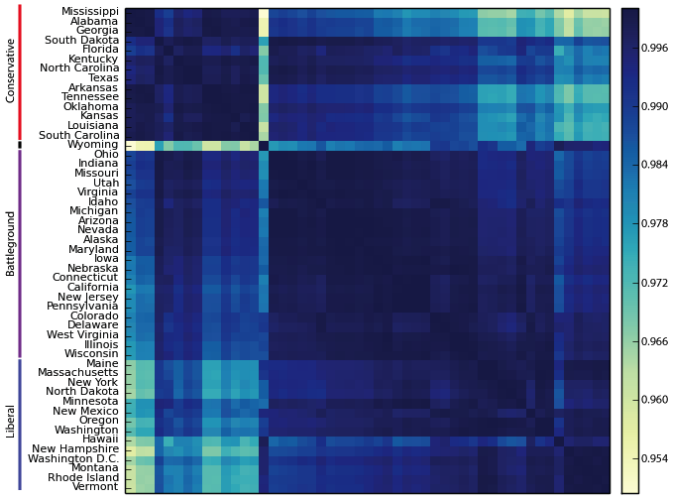

Figure 2. Distance plot of cosine similarity of states, grouped using hierarchical clustering.

Grouping States by Ideology

Grouping states by similarity on our measures helps to validate them in that we should expect states traditionally ideologically similar (e.g., conservative states in the south) to cluster together. Additionally, states that are not strongly in a liberal or conservative cluster suggest states most likely to change ideologically and subsequently legislatively. We start by constructing a vector for each state using our ideological measures and then for each pair of states compute their ideological distance by calculating the cosine similarity. This provides us with a distance matrix, which we visualize in Figure 2. We also perform centroid hierarchical clustering to group states that are ideologically close, and we visualize the main clusters in the figure.

While there are some states that are strongly on the conservative side or strongly on the liberal side, there are also many states, as seen in the middle section of the distance matrix, that could be characterized as "battleground" states. We also see that Wyoming is a clear outlier, highlighting again that for some states that have low Twitter presence due to a small or predominantly rural population, we may not get a completely accurate representation from Twitter data; we discuss these and other limitations in a later section.

Focusing on the 22 states in the middle cluster, while these states are only 43% of the states in the U.S., they account for 71% of the policy events that happened during our time period, with 91% of these states considering some kind of policy change. In contrast, only 64% of states in the top cluster considered a policy change, and all of the considerations failed, are still pending, or if they passed, were actually anti-legalization policies. None of the states in the top cluster have fully legalized same-sex marriage as of mid-2014, while 12 out of 14 states in the bottom cluster have.

Ideological Change

From our initial clustering, we found three groups that conformed to our understanding of conservative, liberal, and battleground states. Using these three categories, we can consider how states changed over time, including whether they moved from one category to another. We conduct clustering using K-means with a cluster size of 3 over posts from 2011 and 2013 to find which states moved from one cluster to another between those years. In Figure 3, we use Principal Component Analysis (PCA) to reduce the number of dimensions to two and plot the clusters for 2011. We then show with an arrow the states that have moved to a different cluster in 2013. While the dimensions themselves do not hold significance, the relative distance between the points tell us roughly how far away states are from each other ideologically as well as their placement within the clusters.

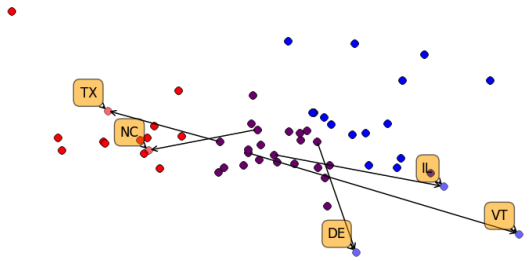

Figure 3. Clustering of states in 2011 using K-means and then plotted to 2 dimensions using PCA. States that transitioned to a different cluster in 2013 are highlighted with an arrow to their new position.

We can see that from 2011 to 2013, 5 states moved clusters, with two states joining the conservative group and three states joining the liberal group from the battleground group. When looking at the policy events that happened in these states, two of the three states moving from battleground to liberal legalized same-sex marriage during this time, while one (Vermont) had already legalized same-sex marriage. In the other direction, North Carolina was the only state during our entire time period to pass an anti-legalization policy, when the state legislature approved in 2012 an amendment defining marriage as solely between a man and a woman. This state is shown as one of the states becoming more conservative.

