Measuring Stylistic Similarity of Political Rhetoric using VSMs and Link Structures

May 17, 2011

Moritz Sudhof
B.S. candidate
Computer Science
Stanford University

Abstract

This paper explores the use of content and link structures to cluster Twitter users based upon how they frame an issue. We present a model that can be used to cluster users relative to a given topic. This model includes many free parameters whose tuning is explored in a case study of tweets relating to the 2011 Wisconsin union protests. In the case study, we find that content and different link types (hashtags and mentions) can be combined to cluster users based upon 1) their general attitude towards the topic and 2) the medium through which the attitudes are expressed.

1 Introduction

The barrier to becoming a content creator on the Web continues to decline, and the amount of content pertinent to any single issue continues to increase. Correspondingly, tools that can help us understand content are becoming increasingly crucial. Previous work has focused on understanding the temporal nature of content creation [1] and the manner in which this content spreads [2]. This research seeks to complement such temporal studies by endeavoring to understand the content based upon its *approach* to a topic.

Given a corpus of documents that all engage with a single topic, this research attempts to meaningfully cluster these documents based upon the style of their engagement with the topic. More specifically, this paper explores methods for leveraging the content and the link structures of tweets to build a model that clusters users.

We begin by introducing Twitter and the unique characteristics that accompany Twitter data (section 2). We then present the model, focusing both on the quantitative methods used and the intuition for those methods (section 3). Finally, as the model includes many free parameters, we test and iterate by presenting a case study based on a corpus of tweets relating to the Wisconsin union protest controversy of February 2011 (section 4).

2 Data

We begin by discussing the properties of Twitter data in general and the data used for this research in particular.

2.1 Twitter

Twitter is a microblogging service that is both a social network and a massive source of real-time information [3]. Twitter users communicate with their "followers" by posting messages, called tweets, that are 140 characters or less. Users can "mention" other users by including a user's username preceded by an "@," and users can tag their messages through hashtags, which are simply words preceded by a "#."

Several features make Twitter a convenient medium to focus on for this research. Twitter is a popular medium (the website serves about 140 million tweets per day), and it is relatively simple to download tweets, either through the API or through public archives. Hashtags simplify the task of bundling tweets that engage with the same topic, making it relatively easy to create a sufficiently large corpus of tweets that all discuss the same topic. Furthermore, Twitter is a personality-driven, informal

source of information. This ensures that tweets are often opinionated and convey information using colorful language.

Finally, the mentions (@user) and hashtags (#hashtags) in Twitter data add internal structure. Mentions link tweets to users, and hashtags link tweets to general concepts. Both constructs are functionally important: mentions ensure that a specific user is included in or addressed by a commentary, and hashtags ensure that a tweet is grouped with all other tweets with that hashtag. More interestingly, though, they are also often highly performative – directed not at the person mentioned but at the audience instead, for purposes of, for example, "calling someone out." This usage makes them potentially useful in defining the topography of the dialogue.

A mention in a tweet is an indication that the tweet is relevant to the mentioned user, and multiple mentions in a tweet are an indication, or a vote, that each mentioned user is relevant to the tweet and thus to each other. Therefore, the link structure of all tweets is comparable to crowd-sourced votes (where the crowd is the corpus of tweets) regarding which users operate in similar corners of the topic space.

The purpose of hashtags, on the other hand, is ever-evolving and more difficult to pinpoint. Tags on Twitter are unique because of their conversational, not just organizational, nature. Studies have found that hashtags are generally used to filter and direct content so it appears in certain streams [4], and increasingly, those streams have been not purely topical but also coded with derision, irony, humor, and love, to name a few. In cases of rich and heavy hashtag usage, tags can be used to summarize the topical and emotive bent of the tweets.

2.2 The Corpus

The corpus used for this research consists of tweets discussing Gov. Walker's proposal to remove the collective bargaining rights of public employees and the protests that followed. The corpus was built by archiving tweets including a *#wiunion* hashtag and downloaded from TwapperKeeper, an online Twitter archiving service, for this research. Although this hashtag archive does not include all tweets related to the Wisconsin union controversy, the *#wiunion* hashtag was used by a wide array of Twitter users with varying opinions to indicate that a tweet was relevant to the discussion. The corpus consists of 219,528 tweets authored by 36,119 users between February 17, 2011 and March 18, 2011.

We chose a political controversy as the basis for our corpus because the political space is ideal for this research. First, it is a space reliably marked by a variety of different, conflicting perspectives, regardless of which topic filter is applied. Furthermore, opinions tend to be resolutely held and forcefully defended, resulting in strong language. Finally, in politics, the way something is said is often more important than what is being said. That is, political rhetoric emphasizes the presentation of information, and political actors often deliberately repeat certain words and phrases in order to define the framing of an issue. Politics is therefore the perfect space to first attempt to measure similarity in the approach to an issue.

3 The Model

Although certain aspects of the model are tailored to the Twitter space and the problem of identifying users to follow, the model was developed with consideration for its ability to be generalized to any set of documents with a link structure.

The model will cluster users relative to a given topic, which means a general corpus must be clustered by topic before the model can cluster users in a topic space. Effectively clustering a corpus by topic is not a problem addressed by this paper. Instead, topic-based corpuses were developed by choosing an issue, manually selecting the top descriptive hashtag(s) Twitter users employ when discussing the issue, and grouping all tweets with those hashtag(s).

3.1 User Extraction

Given a corpus of tweets that all address a similar topic (e.g., all tweets that share a certain hashtag), the model first identifies a set of users to compare. This is a necessary filter in the Twitter space: there are many users contributing to the dialogue on any given topic, but most users are sparse contributors to the discussion in terms of volume, not particularly reliable and influential actors in the space, or both. Extraction of a set of users seeks to ensure that users that are clustered contribute substantially to the dia-

logue in terms of both volume and influence.

Ensuring extracted users contribute substantially in terms of volume is a relatively simple problem, and to solve it, this model simply considers only those users that contribute more than a certain threshold of tweets to the topic corpus (40 tweets was deemed an appropriate barrier for consideration for this application). Measuring influence on Twitter is a more difficult problem that has been studied extensively by Weng and Leavitt et al., among others. The details and controversial aspects of different influence measures are beyond the scope of this paper. It is clear, however, that user influence, or "clout," generally holds across topic domains [5]. Furthermore, since we only seek to exclude from consideration those users who have little or no influence, exact influence rankings of the top users are not necessary. We are therefore comfortable with leveraging measures of the relative influence of Twitter users available through numerous online services (this model utilizes Twitalyzer.com) to quantify a user's "clout."

3.2 Bundling

After a set of interesting users has been defined, the model bundles all tweets authored by a single user into a single document. Documents are then tokenized in a manner sensitive to the vocabulary of Twitter (paying attention to, for example, hashtags, mentions, and links) and transformed into "bags of words." In order to preserve the richness of the language, documents are tokenized in a manner sensitive to the Twitter medium (e.g., with consideration for the preservation of emoticons), and stopwords filtering and term stemming are not employed. After tokenization, each bag of words expresses, in sum, a user's engagement with the corpus topic.

3.3 Content Analysis

The next step is to measure similarity between documents by using vector space models (VSMs) [6]. Given a corpus vocabulary of size n, we can represent the bag of words of document i as an n-dimensional vector v_i, where each dimension represents one term in the vocabulary. The value of dimension j of document vector v_i, or $v_{i,j}$ then depends on the "importance" of term j in document i, where "importance" is quantified by a weighting mechanism. By representing documents as vectors in a Euclidean space, VSMs allow us to measure semantic similarity by leveraging Euclidean or non-Euclidean distance measures.

