What Edited Retweets Reveal about Online Political Discourse

Eni Mustafaraj and **Panagiotis Takis Metaxas**
Computer Science Department
Wellesley College
emustafa, pmetaxas@wellesley.edu

Abstract

How widespread is the phenomenon of commenting or editing a tweet in the practice of retweeting by members of political communities in Twitter? What is the nature of comments (agree/disagree), or of edits (change audience, change meaning, curate content). Being able to answer these questions will provide knowledge that will help answering other questions such as: what are the topics, events, people that attract more discussion (in forms of commenting) or controversy (agree/disagree)? Who are the users who engage in the processing of curating content by inserting hashtags or adding links? Which political community shows more enthusiasm for an issue and how broad is the base of engaged users? How can detection of agreement/disagreement in conversations inform sentiment analysis - the technique used to make predictions (who will win an election) or support insightful analytics (which policy issue resonates more with constituents). We argue that is necessary to go beyond the much-adopted aggregate text analysis of the volume of tweets, in order to discover and understand phenomena at the level of single tweets. This becomes important in the light of the increase in the number of human-mimicking bots in Twitter. Genuine interaction and engagement can be better measured by analyzing tweets that display signs of human intervention. Editing the text of an original tweet before it is retweeted, could reveal mindful user engagement with the content, and therefore, would allow us to perform sampling among real human users. This paper presents work in progress that deals with the challenges of discovering retweets that contain comments or edits, and outlines a machine-learning based strategy for classifying the nature of such comments.

1 Introduction

Retweeting is an important practice on Twitter that makes user engagement with content generated by other users visible, and displays the aggregated users' interest in certain events or topics. Indeed, Twitter itself uses the number of retweets as a way to rank relevant tweets shown in search results. An analysis we performed in a dataset of 50 million tweets (for details refer to Section 2), revealed that 16% of the daily messages in Twitter are retweets. However, the same analysis for two specific data sets of tweets (about political events) revealed a different retweeting pattern, with 43% and 39% (respectively) of all messages being retweets. This high retweeting volume suggests that users of these communities are retweeting more frequently than average Twitter users, and this raises several interesting questions on the nature of this practice inside these communities.

What's in a tweet?	count	%
A retweet (has 'RT @' or 'via @')	102,072	43.5%
A link	100,584	42.9%
A hashtag	83,461	35.6%
A reply (starts with @user)	16,486	7.0%
A mention (contains @user)	6,426	2.7%
None of the above (only text)	45,485	19.4%

Table 1: An analysis of the content of tweets from the MAsen10 data set (collected during the Massachusetts Senate Special Election 2010, explained in Section 2), with a total of 234,697 tweets. Because tweets might contain several of these fields at the same time (e.g., a link and a hashtag), percentages don't sum to 100%. The numbers show presence of a field in a tweet, but don't count multiple instances, e.g., there might be more than one hashtag in a tweet. It can be observed that retweets are dominant in this dataset.

Previous literature on the retweeting practice (Boyd, Golder, and Lotan 2010) has emphasized two major reasons for retweeting: information diffusion and participation in a diffused conversation. In our past research (Metaxas and Mustafaraj 2010), we have uncovered how Twitter was being used to diffuse false information via a combination of replies and retweets. Inspired by that research, a web service that tries to detect in real-time such attempts of information diffusion (called astro-turfing, because they are engineered to resemble genuine grass-root support for a cause or event) has been built (Ratkiewicz et al. 2010), demonstrating that these attempts are common, especially around election time. In the light of published research that claims that Twitter could be successfully used to predict election results (Tumasjan et al. 2010), it is important to quantify and qualify the nature of information diffusion and user engagement around such topics and events in Twitter.

There are several ways in which researchers are addressing questions about information diffusion in Twitter. A common way is to study how URLs spread, as shown in (Bakshy et al. 2011). Another way is to study adoption and diffusion of hashtags (Romero, Meeder, and Kleinberg 2011). The dataset we used in (Metaxas and Mustafaraj 2010) (explained in details in Section 2), showed that retweets were the most common practice of information diffusion inside the community surrounding a political election event, see Table 1, which prompted us to study it more thoroughly.