To summarize our efforts to externally validate our measures, we find strong agreement with established poll-based measures such as Gallup, and we see that traditionally similar states cluster together according to our measures. We also find that states in a battleground cluster were states with the largest percentage of policy events, and that states that changed clusters did so in a way that aligned with policy change, suggesting alignment between a shifting or mixed constituency ideology and higher political activity. In the remaining sections we show that our measures can differentiate states with passing same-sex marriage policies from those with failing policies, further validating our measures.

COMPARING PASSING VERSUS FAILING POLICIES

We seek to understand what is happening within a constituency before a policy decision and compare passing versus failing policies. First, we examine the values of our measures in the time period directly before the policy decision. We collect the average occurrence per post divided by the number of terms in a post for each of our measures in each state and aggregate percentages for each day leading up to a pol-

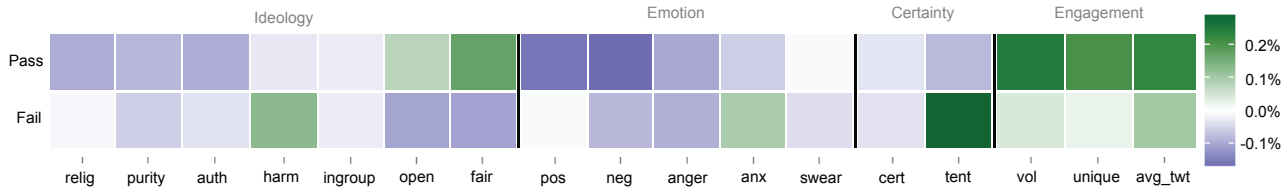

Figure 4. Average percentage of measure's terms within a post relative to the national average in the 7 days before a policy decision.

icy decision in that state. In Figure 4, we show the average value of our measures in the 7 days leading up to the final decision for policies that pass versus policies that fail. As described earlier, our measures are normalized to reflect the value within a state relative to the national average. Generally we see higher engagement and lower emotion when policies pass. Passing policies are also preceded by lower scores for conservative ideological characteristics like *purity* and *authority*, but higher in characteristics like *openness* and *fairness*, findings that correlate with Gallup ideology scores. The differences in moral framing mirror the characterization of the debate over same-sex marriage as a "culture war" pitting different notions of right and wrong against each other.

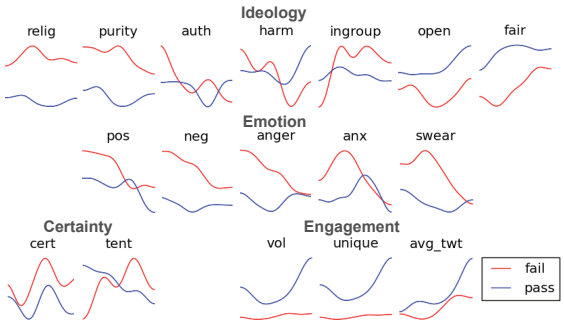

Figure 5. Sparklines showing measures in the 6 months leading up to a policy decision. All measures have been scaled to 0-1.

Figure 5 then shows how those same measures change over time in the six months before the final decision, again comparing state constituencies before passing versus failing policies. We observe that for some measures the values of the two categories are generally different across the entire six months. For instance, we see again that states where legislation fails are consistently higher in *religion* and *purity* scores and lower in *openness* and *fairness*. In some cases however, the measures are not different on average but the *slope* over that time is very different. This is true in the case of *tentativeness*, where the lines cross midway. Other measures have lines that converge by the time of the policy decision, such as for *anger*, though they start out at very different locations. This suggests that to understand whether a policy is going to pass or fail, it would be informative to not just collect the *value* of these measures but also how they *change over time*.

We collect the percentage of measures that were significantly different between passing and failing policies at several points leading up to the final policy decision. As shown in Table 5, measures were often significantly different not only directly before the event but also in the months leading up to the final decision, suggesting that policy outcomes could be predicted reasonably accurately *several months in advance*. We report two thresholds for p, as many of our measures exhibited moderate evidence of difference.

p	6 mos	5 mos	4 mos	3 mos	2 mos	1 mo	1 day
$<.1$	35%	41%	41%	35%	35%	29%	35%
$<.05$	6%	24%	35%	29%	18%	12%	29%

Table 5. Percentage of measures that were significantly different (Welch's t-test) between policies that failed versus passed at different time points before final policy decisions.