The most common method for weighting term vectors for information retrieval is by *tf-idf* [7]. *tf-idf* weighting formalizes the intuition that a rare term has higher information content than an expected term. More precisely, it scales a term frequency (*tf*) by the log of the term's inverse document frequency (*idf*). Under this weighting mechanism, given a document i and a term j,

$$v_{i,j} = \frac{y_{i,j}}{\sum_k y_{i,k}} \cdot \log \frac{|D|}{|i : j \in d_i|}$$

where $y_{i,j}$ is the number of occurrences of term j in document i, $|D|$ is the number of documents in the corpus, and $|i : j \in d_i|$ is the number of documents in which the term i appears.

Although the model weights terms by *tf-idf*, we also weight terms by their odds, which is a measure for how significant a term is for a particular author. Given a document i, the odds of term j appearing is simply

$$O_{i,j} = \frac{y_{i,j}}{n_i - y_{i,j}}$$

where $y_{i,j}$ is as above and $n_i = \sum_k y_{i,k}$. This measure has produced mixed results in emphasizing distinctive, emotive terms [8], and we improve its performance by first filtering statistically insignificant terms by applying the G-test (also known as the log-likelihood ratio test) and filtering using the test's p-values.

The G-test compares observed frequencies to an expected frequency to determine whether the difference in observed frequencies is significant [9]. In general, given term j and a set of documents D, the G-value is computed as follows:

$$G = 2 \sum_{i \in D} O_{ij} \cdot ln(\frac{O_{ij}}{E_{ij}})$$

where O_{ij} is the observed frequency of term j in document i and E_{ij} is the expected frequency. For this application,

$$E_{ij} = \frac{\sum_{i \in D} O_{ij}}{|D|}$$

If the G-value is above a threshold value (which depends on the degrees of freedom of the data, or the

number of documents being compared), the term is statistically significant. The G-test has been shown to be effective in several computational linguistics applications [10], and the model applies it to document vector pairs to filter insignificant terms before measuring similarity.

After document vectors are weighted, the cosine similarity measure, which has proven a reliable similarity metric for textual analysis[11], is used to quantify similarity.

Given document vectors $\mathbf{x}$ and $\mathbf{y}$ with n dimensions, the cosine similarity is defined as

$$\begin{aligned} cos(\mathbf{x}, \mathbf{y}) &= \frac{\mathbf{x} \cdot \mathbf{y}}{\| \mathbf{x} \| \| \mathbf{y} \|} \\ &= \frac{\sum_{i=1}^{n} \mathbf{x}_i \cdot \mathbf{y}_i}{\sqrt{\sum_{i=1}^{n} \mathbf{x}_i^2 \cdot \sum_{i=1}^{n} \mathbf{y}_i^2}} \end{aligned}$$

The intuition behind the cosine similarity measure is that it is not the Euclidean distance between two vectors but the angle between them that defines their similarity. This makes it a fitting similarity measure for documents. The difference between similarity measures that normalize for vector magnitude, though, is commonly thought to be insignificant in information retrieval applications [12], and we chose cosine similarity as it is a familiar and popular metric for natural language applications.

The fact that all documents address the same general content and are authored by often colorful personalities yields two interesting results: there are many content terms that are shared by all documents, and each document has idiosyncrasies that are more indicative of the author's writing style than the author's approach to the issue. In the case study in section 4, we will examine whether *tf-idf* or odds weighting performs better for this application.

3.4 Link Analysis

Parallel to the VSM analysis, the model leverages the link structures to measure link-based similarity between users.

3.4.1 Mentions

The mentions link structure is essentially a bipartite graph with tweets as nodes on one side linking to nodes of users on the other side. We use the SimRank algorithm, a measure of structural-context similarity, to measure similarity between nodes. Consider two distinct nodes a and b where $I(a)$ and $I(b)$ are the set of in-neighbors of a and b. SimRank is a recursive algorithm that defines the similarity $s(a, b)$ as

$$s(a,b) = \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{i=1}^{|I(b)|} s(I_i(a), I_j(b))$$

where C is a decay constant between 0 and 1 [13]. Similarities are measured by initializing self-similarities to 1 and all other similarities to 0, then iteratively computing similarities until they converge.

Consistent with our bundling of all of a user's tweets into a single document, we also condense the graph by grouping the source tweets by author and weighting an edge from source user s to destination user d by the number of times s mentions d divided by the total number of mentions by user s. To account for edge weighting, we now define similarity as

$$s(a,b) = \frac{C}{W(a)W(b)} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} W_{(I_i(a),a)} W_{(I_j(b),b)} s(I_i(a), I_j(b)) \quad (1)$$

where $W_{(I_i(a),a)}$ is the weight of the edge from node $I_i(a)$ to node a.

We are only interested in the similarity scores of a subset of all users in the topic space, but the recursive structure of the algorithm requires us to measure every relevant pair-wise similarity in order to measure the similarity between just one pair of nodes in a strongly connected graph. Furthermore, the space needed to store the similarities between all pairs of authors for large corpuses can quickly exceed the amount of main memory on most machines. Therefore, we use the Random Surfer Pair model to measure similarity between users[14]. Intuitively, random walks measure similarity between nodes a and b by measuring how long it takes for two random walkers, one starting at a and one at b, to meet if they walk in lock-step. Experimentally, as few as 1000 walks have been shown to reasonably approximate

similarity between two nodes. This algorithm is generalizable to any directed graph.

3.4.2 Hashtags

Link analysis of the hashtag link structure is directly comparable to the mentions link structure analysis except that with the hashtag bipartite graph, we are interested in similarity between source nodes, not destination nodes. However, since hashtags are not links as much as indicators of topical and emotive alignment, modeling them as links may not be the most appropriate analysis. Conceptualizing hashtags as the set of all of the topical and emotive spaces a user inhabits suggests using a different similarity metric: the Jaccard similarity coefficient.

Intuitively, the Jaccard similarity coefficient measures the similarity between two sets based upon how much the two sets *do* overlap compared to how much the two sets *could* overlap. Formally, the Jaccard similarity of sets A and B is measured as

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Applied to two sets of hashtags, this similarity measure quantifies the emotive and topical overlap of two users.

As a set-based similarity measure, the Jaccard similarity coefficient does not discriminate between set elements; all hashtags are created equal. Users use some hashtags much more than others, however, and this difference in use is significant. We alter the above formula slightly to account for the relative importance of hashtags. Given a hashtag i and a user a, let us define the weight w_{ai} as

$$w_{ai} = \frac{f_{ai}}{\sum_j f_{aj}}$$

where f_{aj} is the frequency of hashtag j in all tweets authored by user a. Weights are therefore simply normalized frequencies (and the weights for all tags for any given user sum to 1). We can therefore define weighted Jaccard similarity coefficient as

$$J_w(A,B) = \frac{1}{2} \sum_{i \in |A \cap B|} min(w_{ai}, w_{bi})$$

The case study below will analyze the difference between clustering based upon SimRank similarities and weighted Jaccard similarities in an attempt to understand the significance of hashtags.

3.5 Clustering

Finally, hierarchical agglomerative clustering (HAC) is used to cluster the documents. HAC views each element as a singleton cluster at the outset and clusters elements by successively merging clusters until all elements belong to the same cluster. At each iteration, the clusters to merge are chosen by identifying clusters that are "closest," where distance between clusters can be defined as the shortest distance, the longest distance, or the average distance between the clusters, among others. Experimentally, we did not find significant differences in clustering based upon the type of distance measure chosen, and this paper will not analyze their tradeoffs.