There are at least two technical challenges to be overcome from the start:

- The dataset was collected in January 2010, when the retweeting practice was not yet adequately supported by the Twitter API. This means that to find retweets of a given tweet, we need to use text similarity techniques to match potentially related tweets.
- While a large number of users retweet a tweet verbatim, another considerable quantity includes a comment while preserving the original text to different extents. Since there is no universal retweeting style, finding which part of the tweet text belongs to the original sender and which to the retweeter is not trivial.

1.1 Finding Retweets

Twitter recognized the problem of lacking support for retweets and introduced a retweet button and API support for retweets (in the form of a new field `retweeted_status`, which enables linking to the original sender and original text. The `retweeted_status` field makes it easy to compare the original tweet with the retweeted version and also to keep track of how many times a tweet was retweeted, by which users, etc. However, because the retweet button of the Twitter web client doesn't allow a user to edit (or quote) the original tweet, not all users make use of it. Our analysis of 50 million tweets revealed that in the group of retweets, 42% contained the `retweeted_status` API field, but 58% did not. In addition, datasets collected before these changes don't have information that can match a retweet with the original tweet. Solving the problem of finding all retweets (which don't use API information) for any given tweet is important, especially, since the phenomenon is so widespread. However, we do not address this problem in this paper. For our purposes, we created a set of heuristics to combine n-gram analysis and user mentions to perform the matching of tweets with their retweets.

1.2 Verbatim versus Edited Retweets

Keeping in mind the two reasons for retweeting, *information diffusion* and *participation in a diffuse conversation*, it is to be expected that at least two types of retweets exist:

(a) retweets that copy verbatim the original tweet and only add the original author. These serve mostly the goal of information diffusion.

(b) retweets that quote the original tweet and its author, but that add other text (in the form of a comment) with the goal to participate in the diffused conversation. Because of the 140 character limit on tweets, users must make different kinds of edits to a tweet to free space for their comments. We use the term edited retweets to refer to all retweets not covered by the verbatim retweet definition.

Sender	Tweet
da***	Scott Brown supported by Rush Limbaugh & Sarah Palin. *He's way too extreme for Massachusetts* #MAsen
Jo***	**Sounds good to me!** RT @da***: Scott Brown supported by Rush Limbaugh & Sarah Palin. *Perfect for America!* #MAsen

Table 2: An original tweet and an edited retweet. Notice that the retweeter has deleted the last sentence (italized) and inserted a new one, which completely changes the meaning of the original tweet. Then, it has added an own comment (bold) to show agreement.

Our working hypothesis is that retweeting verbatim displays potential complete agreement with the original sender by the part of the retweeter. Discovering such retweets might be important in terms of discovering influental users who enjoy the trust of a community.

On the other hand, the nature of retweets that contain an edited version of the original tweet and additional text is an issue that needs to be explored in the data at hand. One of the most intriguing cases we have come across in shown in Table 2.

The retweeter has deliberately changed the meaning of the original tweet, by deleting a sentence and replacing it with another. Because the tweet contains a hashtag, it was theoretically possible to be seen by users who don't follow the original poster, but keep track of the hashtag stream. Everyone who sees only the retweeted version, and is not aware of the political orientation of the posters, would assume from this message that the two posters agree, which is not true.

By misquoting original tweets in the process of retweeting, it is possible to spread false ifomation and damage the credibility of the original poster. Twitter has partially responded to such a concern, with the introduction of the retweet button. However, since a large majority of users (58%) still retweets in other ways, which allow the editing of original retweets (often necessary, to overcome the limits of 140 characters), the introduction of a feature that facilitates finding the original tweet for a retweet, would be a useful addition to the Twitter eco-system.

Luckily, the example shown in Table 2 is not representative of the kind of editing process during retweeting. Far more common are commenting on the tweet or curating the tweet (with curation is meant the introduction of new hashtags, links, or mentions). We discuss the editing process in Section 3 and efforts to classify the agreement or disagreement nature of introduced comments in Section 4.