Finally, we examine how the composition of people talking about same-sex marriage changes over time before a policy decision. Looking at 6 months prior to the final decision, we break down the time into 24 one-week-long bins and collect the unique users in each week for each policy event. We then calculate the percentage of user overlap in comparison with every other week. We average over all our events and compare passing versus failing policies in Figure 6. We can see that passing policies have overall greater overlap, with 6% overlap sustained for many weeks and even several months before the policy decision is made. One way to interpret this is that it indicates there are a greater percentage of users that are passionate about an issue, or users that will make continual reference to same-sex marriage, as opposed to a single reference. We encapsulate this in our *average posts per user* measure, calculated over the time period of a month.

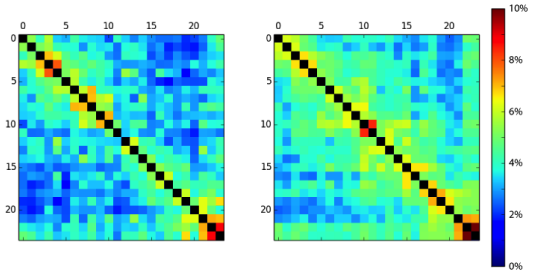

Figure 6. Degree of overlap in unique users when comparing 2 different weeks in the 24 weeks before failing (left) vs. passing (right) policies. The week directly before the policy decision is week 24.

Interestingly, we do not observe a large shift in the composition of people at any point in time, including leading up to the final decision. Instead, the people posting close to the date of the policy decision have relatively high overlap even with the people posting 24 weeks before, for both passing and failing policies. This provides evidence that the discussion of same-sex marriage is *not* getting co-opted, even with a policy decision looming and possible regional or national

AV	I	DC	I
anxiety	0.145	harm	0.132
fairness	0.085	unique users	0.092
unique users	0.056	religion	0.080
post volume	0.042	swear words	0.068
authority	0.023	post volume	0.042

Table 6. Top 5 most important feature values directly before policy decision (AV) and average feature change in the two months prior (DC).

attention. Instead, the discussions often involve people with *sustained interest* over long periods of time.

PREDICTING POLICY DECISIONS

We now focus on building classifiers to predict the outcome of a potential policy change given observations of our measures of morality, personality, emotion, certainty, and engagement within constituencies in the time leading up to a policy decision. We frame our prediction task as a binary classification problem to predict whether a particular policy will *pass* or *fail*. To start, we use observations of our measures only within the state-level constituency that the policy would affect. In later sections we incorporate influence from other states, as well as compare our models to using traditional polling data, and analyze their performance over time.

Using the measures defined previously, we construct two features for each measure. These features are informed by our earlier exploration of constituency voices leading up to policy decisions. The first is the **average value (AV)** of the measure in the 7 days before a decision, which encapsulates the population's general feelings about the issue *directly* before the decision is made. The second is the **average daily change (DC)** in the measure over the course of two months prior to legislation, which captures the direction and degree that the constituency is *changing* leading up to a decision. The time windows of one week for AV and two months for DC were found through experimentation. In total, we have 34 features, 17 AV and 17 DC, to characterize each of our policy events.

We experiment with four different classification algorithms and compare the performance. The algorithms we choose are Logistic Regression (LR), Adaptive Boosted Decision Trees (ADT), Random Forests (RF), and Support Vector Machines (SVM) with a radial-basis function kernel. We use 5-fold cross validation over 46 policy events and repeat trials 50 times for each experiment, averaging the results. We also perform a tree-based feature selection by setting a threshold on the feature importances calculated by a Decision Tree classifier. The calculation we use is the Gini importance (I), which computes for each feature the normalized total reduction of the criterion brought by that feature. In Table 6, we list the top 5 features using this metric for the categories AV and DC and note that the most important features for DC versus AV are often not the same. For instance, *anxiety* was the most distinguishing feature in the week before the policy event, possibly reflecting worry about whether the policy would pass, while *harm* was most important in terms of the trend over time.