HAC provides several attractive features for this application. The dendrograms produced by HAC are ideal for visual evaluation, as the distance between clusters and cluster quality are visually apparent, as is the sequence of cluster mergers. Furthermore, HAC allows us to combine similarity measures. Since HAC does not require the ability to compute centroids of document sets in order to cluster documents, we can easily combine content-based similarities and link-based similarities before clustering, where link-based clustering can be based on either mentions or hashtags.

We combine similarities derived from different analyses by assigning a weight w to one of the similarity measures, s_1, to reflect its "value" relative to the other similarity measure, s_2. The combined similarity is then simply the weighted sum of both similarity measures, or $ws_1 + (1-w)s_2$. Since we do not have labeled data for this task, we will determine the weight w, or relative "value" of a similarity measure, heuristically. That is, we let w be a range of values between 0 and 1 and choose the value that maximizes the clustering coefficient of the resulting combined similarity. Intuitively, the clustering coefficient measures the degree to which nodes in a graph are highly interconnected. A high clustering coefficient indicates that there are more clusters and groups than in a purely random network. We measure a clustering coefficient by creating a graph with all documents as nodes and only the greater half of pair-wise similarities as edges, and measuring the fraction of all possible triangles of the graph that ex-

ist. We evaluate the effect of combining content- and link-based similarities in the case study.

4 Case Study

We now evaluate the model by applying it to the Wisconsin union corpus.

4.1 Political Background

On Feb. 11, 2011, Wisconsin Governor Scott Walker proposed removing most public employee collective bargaining rights and tightening the requirements for union certification as part of an effort to close a projected $3.6 billion budget deficit. The legislation also proposed requiring state employees to contribute 5.8% of their salaries to pension costs and to pay 12% of their own health insurance premiums. Although Democrats and unions agreed to require state employees to contribute more towards pensions and premiums, they ferociously opposed the new union restrictions, especially the removal of collective bargaining rights. For overviews of the timeline of events, see [15] [16] [17].

4.1.1 The Debate

The controversy surrounding the bill soon transcended both budget politics and Wisconsin state borders. By Democrats and union supporters the bill was seen as a union-busting measure meant to strip workers of basic rights, and by Republicans and supporters of Gov. Walker, it was viewed as a cost-saving measure that proposed only modest and fair changes for public employees.

To Democrats, Scott Walker was using the budget deficit as an excuse to push an anti-labor agenda. They viewed the bill as a dangerous step towards seriously curtailing the rights of workers in favor of businesses. Supporters of Gov. Walker, conversely, argued that union supporters were overreacting and represented a dangerous culture of entitlement often referred to as indicative of a "nanny state."

4.1.2 The Role of Social Media

Although it is difficult to quantify the effect of social media on the debate, it is clear that social media played an integral role in shaping, not just commenting on, the controversy. After leaving the state to block a vote, Democratic state senators communicated with supporters at home primarily via Twitter, and Gov. Walker used his personal Twitter account (@scottwalker) to defend his proposals. Finally, Twitter and Facebook were the primary tools used to organize protests and disseminate information rapidly [18].

4.1.3 Note on the Corpus

On Twitter, many hashtags were used to tag tweets dedicated to the Wisconsin protests, but none were used as prevalently as *#wiunion*. This hashtag was introduced on Feb. 11 by Kristian Knutsen of *Isthmus: The Daily Page*, an alt-weekly based in Madison, Wisconsin. According to *The Daily Page*, the tag was meant to accompany any tweets referencing the showdown [19]. During the protests, *#wiunion* was consistently a trending topic nationally and even internationally for a brief time.

Since the corpus consists only of those tweets tagged with *#wiunion*, it does not encompass the entirety of the conversation regarding the protests in Wisconsin. Furthermore, by volume of tweets, the corpus is biased in favor union supporters. This dynamic is understandable given that the legislation threatened to curtail union rights, and the defiance and accompanying energy began with pro-union voices, with anti-union voices arising only in response.

This corpus is well-suited for this case study, however, because tweets in the corpus are at once binary and multi-dimensional. The issue is very polarizing, and users using the *#wiunion* hashtag come out either firmly in favor of the unions or firmly opposed. This rigid polarity makes a qualitative assessment easier as it reduces the dimensionality of the discussion: on a basic level, users can be determined to be either in support of the unions or Gov. Walker.

The corpus is still multi-dimensional, however, in terms of the many different ways users convey their support or opposition. By highlighting the nobility of the protesters, for example, a user voices support for the unions differently from the user that primarily operates by comparing Gov. Walker to a fascist. (For examples of different approaches to the controversy represented in the corpus, see Table 1.) Therefore, this corpus allows the case study to test the model's effectiveness on a corpus with a rich variety of approaches to an issue while also simplifying

a qualitative assessment.

For this case study, we will focus on a set of eight users. This limited focus ensures that we can retain oversight during the case study, aiding our qualitative assessment. After narrowing the user base of the corpus to only those users that authored at least 40 tweets in the corpus, we extract a set of eight users to analyze more closely by choosing the users with the highest influence ratings. We also manually ensure that this set of users includes supporters of both the unions and Gov. Walker and expresses that support in a variety of ways. (For a qualitative summary of the users' tweets in the corpus, see Table 2.)

4.2 Term Vector Weighting and the G-test

By definition, *tf-idf* weighting emphasizes the terms that are rarest across the corpus, and odds weighting emphasizes the terms that are used most often by a user. *tf-idf* weighting takes the whole corpus into account, whereas odds weighting only considers a single document. Intuitively, both try to capture a user's distinct approach: *tf-idf* emphasizes the unusual terms a user uses, thus highlighting the terms that distinguish a user from the rest of the corpus (note, of course, that a high frequency term with a low inverse document frequency could still be emphasized by *tf-idf*), and odds weighting emphasizes the terms used most often by the user.

To see which weighting mechanism best expresses a user's engagement with an issue in practice, we compare weighted vectors qualitatively. Given authors a and b and weighted vectors v_a and v_b, the difference vector $v_d = v_a - v_b$ can be used to identify the features that distinguish the two authors given the weighting chosen. Terms with high v_d values are those that most distinguish a from b, and terms with low v_d values are those that most distinguish b from a.

Consider users *aflcio* and *diggrbiii*. We will compare these two users because they are on opposite ends of the political spectrum. Comparing them therefore highlights the general distinctive characteristics of *aflcio* and *diggrbiii*, not just those characteristics that distinguish them (this intuition was confirmed by examining other comparison vectors). Table 3 displays the most distinctive features for both *aflcio* and *diggrbiii* for both weighting mechanisms.

A quick perusal of the table shows, however, that neither weighting mechanism is very effective. Odds weighting, with its emphasis on words used frequently by an author, results in the elevation of many function words such as *on*, *from*, *with*, *the*, and *that*, which, though relevant to authorship attribution, tell us very little about the author's engagement with an issue. In contrast, *tf-idf* weighting, with its emphasis on very rare words, results in the elevation of terms that are used only once or twice and cannot be used as reliable indicators of a user's approach, such as *eaglery* and *chyron*. This problem is the motivation for the application of the G-test. Applying the G-test to a pair of users filters all features such that only those with a G-value greater than 3.8, which corresponds to a p-value of less than 0.05, remain. Table 4 displays the most distinctive features after application of the G-test. (The number of distinctive features is smaller as many terms are removed from consideration by the G-test.)