2 Data Collection

In this paper, we refer to three different data sets: MAsen10, #tcot and #p2, and November2010. These three datasets were collected in different ways and contain different kinds of data. Our most important dataset is MAsen10 and most of the analysis in this paper is based upon that set.

Come to NHTPP to make phone calls. Help get Scott Brown (R) elected to the Mass. U.S. Senate . http://bit.ly/5rHMBn
RT @twteaparty: RT @coo***: Come to NHTPP to make phone calls Help get Scott Brown (R) elected to the Mass. U.S. Senate . http:/ ...
RT @By***: Massachusetts: The Scott Brown ’moneybomb’ keeps exploding. http://tinyurl.com/ycvzmfj
RT @Te***: @Ma*** RETWEET URGENT URGENT REVOLUTION SCOTT BROWN MASSACHSETTS SENATE DONATE VOTE JAN 19 FOR NO GOV HC FREEDOM 1ST

Table 3: Examples of tweets that show the use of Twitter to coordinate support for Scott Brown. Some account names are obfuscated to protect users’ privacy.

2.1 The MAsen10 Dataset

On January 19, 2010, the US state of Massachusetts held a special election to fill the vacant seat of Senator Ted Kennedy, who had passed away in August 2009. Prior to this election, Democrats had a filibuster-proof majority in the US Senate (60 to 40) and were in the process of passing the Health Care reform that President Obama had promised during his presidential campaign, a reform strongly opposed by republicans in the House and Senate. On the ballot were the names of the democrat Martha Coakley, the acting attorney general of Massachusetts, and republican Scott Brown, a state senator. Because Massachusetts is traditionally one of the most democratic-leaning US states, there was initially no doubt that the democratic candidate would win the election. However, during the campaign, Scott Brown positioned himself as the potentially 41st vote in the Senate that will end the dominant position of democrats (who controled the White House, the Senate, and the House). This positioning brought him endorsements from Tea Party activists and other conservative groups who contributed financially to his campaign, especially through online donations (sometimes called money-bombs). Twitter was used extensively for coordination of such support, as some example tweets in Table 3 show.

Tweets were collected during the week leading to the election, between Jan 13-20, using the Twitter Streaming API, configured to retrieve near real-time tweets containing the names of either of the two candidates. The obtained corpus comprised 234,697 tweets contributed by 56,165 different Twitter accounts. Some other statistics about this dataset are shown in Table 1. The retweeting activity consitutes 43.5% of all tweets. Retweets were discovered by filtering tweets containing expressions such as ’RT @’ or ’via @’ (with or without space in between).

2.2 The #tcot and #p2 dataset

Our analysis of the MAsen10 dataset revealed that while the most used hashtag was #masen (Massachusetts Senate election) for a total of 49,710 times, at number two was positioned #tcot with 34,073 appearances and at number three #p2 with 10,884 appearances. These two hashtags stand for the two opposing political groupings in the United States: conservatives (tcot) and progressives (p2). Since these hashtags are used to mark political conversation, we used again Twitter Streaming API to gather all tweets containing one of these two hashtags, during a 20 days period in Aug-Sep 2010 (before the primary elections in several US states). A total of 652,067 tweets was collected. In this paper, this dataset is mentioned only in relation to the retweeting activity, which amounts to 39% of all tweets. This is a further proof that communities of users interested in politics engage in retweeting much more frequently than average Twitter users.

2.3 The November2010 dataset

This dataset is courtesy of the Center for Complex Networks and Systems Research at the Indiana University School of Informatics and Computing, who is whitelisted by Twitter to have access to the “gardenhose” API, which allows realtime accees to a random set of 10% of all tweets worldwide. The dataset contains almost 50 million tweets[1] collected during Oct 26 - Nov 1, 2010, the week preceding the US 2010 Midterm Congressional Elections of Nov 2, 2010. We used this dataset to find the daily average of retweets in such a large random sample of tweets. The average is pretty constant from day to day, minimum average was 16.13% and maximum 16.60%. We also used the dataset to calculate the percentage of retweets that contain the field `retweeted_status` (which allows to access the original tweet), and those that don’t. The proportions are 42.16% versus 57.84%.