Results

In Table 7, we report accuracy, precision, recall, F1, and area under the curve (AUC). As can be seen, Adaptive Boosted Decision Trees performs the best across the board, with on average 80% accuracy, while Logistic Regression performs the worst with 76% accuracy. The best classifier represents a 19% increase over the baseline performance of 61% if we simply pick passing for every policy event. We group our features by our different measures and report pseudo-R^2, a goodness-of-fit statistic from logistic regression as well as precision, recall, F1, AUC, and accuracy for an ADT classifier in Table 8. We break down our emotion-related measures into Sentiment, containing *positive* and *negative* affect, and Emotion, containing *anxiety*, *anger*, and *swear words*. We see from the pseudo-R^2 and accuracy values that the Engagement, Emotion, and Morality measures are the most predictive, while Sentiment on its own is not very predictive.

Algorithm	Precision	Recall	F1	AUC	Accuracy
LG	0.846	0.793	0.787	0.712	0.761
ADT	0.870	0.827	0.824	0.912	0.803
RF	0.834	0.823	0.793	0.886	0.763
SVM	0.828	0.827	0.796	0.722	0.756

Table 7. Performance of classifiers to predict passing and failing policy decisions.

Measures	R^2	Precision	Recall	F1	AUC	Acc.
Engagement	0.302	0.770	0.687	0.694	0.641	0.679
Emotion	0.176	0.657	0.853	0.734	0.617	0.652
Morality	0.443	0.707	0.647	0.629	0.688	0.584
Certainty	0.044	0.553	0.673	0.602	0.446	0.505
Personality	0.029	0.564	0.607	0.567	0.492	0.478
Sentiment	0.009	0.490	0.493	0.481	0.408	0.415

Table 8. Goodness-of-fit of logistic model and performance of ADT classifier using only one category of our measures.

Comparison to Surveys

To compare our prediction results with a proper baseline, we turn to the current gold standard, which is to use public opinion surveys. We manually gathered 204 state-wide polls taken from 2011 to mid-2014 and conducted by reputable polling organizations. We exclude surveys that offered a choice between same-sex marriage, civil unions, and no legal recognition. Instead we only collect survey results for the question of "Do you think same-sex marriage should be legalized?" and use a simple majority ruling to code the outcome of each survey. We experimented with several ways to make predictions using the survey data, but achieved the best accuracy of 70% when we use the most recent in-state poll as a predictor of a coming policy decision.

There exist more sophisticated ways to predict events using survey data that take into account many additional factors. However, given the 10% improvement of our model over the survey results, we believe that social media analysis and surveys are at the very least comparable in accuracy. When using survey data however, the number of predictions we can make decreases by 18% because of lack of data in many states, and this goes down further when we consider only polls near in time to an event. Thus, social media analysis is a way to fill the gap when little or no polling data is available.

Incorporating Voices from Other States

So far, we have worked with features that isolate the perceptions of the people within the state for which we are doing

Weighting	Precision	Recall	F1	AUC	Accuracy
Geography	0.867	0.933	0.894	0.910	0.871
Ideology	0.876	0.900	0.872	0.948	0.846

Table 9. Performance of model including without-state features with different ways of weighting each state.

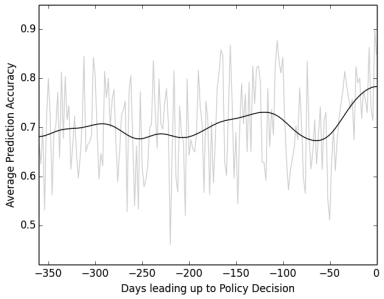

Figure 7. Prediction accuracy over time, 360 days leading up to a policy decision. Raw values are in gray while smoothed values are in black.

predictions, irrespective of the national conversation. However, the issue of same-sex marriage has had national coverage in the past 4 years, and many events were national in nature, such as when President Obama declared support for same-sex marriage. Because states do not live within a vacuum, the opinions of people in other parts of the U.S. may influence the policy decisions made within a state. Some research has shown that policies can diffuse across states that are geographically close [2], while other research has shown that it can diffuse across states that are ideologically similar [18]. To understand the influence of other states on a state's policy decisions, we construct an additional set of features to add to our model that encapsulate *without-state* expressions. For each target state, this new set of features is the average value of each feature across the remaining states.