The quality of both sets of distinctive terms is improved by application of the G-test, but odds weighting seems more effective at emphasizing the terms that express the author's approach to the issue. Qualitatively, tweets authored by *aflcio* generally fall into three categories: those that express defiance at Gov. Walker's actions, those that express support for protesters and laborers, and those that relay information about on-the-ground events. All of these categories are represented by sets of the distinctive terms: *#statesos*, *#notmywi*, and *#standupoh* express the defiance, *#wearewi*, *solidarity*, *workers*, *rally*, and *labor* express the support of the protesters and workers, and *now*, *today*, and *capitol* are terms used when relaying information about current on-the-ground events. *diggrbiii*'s distinctive odds-weighted terms match the general sentiment of the tweets less facilely, but some differences between *aflcio* and *diggrbiii* are notable. Instead of *now* and *today*, *diggrbiii*'s term list includes *was*, which, in conjunction with *#tcot* (top conservatives on Twitter), *#tlot* (top liberals on Twitter), and *#debunked*, accurately reflects the more after-the-fact, armchair commentary *diggrbiii* engages in. (Note: *#wiunion* is included as a distinctive feature of *diggrbiii* because each tweet includes the hashtag and

aflcio authored so many more tweets than *diggrbiii* to make the difference in occurrences statistically significant.)

The distinct terms resulting from a *tf-idf* weighting do not reflect the general body of tweets as well. For example, the terms *vp* and *biden* are included by *aflcio* as quote-attribution, but it is the sentiment of the quote, not who said it, that tells us about *aflcio*'s approach to the discussion. Since very few other Twitter users mentioned Vice President Biden in tweets relating to Wisconsin, though, *tf-idf* weighting makes these terms "characteristic" to *aflcio*'s content. Furthermore, although the G-test filtered many of the rare hashtags and mentions included by *aflcio*, some were not filtered and were again elevated to being very "characteristic" by their rarity.

Table 3: Effect of odds and tf-idf weighting

	odds	tf-idf
aflcio	#statesos	#statesos
	#wearewi	#notmywi
	#notmywi	afl-cio
	to	#standupoh
	solidarity	blog
	capitol	biden
	afl-cio	vp
	#standupoh	#wearewi
	of	@melissaryan
	now	#humanrights
	#wiunion	statehouse
	workers	@wisaflcio
	rally	unbelievable
	on	pres
	blog	@tulaconnel
	one	#inunion
	from	officers
	with	text
	new	@defendwisconsin
	today	date
diggrbiii	#tcot	@diggrbiii
	the	@karoli
	?	#debunked
	@karoli	#tcot
	that	switch
	i	bait
	what	@derekahunter
	was	doctors
	@diggrbiii	@brentteichman
	#tlot	proving
	right	clinton
	would	classy
	and	#kochspiracy
	union	#tlot
	teachers	democrat
	#debunked	@mmfa
	doctors	eaglery
	democrats	@readyaimshoot
	this	chyron
	you	loony

4.3 Content-based Clustering

We are confident that the combination of the G-test and odds weighting outperforms *tf-idf* in terms of capturing a user's emotive approach to an issue, but the clustering based upon the cosine similarity of odds weighted vectors does not exactly match our intuition (see Figure 5). It is, however, promising. Ignoring the clear outlier, *sunshineejc*, the clustering successively merges first the pro-union users before adding the pro-Walker duo *brooksbayne* and *diggrbiii* and merging with the final pro-Walker user, *conservativeind*. There are not two distinct clusters, but the successive agglomerations generally matches our intuition.

The severity of *sunshineejc*'s outlier status, however, is disturbing. Qualitatively, *sunshineejc*'s approach to the debate is not as radically different from the approaches of all other users as this clustering suggests. *sunshineejc*'s tweets frame Republicans as defenders of rich people aiming to decimate the working class. The rhetoric is much more vitriolic than that of, for example, *supermanhotmale*, but the two users are certainly similar. An examination of the odds-weighted term vectors reveals that the terms that most distinguish *sunshineejc* from all other users include terms such as *hate*, *poor*, *#corporatewelfare*, *entitlements*, *#koch*, and *broke*. These terms suggest that *sunshineejc*'s singular focus on class distinctions and class anger does not fit into the framework provided by the other users. To hu-

	odds	tf-idf
aflcio	#statesos	#statesos
	#wearewi	#notmywi
	#notmywi	#standupoh
	solidarity	afl-cio
	capitol	blog
	#standupoh	vp
	afl-cio	biden
	workers	#wearewi
	rally	@melissaryan
	now	#humanrights
	today	statehouse
	labor	@wisaflcio
	#solidaritywi	@tulaconnell
	#solidarity	officers
diggrbiii	#wiunion	@diggrbiii
	#tcot	@karoli
	@karoli	#debunked
	in	#tcot
	#tlot	doctors
	@diggrbiii	#tlot
	was	democrats
	democrats	was
	#debunked	have

Table 4: Effect of odds and tf-idf weighting, with G-test

man readers, *sunshineejc*'s attitude is clearly similar to that of, for example, *supermanhotmale*, but *sunshineejc*'s vitriolic rhetoric and singular class focus radically differentiate it from all other users simply because no other users use comparable rhetoric. Although this case reflects a limitation of this content-based approach, the limitation is exacerbated dramatically by the fact that none of the other seven users focus on the class dynamics of the issue at a level comparable to *sunshineejc*, which naturally means *sunshineejc* is an outlier.

4.4 Hashtags

Hashtag-based clustering yields dramatically different results depending upon whether SimRank or the weighted Jaccard similarity coefficient is used. Using our notions of support for or opposition to unions as the reference for successful clustering, weighted Jaccard similarity clustering outperforms SimRank clustering. In fact, the Jaccard clustering matches the expected clustering exactly except for *sunshineejc*, which is again a clear outlier. This outlier status is attributable to *sunshineejc*'s liberal use of hashtags. Each tweet from *sunshineejc* includes at least four hashtags, and some include up to eight. The other users, who use at most half that number of hashtags, can thus overlap with at most half of *sunshineejc*'s hashtags.

SimRank clustering should not be dismissed, however. Although the clustering does not match the expected clustering exactly, clustering quality is greater. The two clusters have greater internal similarities and a lesser external similarity. It is interesting to note that *sunshineejc* switched from being an outlier to becoming a member of the closest pair of users. This switch is understandable, however, given that SimRank, unlike the weighted Jaccard similarity coefficient, does not "penalize" a user (reduce similarities between the user and all other users) for using hashtags no other users use, as SimRank random walkers always randomly teleport when reaching a dead-end. In fact, *sunshineejc*'s liberal use of hashtags opens many paths between *sunshineejc* and all other users, improving its structural-context similarity with other users. Conversely, *supermanhotmale*'s appearance as an outlier is attributable to the relatively scarce use of hashtags. The reasons for *sunshineejc*'s entry into the pro-Walker cluster and the corresponding entry of *diggrbiii* into the pro-union cluster are not immediately obvious, however.

Although hashtag-based clustering produces intuitive results, we do not feel that hashtags by themselves should be used to cluster users for several reasons. First, hashtags are performative – they represent conscious decisions on the part of the user to include a tweet in a particular conversation. Theoretically, this makes them less reliable as indicators of a user's actual emotive alignment than links (which come from other users) and raw content (which is the actual currency of the user's communication). Relying excessively on hashtags also critically limits the potential of this model to be generalized to other domains. Furthermore, we have seen how a user's particular hashtag usage can dramatically influence that user's similarity to other users. Since hashtag usage is particular to a user, we cannot trust hashtag analysis to be consistent.