3 Finding Edits in Retweets

Twitter users retweet in different ways. An additional problem is introduced when they need to comment on the original tweet, especially if its length is near the limit of 140 characters. This problem is solved differently by different users. In order to be able to study a variety of such practices, we decided to find a large set of edited retweets for a single tweet, so that all retweets were based on the same original content. Our intuition is that tweets sent by Twitter accounts with a large number of followers are more likely to be retweeted from a diverse set of users, thus, more likely to be edited to accomodate commments. The Twitter Streaming API returns tweets with a lot of contexual information, such as the number of followers of the sender. This allowed us to rank all tweets based on the number of the followers for the tweet sender. Two tweets from @BarackObama were at the top, followed by four tweets by @cnnbrk (CNN Breaking News) and four by @nytimes (the New York Times newspaper). To check whether the heuristic of finding retweets with ’RT @’ and ’via @’ offers appropriate coverage for retweets, for this small group of tweets, we decided to find retweets based on n-gram similarity. This technique discovers both retweets and other un-attributed copies of the tweet. Some of these copies belong to bot accounts, which mimic humans on Twitter by simply copying the text of tweets of well-known accounts. We found a dozen of identical copies of a tweet by @BarackObama, without any reference to him. A few copies were edited to change the meaning, for an ex-

[1] The precise number is 49,152,948 tweets.