Rather than weighting each of the other states equally, we compare two ways of weighting them: using geographic proximity and ideological similarity. Geographic proximity is calculated using the great-circle distance between the average latitude and longitude of two states. For ideological similarity, we use the ideological distance calculated earlier from our ideological features, shown in Figure 2. We use only data from before the event to calculate measures, limiting data to the two months prior. The geographic weights and ideological weights are only weakly correlated (ρ=0.094, $p<.0001$). As seen in Table 9, adding the without-state features weighted by geographic distance provides the best overall prediction, improving on the accuracy of our model with only within-state features by 7% and of poll-based models by 17%. Weighting the without-state features by ideological similarity attains higher precision than weighting by geography, but recall is lower, leading to a lower overall accuracy.

Performance over Time

Finally, we examine how our method performs over time leading up to the final policy decision. Looking at the year prior to the event, we make a prediction every seven days using our best model and using as inputs our features at that point in time. We calculate prediction accuracy at all these time intervals and then present the raw and smoothed results in Figure 7. We can see that prediction generally improves as we approach the date of the policy decision from the year prior, although the raw data is quite noisy. This highlights that our prediction may need to be averaged over time for best results. We also note that our model can achieve an accuracy above 70% *several months* before the final decision.

DISCUSSION

Through a case study of same-sex marriage, we demonstrate how analysis of language and activity on social media allows us to characterize a population's ideology. Using measures we extracted from social media text, we were able to group states into ideological camps and observe how they shifted over a period of 4 years. For instance, we identified states like Texas and North Carolina becoming more conservative and Delaware and Illinois becoming more liberal. Then we used these measures to examine the link between policy decisions and prior constituency opinion. We found that policies that passed had a greater percentage of people with sustained interest over time, had greater overall engagement levels, and had significantly higher levels of language related to *fairness* and *openness* before the decision. On the other hand, states with policies that failed had higher levels of *anxiety*, *religion*, and *tentativeness*. These findings align with previous research characterizing the same-sex marriage debate as a "culture war" [1], where proponents advocate for it in terms of fairness morality, while opponents argue against it in terms of religious morality.

The multifaceted nature of our measures derived from social media highlights the possibility of augmenting or replacing traditional poll-based measures. This moves the level of understanding of a constituency from simple pro/con values typically seen in surveys to a nuanced understanding of aspects of ideology, emotion, and issue engagement. We also demonstrate that relatively accurate predictions can be made several months in advance of the final decision. In contrast, a survey can only measure public opinion at a certain point in time and is also often slow and expensive to distribute. While pro/con surveys have been shown to correlate with issue sentiment [29], we saw that sentiment was the least predictive measure within our model. This suggests that the additional measures we compute can further characterize constituencies. Some of the measures we collected, such as moral framing or level of certainty, also often occur implicitly in natural, everyday speech, and may be difficult to collect via a survey.

Finally, our work contributes to the literature on the interplay of constituencies and governmental policies. We show that the expressions of a constituency as measured on Twitter was indeed predictive of same-sex marriage policy change, confirming prior research [22]. We also found that both the value of our measures as well as how they changed over time were predictive of policy change, suggesting that policymakers may be attuned to both the views of their constituents as well as general trends. This may be because policymakers often must consider how their actions will be viewed several years in the future.

In addition, we found that constituencies of other states help to predict the policies within a state. Specifically, we found

that weighting the importance of other states by geographic distance provided greater predictive power than weighting them by ideological similarity. This suggests that the views of geographically proximate states carry greater influence than states that are far away but ideologically similar. This may be due to the distribution of regional news, policymakers with regional communication networks, or people crossing state boundaries for commuting, visiting, or moving.

Design Implications

This research bears implications for analysis of political discourse and design of applications targeting political expression. Research has found increased polarization in recent years, with many studies blaming "filter bubbles" [24] in search, news, and social streams. Using the measures that we have selected, we can gain not only an understanding of how a particular population stands on an issue but also the moral and personal lens through which they approach the issue. By presenting opposing opinions in this light, applications may be able to foster greater understanding and empathy across divides, further humanizing opposing side.