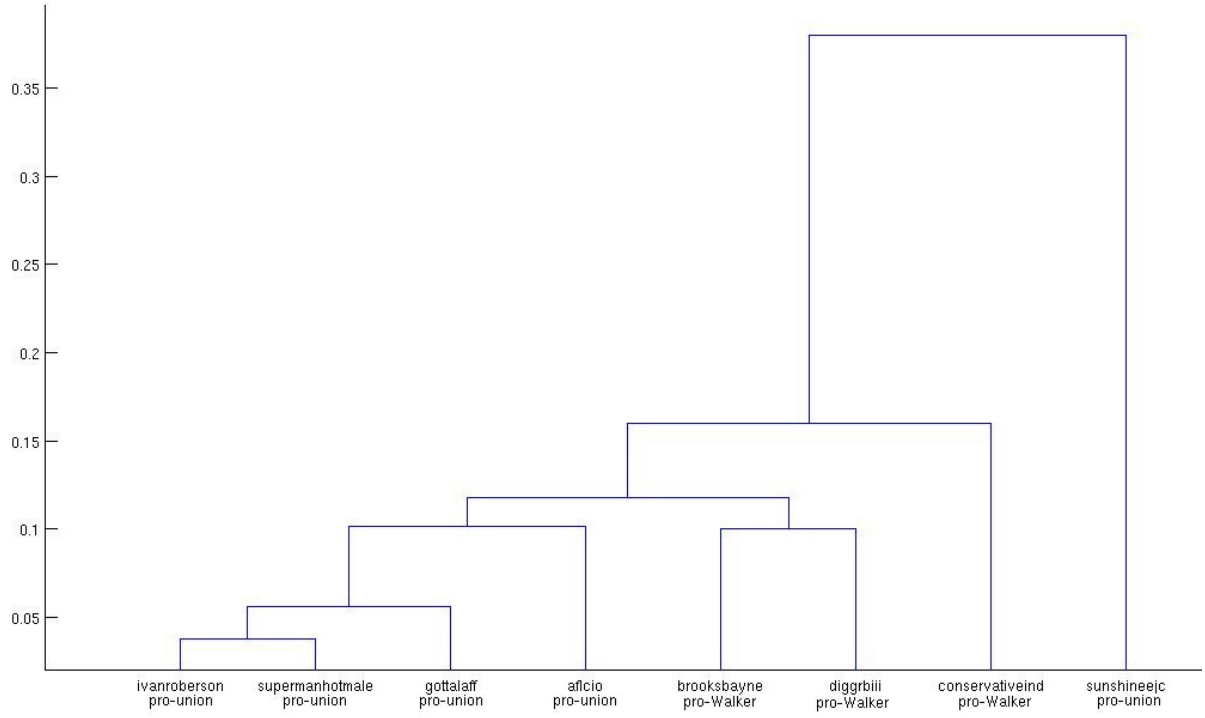

Figure 1: Odds-weighted cosine similarity-based clustering, with G-test

4.5 Mentions

Mentions-based clustering does not produce intuitive results. In fact, even with our knowledge of the users, it is difficult to substantiate an intuition behind any aspect of the clustering (see Figure 2).

It is possible that this seemingly non-meaningful clustering results from the binary use of mentions. We have observed that mentions encode much greater polarity than hashtags, especially when the dialogue is characterized more by attacking others and defending one's self than by the exchange of information. Users direct tweets at other users they either emphatically agree or forcefully disagree with, and little in between. In the aggregate, this mixture may confuse the attempt to measure similarity in approach as it is impossible to determine whether a user is mentioned in agreement or anger.

Analyzing the mentions-based clustering is complicated by the nature of the similarity measure. The effectiveness of this qualitative assessment relies on our ability to understand what we're clustering and how we're clustering it. That is, when we have an intuitive sense for how users *should* cluster, and when we have the ability to determine why our methods cluster users the way they do, we can evaluate both the performance of our model and the reasons behind this performance. Since we focus our analysis on a clustering of only eight users, understanding these users' approaches to the debate has been manageable. Furthermore, since comparing users has relied solely on the content produced by those users, understanding why the model clusters the way it does has also been possible.

Cultivating an intuitive sense for how users should cluster based upon mentions is not a manageable task, though, as it requires understanding not the eight users we have chosen to focus on but rather the manner in which all other users in the corpus engage with these eight users. Similarly, SimRank's recursive structure makes it exceedingly difficult to determine why users cluster as they do. On a superficial level, we understand that SimRank finds users

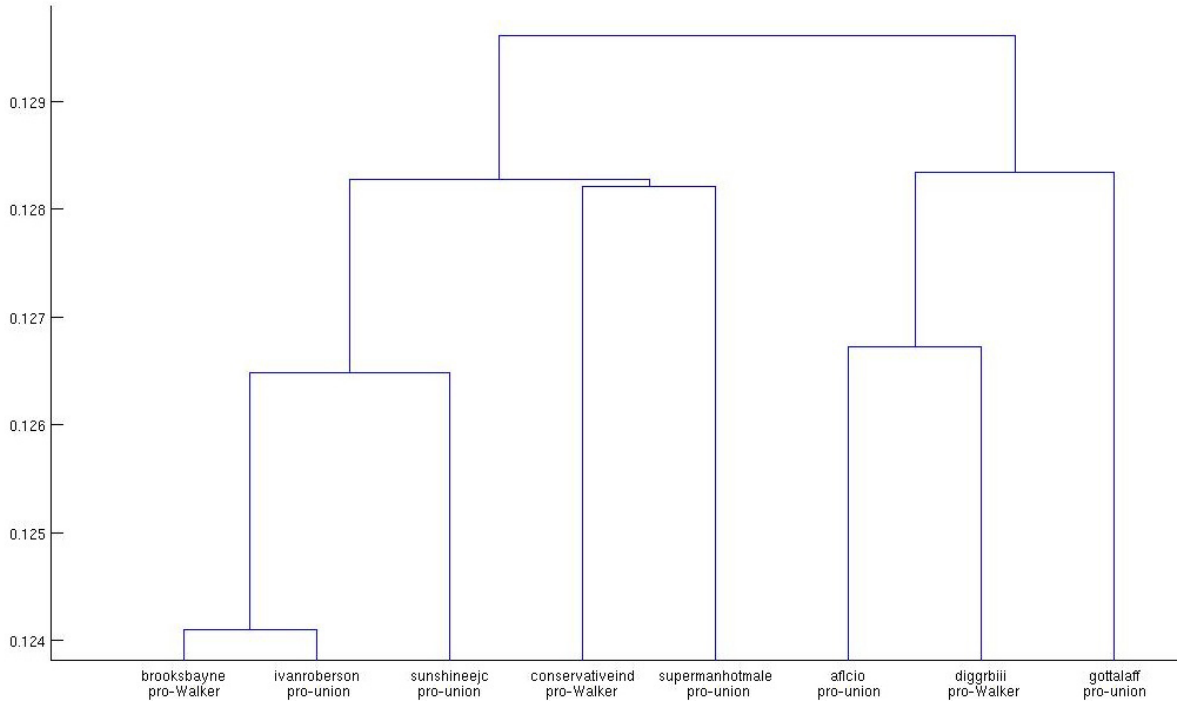

Figure 2: SimRank-based mentions clustering

a and b similar if users that mention a and b are similar. Concretely understanding why certain users are similar and how that similarity cascades to increase other similarities, however, is difficult.