0	Some in Mass. got ballots already marked for Republican Brown, Dem. candidate Coakley’s camp says. http://bit.ly/7HiUZe
1	RT @cnnbrk: Some in Mass. got [...] camp says. http://bit.ly/7HiUZe **&lt;— ?**
2	RT @cnnbrk: Some in MA got [...] for GOP [...] camp says. http://bit.ly/7HiUZe **&lt;– and it begins**
3	RT @cnnbrk: Some in Mass. got [...] ~~candidate~~ Coakley’s camp says. http://bit.ly/7HiUZe **BS starts already**
4	**She’s smoking the good shit:** RT @cnnbrk Some in Mass. got [...] camp says. ~~http://bit.ly/7HiUZe~~
5	cnnbrk: Some in Mass. got [...] camp says. ~~http://...~~ **http://bit.ly/8IUucF**
6	**should I be surprised by this?** RT @cnnbrk: Some in MA got [...] camp says. ~~http://bit.ly/7HiUZe~~
7	**Rigged!** Some in Mass.got [...] camp says. http://bit.ly/7HiUZe (via @cnnbrk)
8	**Oh HELL NO.** RT @cnnbrk: Some in Mass. got [...] Repub [...] camp says. http://bit.ly/7HiUZe
9	RT @cnnbrk: Some in Mass. got [...] camp says. **ht (cont) http://tl.gd/4d4d7**
10	**Hahaha...what is this Iran? Sour grapes.** RT @cnnbrk: Some in Mass. got [...] Coakley~~’s camp says. http://bit.ly/7HiUZe~~
11	**Crap** rt @cnnbrk: Some in Mass. got [...] camp says. http://bit.ly/7HiUZe
12	**Sounds like dirty politics.** Some in Mass. got [...] camp says. ~~http://bit.ly/7HiUZe~~via @cnnbrk
13	**RT really its come to this?!**@cnnbrk: Some in Mass. got [...] camp says. ~~http://bit.ly/7HiUZe~~
14	RT @cnnbrk Some in Ma got [...] 4 Rep [...] camp says. http://bit.ly/7HiUZe **/ur article doesn’t match ur post!?**
15	RT @cnnbrk: Some in Mass. got [...] camp says~~. http://bit.ly/7HiUZe~~ **// That’s awesome!**
16	**and let the accusations begin:** RT @cnnbrk Some in MA. got [...] camp says.http://bit.ly/7HiUZe
17	**Oh come on!** [cnnbrk] Some in Mass. got [...] camp says. http://bit.ly/7HiUZe
18	Some in Mass. got [...] camp says. ~~http://bit.ly/7HiUZe~~(via @cnnbrk)**Cheating Pubs!**
19	**O please-go suck a hanging chad and stfu!**RT@cnnbrk: sum in Mass. got [...] 4 Repub [...] camp says. ~~http://bit.ly/7HiUZe~~
20	**WTF!!** RT @cnnbrk Some in Mass. got [...] camp says. http://bit.ly/7HiUZe
21	RT @cnnbrk Some in Mass. got [...] camp says. http://bit.ly/7HiUZe **Dems blow it again**
22	**Classy** RT @cnnbrk: Some in Mass. got [...] camp says. http://bit.ly/7HiUZe
23	**Where does article say this??** @cnnbrk [...] for ~~Republican~~ Brown, ~~Dem. candidate~~ Coakley’s camp says. http://bit.ly/7HiUZe/
24	**Whine ... whine** RT cnnbrk Some in Mass. got [...] camp says ~~. http://bit.ly/7HiUZe~~
25	**What?!**RT @cnnbrk: Some in Mass. got [...] camp says. http://bit.ly/7HiUZe.
26	**it’s already started** RT @cnnbrk Some in MA got [...] Rep Brown, Dem [...] camp says. http://bit.ly/7HiUZe
27	**Let the craziness commence** RT @cnnbrk Some in Mass. got [...] camp says. ~~http://bit.ly/7HiUZe~~
28	**Wow** RT @cnnbrk: Some in Mass. got [...] camp says. http://bit.ly/7HiUZe
29	**same ol** RT @cnnbrk Some in Mass. got [...] 4 [...] camp says. http://bit.ly/7HiUZe
30	**Grrr!** RT @cnnbrk: Some in Mass. got [...] camp says. http://bit.ly/7HiUZe
31	**WTH!** RT @cnnbrk: Some in Mass. got [...] camp says. http://bit.ly/7HiUZe
32	**I doubt it.** RT @cnnbrk: Some in MA got [...] Rep. [...] camp says. http://bit.ly/7HiUZe
33	**WHAT??** RT @cnnbrk Some **voters** in Mass got [...] **pre-**marked for Rep [...] camp says. http://bit.ly/7HiUZe
34	**Boo!** RT @cnnbrk: Some in Mass. got [...] camp says. http://bit.ly/7HiUZe
35	**Here we go**RT@cnnbrk: Some in Mass. got [...] camp says. http://bit.ly/7HiUZe
36	**4 real?** RT @cnnbrk: Some in Mass. got [...] camp says. http://bit.ly/7HiUZe
37	**The excuses begin.** RT @cnnbrk: Some in MA got [...] Repub. [...] cand. Coakley’s camp says. http://bit.ly/7HiUZe
38	**Result delay ploy?...**RT @cnnbrk: Some in Mass. got [...] camp says. ~~http://bit.ly/7HiUZe~~
39	RT @cnnbrk Some in Mass. got [...] camp says. ~~http://bit.ly/7HiUZe~~ **http://tinyurl.com/yau3uwk**
40	**Um, what?** RT @cnnbrk: Some in Mass. got [...] Repub. [...] camp says. http://bit.ly/7HiUZe
41	Some in Mass. got [...] camp says. http://bit.ly/7HiUZe (via @cnnbrk) **- hmm**

Table 4: One original tweet, followed by 41 edited retweets. Attribution to original sender is in orange, comment by retweeter in bold aquamarine, edit operations in text in magenta. For space reason, we use [...] to replace verbatim text.

User	Tweet text	# tweets	Label
UserA	Scott Brown is a winner tonight	1	none
UserB	**The People won too!!** RT @UserA: Scott Brown is a winner tonight	15	C
UserC	**The People won too!!** RT @UserA: Scott Brown is a winner tonight (via @UserB)	15	C
UserD	RT @UserA: Scott Brown is a winner tonight	16	C
UserE	**Awesome! “America Needs Change” right Obama supporters?!? How’s that for change?** RT @UserA Scott Brown is a winner tonight	1	none

Table 5: A group of 4 retweets around a single tweet. Because the political orientation of some users can be inferred from their past behavior, and there is one verbatim retweet in this group which signals agreement with the original tweet, it can be inferred that other retweets also show agreement.

ample, compare the following two tweets, were the edits are shown in red italic:

This is it-the polls are open in MA. Your calls can help send Martha Coakley to the Senate: http://bit.ly/7-5 #MASen Pls RT