Additionally, when it comes to presenting diverse information, research has found that people can react adversely to opposing information [27]. Recent research suggests that first broaching intermediary topics that people have in common can be a way to ease people into reading divergent opinions on sensitive topics [16]. Perhaps applications could present disparate opinions but keep one aspect in common with the user, such as their moral framing. For instance, while same-sex marriage is often opposed for religious reasons, people have also used religion to argue for it.

Finally this research suggests tools for policy analysts, activists, and political organizations. In recent years, social media has become an important place for political activism [13] and political discourse. We can imagine our analyses being used, for instance, on a social media dashboard to help people monitor voices mentioning different issues over time. This could help activists and political organizations better allocate their resources to certain groups, observe large-scale shifts in perceptions and opinions as they are happening, and target or frame their message to speak to certain populations. This tool could also be useful for the general public during times of political uncertainty.

Limitations and Future Work

We now turn to discuss some limitations of our methods and datasets used as well as promising future work. First, some limitations arise because we used a lexicon-driven approach, specifically dictionaries taken from LIWC, to calculate many of our measures. We can only measure self-stated terms using this method, and we also did not account for negation, sarcasm, and irony. While these issues may be adding noise to our data, we believe our findings still hold because we consider tens of thousands of posts on average per state, and we observe these measures over a long period of time. Additionally, the conventional method of using surveys requires users to explicitly consider their opinions on issues, which may bias results in a different way, while we collect implicit signals from conversations on Twitter.

The use of Twitter data as a proxy for the voices of constituencies also has some limitations. Twitter tends to be biased towards urban areas [19] and towards more technologically-literate populations. This may lead to some of our measures not accurately representing the constituency of a state. Also, we have no way of differentiating age or ability to vote on Twitter. Restricting our measures to registered voters might provide a more accurate picture of a voting population as opposed to an entire population, which would be useful for certain questions. Finally, this research illuminates the *correlation* between constituency opinions and policy decisions. We cannot make any claims using our methods that policy outcomes or the decisions of policymakers are *caused* by the opinions of a constituency. Overall, despite imperfections in our data, that we were able to differentiate and even predict passing and failing legislation provides ecological validity.

For this work, we chose to focus on same-sex marriage because we felt it was an unprecedented opportunity to study an important movement in the midst of major political battles. However, while 46 policy events from one issue is a great deal given the time range, it is not a lot of events to use for classification. In the future, it would be interesting to study how our measures, methods, and findings generalize to other issues such as marijuana legalization or gun control. This was one reason we did not build a bag-of-words classifier, and instead focused on dimensions like *fairness* that are known to underlie moral framings that research has shown drive stances on many issues [5]. This will allow us to generalize more easily in the future. We also only focused on policies *for* same-sex marriage legalization. It would be interesting to see how we could incorporate policies *against* same-sex marriage in our model and if we would need to weight any of our measures differently. Finally, there exists research that looks at the *impact* of policies on constituencies, finding evidence that the influence may also go in the other direction. Another promising area for future work could be examining the long-term impacts of policies on a constituency using our methods.

CONCLUSION

We conducted a large-scale quantitative analysis of expressions on social media on the issue of same-sex marriage. We explored several attributes including moral framing, personality, levels of emotion, degree of certainty, and engagement to characterize constituencies over time and leading up to policy decisions. We found that we could predict whether a state-level policy would pass or fail with 80% accuracy using as input our measures within the constituency before the final decision. Our accuracy improves to 87% when we add measures from outside the state, weighted by geographic proximity. The models we built constitute a 17% absolute increase over the current gold standard of using public opinion surveys. We believe that the measures and the models we have described could be useful for technological applications for recommending content and finding common ground on controversial issues. They could also be useful for helping policy

analysts, activists, and political groups monitor constituency opinions and target messages on important issues.