Although we cannot find meaning in the clustering produced by mentions-based similarity measures, we do not believe that the mentions-based clustering is absent of meaning – we simply do not know what it is yet. In the next section, we evaluate whether combining mentions-based similarity measures with content-based similarity measures can improve the clustering and shed light on the meaning encoded in a mentions-based clustering.

4.6 Combining Similarity Measures

We examine two dendrograms based on combined similarity measures, starting with the content-hashtag similarity measure. The optimal clustering coefficient was achieved at weighting hashtag-based Jaccard similarities by 0.7 and content-based cosine similarities by 0.3, then summing them. In this clustering (Figure 5), *sunshineejc* is no longer a complete outlier but is instead included in the wrong cluster. It is unclear whether this is an improvement, but the change does demonstrate the potential of using similarity measures that leverage different aspects of the content to help contextualize a user whose content is considerably different from all other content. More promisingly, the content-hashtag clustering does cluster users into two distinct clusters. This improvement is unsurprising given the quality and structure of the clustering based purely on hashtags.

For the content-mentions clustering (Figure 6), the optimal cluster coefficient was again achieved with a weight of 0.7 on the mentions-based SimRank similarities and a weight of 0.3 on the content-based similarities. At first glance, the clustering does not seem meaningful. A closer look at the users clustered, however, reveals that the users in the first cluster, *ivanroberson*, *supermanhotmale*, and *aflcio*, are the only users who dedicate a significant proportion of their tweets to on-the-ground reporting and commentary, while all other users engage in the

conversation in an abstract, armchair-commentary style. This is significant because while previous clusterings identify the users that have similar base attitudes towards the conflict, they do not distinguish users based upon *how* they expressed these attitudes. That is, no previous clustering distinguished between attitudes that were expressed via a discussion of the actual mechanics of the protests or via a more general commentary on the theoretical underpinnings of the controversy.

Note that since all of the users in the first cluster (*ivanroberson*, *supermanhotmale*, and *aflcio*) support the unions, it is unclear whether the clustering would have been as effective had some on-the-ground commentary users supported Gov. Walker. Furthermore, it is unclear whether this performance generalizes to other topics or domains or whether this is a pecuiliarity of the Wisconsin dataset that we have discovered.

5 Looking Ahead

We chose the Wisconsin corpus for our case study for a reason. The incredible polarity of the issue – the fact that a basic "a vs. b" narrative structure holds across the corpus – allows us to use "pro-union or pro-Walker" as a reasonable approximation of a user's frame. This quality provides us with a reference for evaluating the results of our model.

The intuitive meaning captured by the content-mentions clustering reminds us, however, that this is a simplification of the actual topography of the space. It reminds us of the core difficulty of this problem: it is hard to define, let alone quantify, a user's "approach" to an issue. Basic political bias is certainly a part of it, but there are many more increasingly subtle aspects of a user's tweets that define how that user "frames" an issue. Whether the user engages with the controversy from the front lines or from an armchair is just another valid aspect of a "frame."

Given that a user's approach to an issue is multifaceted, we cannot summarize it with one similarity measure. Instead, we propose developing a system that clusters users hierarchically based upon different similarity measures corresponding to different aspects of the user's approach.

The current model's ability to achieve this goal is rudimentary but demonstrates the concept well. Leveraging both content-hashtag clustering and content-mentions clustering, we could potentially engage in two-staged clustering in order to express the topography of the space as meaningfully as possible: content-hashtag clustering can be used to divide users into to clusters based upon their attitudes towards the topic, and the clusters produced by this clustering can be further subdivided by how the attitude is conveyed by using the results of a content-mentions clustering. With further research, we can identify other features that define a user's "frame" and leverage other aspects of the data to tell us how the users relate to one another with regard to that feature.

The narrowness of the scope required for a meaningful qualitative analysis will always limit the effectiveness of such an analysis and skew our conception of the model's performance. Therefore, it is important to develop a way to apply a comprehensive quantitative analysis to the model to verify and augment the qualitative analysis. There are two strategies that could be used to qualitatively assess the model. First, we could use hashtag-based clustering as a pseudo "ground truth" on which to evaluate the more general content-based clustering results. Second, we could leverage crowd-sourcing platforms such as MechanicalTurk to compile assessments of how meaningful different, more comprehensive, clusterings are.

Furthermore, additional research is needed to determine whether this model is generalizable. Twitter is a unique medium, and the model currently leverages much of that uniqueness, particularly the interesting link structures that result from the mix of performative and functional roles that links play. Research evaluating the model's performance in clustering content that is less polarized with link structures that are more functional would be valuable.

Much work is still required, but this research already demonstrates the powerful symbiosis of quantitative and qualitative analyses. The development of an effective model and the understanding of social data are co-dependent, iterative processes: as we refine our understanding of the data, we can improve our model, and vice-versa. We began the case study with an understanding of how eight users qualitatively engage with the Wisconsin protest contro-

versy, and we ended with a general model that we feel effectively clusters users based upon the user's engagement with the issue. We can now, in turn, leverage the model to help us understand how the 36,111 users we *didn't* analyze in the case study relate to each other.

Social phenomena, though always characterized by rich complexities, are now accompanied by ever greater mountains of data. To understand the complex dynamics of an issue, it is no longer enough to simply code up an intuitive model or study the raw data. Only an understanding of the data can fuel the development of an effective model, and only with the help of a model can we identify structure and meaning in mountains of data.

Notes

[1]Gtz M, Leskovec J, McGlohon M, and Faloutsos C. 2009. Modeling blog dynamics. *Proceedings of the Third International ICWSM Conference*, pages 2633, San Jose, CA, May. Association for the Advancement of Artificial Intelligence.

[2]Leskovec J, Backstrom L, and Kleinberg J. 2009. Meme-tracking and the dynamics of the news cycle. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 497 506, New York. ACM.

[3]Teevan J, Ramage D, and Morris MR. 2011. #twittersearch: A comparison of microblog search and web search. *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM 11, pages 3544, New York. ACM.

[4]Huang J, Thornton KM, and Efthimiadis EN. 2010. Conversational tagging in twitter. *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, HT 10, pages 173178, New York, NY, USA. ACM.

[5]Cha M, Haddadi H, Benevenuto F, and Gummadi KP. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, Washington DC, USA, May.

[6]Turney PD and Pantel P. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141188.

[7]Jones KS. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:1121.

[8]Monroe BL, Colaresi MP, and Quinn KM. 2008. Fightin words: Lexical feature selection and evaluation for identifying the content of political conflict. Political Analysis, 16(4):372403.

[9]Woolf B. 1957. The log likelihood ratio test (the g-test). *Annals of Human Genetics*, 21(4):397409.

[10]Dunning T. 1993. Accurate methods for the statstics of surprise and coincidence. *Computational Linguistics*, 19(1):6174.

[11]Turney PD and Pantel P. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141188.

[12]Rijsbergen CJ. 1979. *Information Retrieval*. Butterworth.

[13]Jeh G and Widom J. 2001. Simrank: A measure of structural-context similarity. Technical Report 2001-41, Stanford InfoLab.

[14]Jeh G and Widom J. 2001. Simrank: A measure of structural-context similarity. Technical Report 2001-41, Stanford InfoLab.

[15]Hinton E and Kleefeld E. 2011. The wisconsin union struggle timeline. *Talking Points Memo*.

[16]Isthmus: The Daily Page. 2011. Wisconsin capital protests: A 13-day timeline. *Isthmus: The Daily Page.*

[17]Wisconsin State Journal. 2011. Timeline of budget bill protests. *Wisconsin State Journal*.