This is it-the polls are open in MA. Your calls can help send Martha Coakley *back to the DA office and @ScottBrownMA to the Senate* #MASen

To find the verbatim retweets in the set of all retweets, we created combinations of the original tweet text with common ways of signalizing retweets ('RT @user' or 'via @user). Removing them from the set leaves us with the set of edited retweets. The ratio of edited retweets to all retweets ranged from 5 to 30%. This shows that users most frequently engage in information diffusion by retweeting verbatim. However, since usually the volume of tweets on a topic is very large, a ratio of edited retweets up to 30% indicates a sizable participation in diffused conversations, which is not to be ignored. In the following, we analyze all the retweets for the tweet with the largest number of edited retweets (41 out of 139 retweets). This tweet was sent by @cnnbreak and is shown at row 0 in Table 4. For the sake of completeness, we include in the table all edited retweets, using colors to distinguish different parts of the editing process. There are several observations to take away from this table:

1. The majority of these retweets contains the attribution 'RT @user' (33 out of 41), though with some minor variations concerning the use of colon, @ sign, and white space.
2. The majority of comments are added at the beginning, and 25 of them can be extracted by simply splitting at the 'RT @' phrase. However, for the other comments, several unique rules that combine splitting with the comparison of the original tweet with the retweet might be needed.
3. In 13 retweets, the article URL was dropped from the text (shown by striked-through text in magenta). This should be taken into consideration when information diffusion is studied by following the spread of links, and maybe an approach that combines unique phrases with URLs might offer better coverage of the phenomenon.
4. In 11 retweets, the users have opted to edit the internal text of the tweet (by shortening words or deleting them), in order to preserve the article URL. These kind of edits complicate the issue of distinguishing between added comments and necessity edits, especially in cases when the comments are not at the start of the tweet.
5. Comments are generally short, averaging 2.8 words per commment. 14 comments consist of a single word, and yet, that is often sufficient to convey the sentiment or opinion of the user on the reported story, which ranges from disbelief or concern to dismissal and derision.

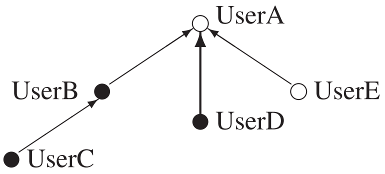

Figure 1: A retweet graph. Nodes are users and arrows are retweets. Filled nodes indicate users for which we know the political orientation, in this case C for conservatives.

From these observations, it follows that to be able to know what real people think about news events, people, policy issues, etc., we will need to extract their comments from edited retweets. Once we have such comments, they can be seen as individual contributions in the diffused conversation about the same topic (the tweet beeing retweeted). We can then proceed performing classification of such "dialogue" acts as in (Stolcke et al. 2000). However, the categories for dialogue acts described in (Stolcke et al. 2000) do not necessarely apply to this kind of diffused conversation. Since our goal is to be able to answer questions such as: is it possible to predict an election, or is there support for a specific policy issue, the categories of speech acts of interest to us are agreement and disagreement, as well as positive or negative opinions. The biggest hurdle then, is to create a domain-specific traning set of instances to learn a supervised classifier. We discuss this question in more detail in the next section.

4 Classifying Comments

Examples in Table 4 indicated that comments are usually very short (an average of 2.8 words). If we were to build a learning approach based only in these textual comments, we will need a large training set to be able to provide enough vocabulary coverage. Manually creating such a training set is a laborious task. We believe that we can overcome this obstacle by using the idea central to the co-training algorithm (Blum and Mitchell 1998), the existence of two different indenpendent views for each instance: a textual view (the text of the comment) and a structural-behavioral view (the friendship network of the sender and her previous retweeting behavior). Here are the knowledge sources that allow us to make use of the second view:

- For all Twitter users, we can find the list of who they follow.
- For all Twitter users, we can access historical tweets and find out whether they have retweeted verbatim another user.
- For Twitter users who have used certain hashtags in their past tweets, we can infer their political orientation.