REFERENCES

1. Ball, C. Moral foundations for a discourse on same-sex marriage. *Geogetown Law Journal 85* (1996), 1871.
2. Berry, F., and Berry, W. State lottery adoptions as policy innovations. *American Political Science Review* (1990), 395–415.
3. Bollen, J., Mao, H., and Pepe, A. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *Proc. ICWSM* (2011).
4. Carney, D., Jost, J., Gosling, S., and Potter, J. The secret lives of liberals and conservatives. *Political Psychology 29*, 6 (2008), 807–840.
5. Clifford, S., and Jerit, J. How words do the work of politics. *The Journal of Politics 75*, 03 (2013), 659–671.
6. Cohen, R., and Ruths, D. Classifying political orientation on Twitter: It's not easy! *Proc. ICWSM* (2013).
7. Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., and Flammini, A. Political polarization on Twitter. *Proc. ICWSM* (2011).
8. De Choudhury, M., Counts, S., and Horvitz, E. Major life changes and behavioral markers in social media: case of childbirth. *Proc. CSCW* (2013).
9. Diakopoulos, N., Zhang, A., Elgesem, D., and Salway, A. Identifying and analyzing moral evaluation frames in climate change blog discourse. *Proc. ICWSM* (2014).
10. Gallup. *The State of the States*, 2014 (accessed on August 30, 2014). `http://www.gallup.com/poll/125066/State-States.aspx`.
11. Gayo-Avello, D. A meta-analysis of state-of-the-art electoral prediction from Twitter data. *Social Science Computer Review* (2013).
12. Geer, J. *From tea leaves to opinion polls*. Columbia University Press, 1996.
13. Gerbaudo, P. *Tweets and the streets: Social media and contemporary activism*. Pluto Press, 2012.
14. Gilens, M. *Affluence and influence*. Princeton University Press, 2012.
15. Gottipati, S., Qiu, M., Yang, L., Zhu, F., and Jiang, J. Predicting users political party using ideological stances. In *Social Informatics*. Springer, 2013, 177–191.
16. Graells-Garrido, E., Lalmas, M., and Quercia, D. People of opposing views can share common interests. *Proc. WWW Companion* (2014).
17. Graham, J., Haidt, J., and Nosek, B. A. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology 96*, 5 (2009), 1029.
18. Grossback, L., Nicholson-Crotty, S., and Peterson, D. Ideology and learning in policy diffusion. *American Politics Research 32*, 5 (2004), 521–545.
19. Hecht, B., and Stephens, M. A tale of cities: Urban biases in volunteered geographic information. *Proc. ICWSM* (2014).
20. Jost, J., Nosek, B., and Gosling, S. Ideology: Its resurgence in social, personality, and political psychology. *Perspectives on Psychological Science 3*, 2 (2008), 126–136.
21. Laver, M., Benoit, K., and Garry, J. Extracting policy positions from political texts using words as data. *American Political Science Review 97*, 02 (2003), 311–331.
22. Lax, J., and Phillips, J. Gay rights in the states. *American Political Science Review 103*, 03 (2009), 367–386.
23. Levitt, S. How do senators vote? Disentangling the role of voter preferences, party affiliation, and senator ideology. *The American Economic Review* (1996), 425–441.
24. Liao, Q., and Fu, W. Beyond the filter bubble. *Proc. CHI* (2013).
25. Lin, Y., Bagrow, J., and Lazer, D. More voices than ever? Quantifying media bias in networks. *Proc. ICWSM* (2011).
26. McGuire, K., and Stimson, J. The least dangerous branch revisited: New evidence on supreme court responsiveness to public preferences. *Journal of Politics 66*, 4 (2004), 1018–1035.
27. Munson, S., and Resnick, P. Presenting diverse political opinions: how and how much. *Proc. CHI* (2010).
28. Naaman, M., Zhang, A., Brody, S., and Lotan, G. On the study of diurnal urban routines on Twitter. *Proc. ICWSM* (2012).
29. O'Connor, B., Balasubramanyan, R., Routledge, B., and Smith, N. From tweets to polls. *Proc. ICWSM* (2010).
30. Pennebaker, J., Chung, C., Ireland, M., Gonzales, A., and Booth, R. The development and psychometric properties of LIWC2007. *LIWC.net* (2007).
31. Sim, Y., Acree, B., Gross, J., and Smith, N. Measuring ideological proportions in political speeches. *Proc. EMNLP* (2013).
32. Tumasjan, A., Sprenger, T., Sandner, P., and Welpe, I. Predicting elections with Twitter. *Proc. ICWSM* (2010).
33. Yarkoni, T. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality 44*, 3 (2010), 363–373.
34. Zhou, D., Resnick, P., and Mei, Q. Classifying the political leaning of news articles and users from user votes. *Proc. ICWSM* (2011).