[18]Sebastian M. 2011. 3 ways social media is fueling the protests in wisconsin. *Ragans PR Daily*, February.

[19]Knutsen K. 2011. A guide to social media campaigns

against scott walkers agenda for wisconsin public unions. *Isthmus: The Daily Page*, February.

6 Advisers

Dan Jurafsky
Official Adviser
Professor, Linguistics & CS

Chris Potts
Working Adviser
Associate Professor, Linguistics

Approach	Example tweet
front-line support, solidarity	"We are ONE! Show solidarity with teachers fire fighters nurses & all under attack in the states. Text ONE to 235246. #StateSOS #WIunion" –aflcio
protestors as heroes	"'The test of a true good leader is putting yourself on the front line. Thank you.' Thank the #wiunion 14 senators http://on.fb.me/eH0RtV" –aflcio
criticism of Gov. Walker	"RT @Nanbp: @GovWalker #fail MRT @evale72:While telling working families need 2"sacrifice"Walker outsourced jobs 2 India #WIunion" –ivanroberson
focus on Koch brothers	"Hey Walker How Much Money Do The Koch Bros have to pay you to stomp on the good People of America? HOW MUCH? #WIUnion" –supermanhotmale
criticism of media	"Nominate for Best Actor in a Supporting Role: #WIunion coverage by UStream and http://qik.com/Brandzel Brandzel #Oscars" –gottalaff
class struggle	"Republicans dont hate Entitlements They Just hate When Poor people Get Them #KOCH #CorporateWelfare #WIunion #RecallWalker #P2 #sgp #tlot" –sunshineejc
teachers as victims	"Dear #WI Kids Your Teachers Got Fired b/c They Didn't Want a Pay Cut to Give Tax Cuts to the Corporation That Laid Off Your Dad #WIunion" –thenewdeal
teachers as villains	"The protesting teachers of #WI are teaching kids to LIE abandon responsibility & hold others HOSTAGE #wi #wiunion #FIRETHEMALL" –conservativeind
demonization of unions	"#BigLabor 's Legacy of #VIOLENCE http://bit.ly/brDuEr #twisters #tcot #wi #oh #wiunion #ohunion #ocra #p2" –conservativeind
protesters as socialists	"it's hilarious how these socialists think they have a right to an unlimited supply of *your* money. #tcot #wiunion #p2" –brooksbayne
unions as irresponsible, greedy	"RT @lheal: I'm showing #solidarity w/ #wiunion teachers by ignoring my kids until someone pays for my health care & retirement. #p2 #wi" –conservativeind

Table 1: Examples of approaches to the controversy with accompanying example tweets. All tweets are either clearly pro-union or pro-Gov. Walker, but there is a rich variety of ways this support is expressed.

User	Alignment	Notes
aflcio	unions	Relays a lot of "on the ground" information about the protests. Frames unions as heroes. Relays a sense of worker solidarity, excitement regarding the protests, and defiance of Gov. Walker.
brooksbayne	Gov. Walker	Frames unions as greedy and their demands as unfounded. Tone is often caustic and mocking. Uses Marxist language when describing union supporters.
conservativeind	Gov. Walker	Relays similar sentiments to brooksbayne, but does so in a combative, not mocking, way, and without Marxist language. Makes the case that unions do not deserve the bargaining rights.
diggrbiii	Gov. Walker	More focus on Gov. Walker than other union opponents, rhetoric more toned-down.
gottalaff	unions	More commentary on media behavior than other pro-union users, mostly to argue unions are not receiving fair coverage.
ivanroberson	unions	Exclusively retweets. Amalgamation of many styles of union support rhetoric, focusing on relaying front-line comments and criticizing Gov. Walker.
sunshineejc	unions	Frames controversy as class conflict, pitting rich republicans vs. middle and lower class workers.
supermanhotmale	unions	Similar approach to sunshineejc, but anger directed at Koch brothers and Gov. Walker specifically, interspersed with exclamations of support for union protesters.