We make then the following assumptions, which will allow us to provide labels based upon the structural-behavioral view:

- If a user follows the original sender and has retweeted her verbatim, then he agrees with her.
- If a user shares the same political orientation with another user, as observed by their use of certain hashtags, then they agree and vice-versa.

- If two users don't share the same political orientation and they don't follow each-other on Twitter, then they potentially disagree on political issues.
- There are two opposing political positions, conservatives and liberals.[2]

To clarify how these assumptions would help us to provide labels for comments in retweets, let's examine a concrete example. Table 5 contains one original tweet from UserA and four retweets from other users. UserA has only this tweet in the MAsen10 corpus, so we don't know anything about her previous behavior or political orientation. The same situation applies to user UserE, who has written the longest comment. What label should we apply to that comment? The graph of retweets in Figure 1 shows how this will be achieved. The three other users in the retweet graph UserB, UserC, and UserD have been tweeting about the Massachusetts election and they have predominantly used the #tcot hashtag, which allows us to infer their political orientation as convervatives (C). Because the retweet from UserD is a verbatim retweet and he is following UserA in Twitter, we infer that he agrees with UserA and that they share the same political orientation. In this way, we spread the label of political orientation and of agreement in this small local network, which makes possible to acquire three examples of comments with the agreement label.

We are in the course of implementing such an algorithm and hope to be able to report results soon. Our initial explorations have shown that we will be able to find more examples of agreements than disagreement, since the two political groupings rarely communicate with each-other. However, the use of hashtags in tweets, allows the two groups to read the tweets of each-other, and from time to time, it's possible to find examples of disagreements, as shown by the following pair of tweets:

`BROWN BULLYING TACTICS WATCH: Reports of Brown signs going up on property of Coakley supporters. Hint: yard signs don't vote, Scott. #MASen`

`RT @Dem*** Brown signs goingon prop. of Coakley supporters. Hint: yard signs don't vote|`*`they dont?Dems stopped that practice?`*

By making use of the two political datasets, we predict to be able to accumulate enough training examples for both categories.

5 Conclusions

Retweeting is a common practice especially in communities of Twitter users interested in political events and topics. While the majority of retweets are verbatim copies of the original tweets, there is a sizable group of retweets that contain comments, which, if collected and analyzed, can reveal interesting insights about online political conversation. In this paper, we discussed the inherent difficulties of finding edited retweets and extracting comments from them. We then outlined an approach of making use of the structural-behavioral features of Twitter users, in order to acquire labels for such comments, which would allow us the training of a supervised classifier.

6 Acknowledgement

We are grateful to Daniel Gayo-Avello for fruitful discussions and to the anonymous reviewers for their thoughtful and valuable comments.

References

Bakshy, E.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Everyone's an influencer: quantifying influence on twitter. In WSDM, 65–74.

Blum, A., and Mitchell, T. 1998. Combining labeled and unlabeled data with co-training. In COLT, 92–100. Morgan Kaufmann Publishers.

Boyd, D.; Golder, S.; and Lotan, G. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In HICSS, 1–10.

Metaxas, P. T., and Mustafaraj, E. 2010. From obscurity to prominence in minutes: Political speech and real-time search. In WebSci10: Extending the Frontiers of Society On-Line. http://bit.ly/h3Mfld.

Ratkiewicz, J.; Conover, M.; Meiss, M.; Gonçalves, B.; Patil, S.; Flammini, A.; and Menczer, F. 2010. Detecting and tracking the spread of astroturf memes in microblog streams. CoRR abs/1011.3768.

Romero, D. M.; Meeder, B.; and Kleinberg, J. M. 2011. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In WWW, 695–704.

Stolcke, A.; Ries, K.; Coccaro, N.; Shriberg, E.; Bates, R.; Jurafsky, D.; Taylor, P.; Martin, R.; Van, C.; and dykema Marie Meteer, E. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics 26:339–373.

Tumasjan, A.; Sprenger, T.; Sandner, P. G.; and Welpe, I. M. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In Proc. of 4th ICWSM, 178–185. AAAI Press.

[2] This is particularly true for US Politics, with its two-party political system. We acknowledge that the political spectrum is continuous, but we will treat that problem in future work.