Table 2: Qualitative assessment of user's tweets

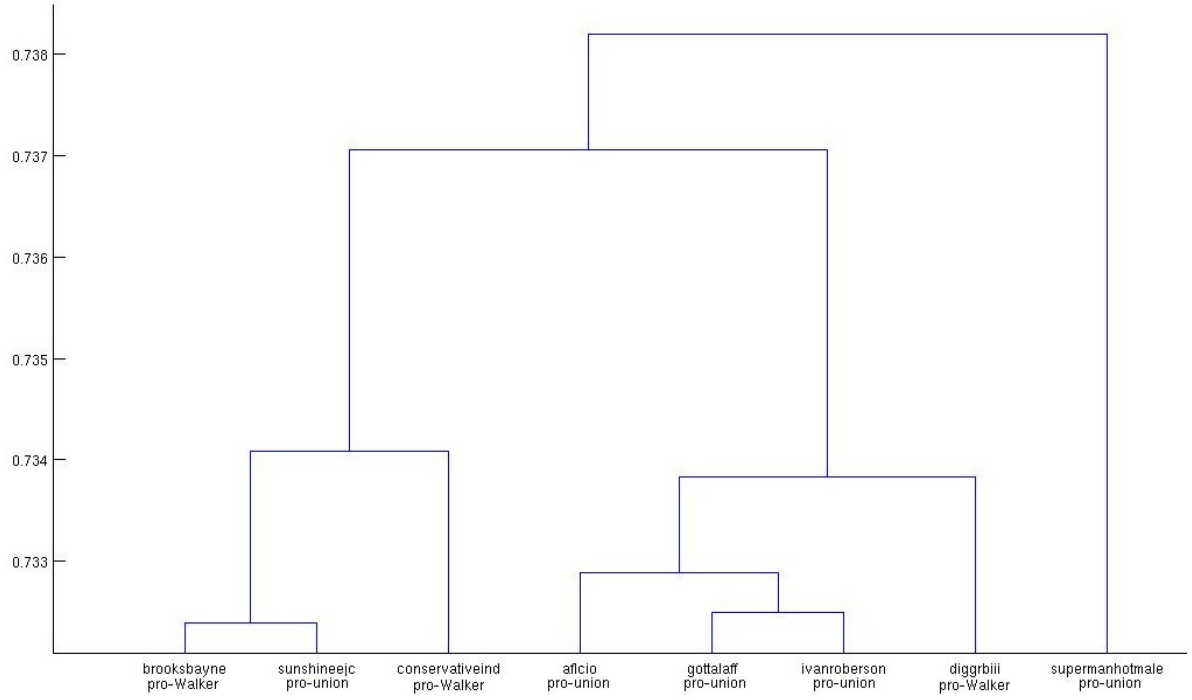

Figure 3: SimRank-based hashtag clustering

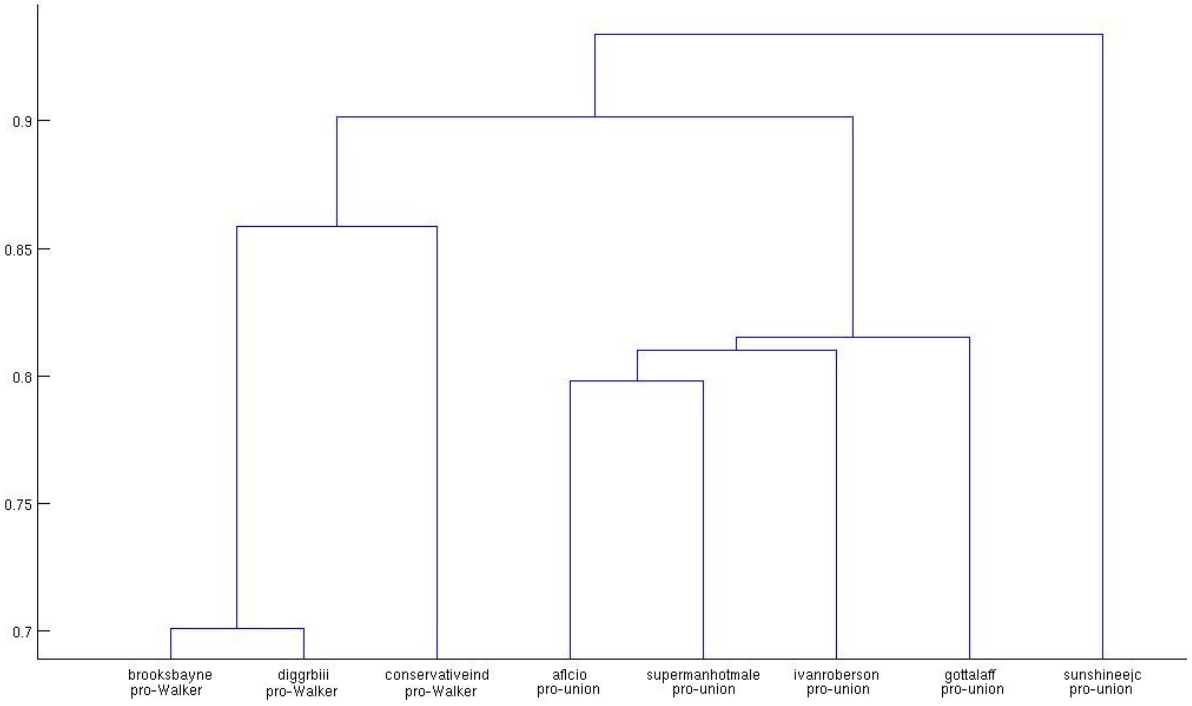

Figure 4: Weighted Jaccard similarity coefficient-based hashtag clustering

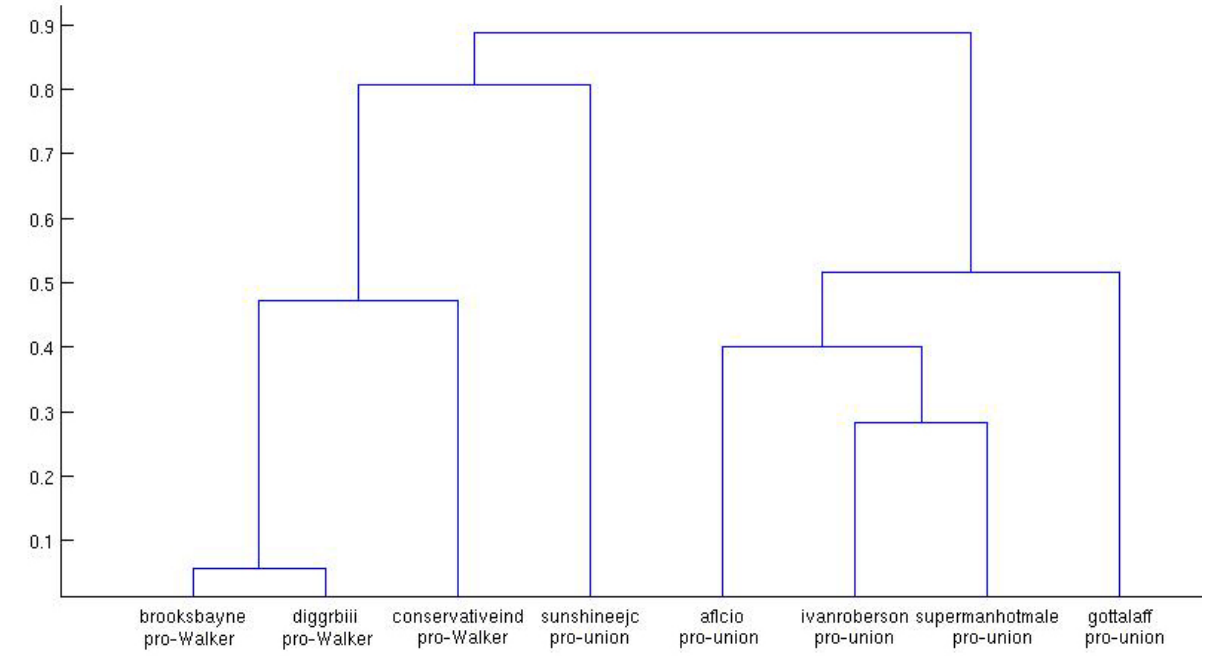

Figure 5: Clustering based on a combination of hashtag- and content-based similarity

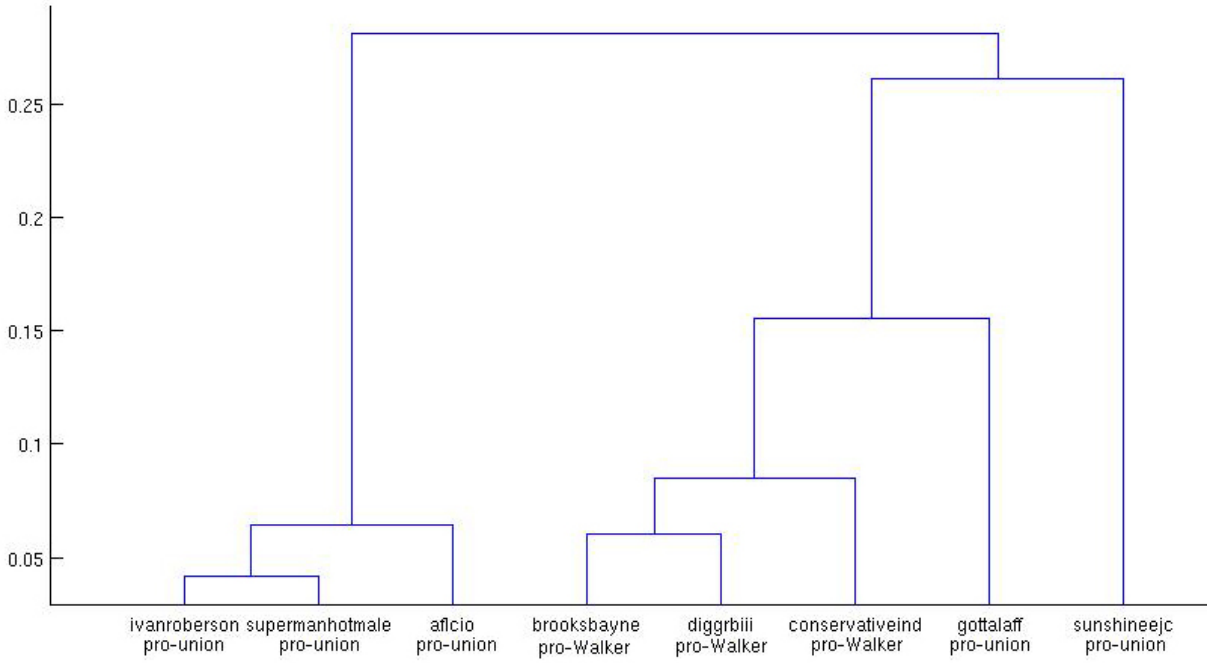

Figure 6: Clustering based on a combination of mentions- and content-based similarity