Characterizing Deleted Tweets and Their Authors

Parantapa Bhattacharya and **Niloy Ganguly**
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur, India
{parantapa,niloy}@cse.iitkgp.ernet.in

Abstract

This paper provides a detailed characterization of tweets posted and then later deleted by their authors, in the Twitter microblogging platform. Our characterization shows significant personality differences between users who delete their tweets and those who do not. We find that users who delete their tweets are more likely to be extroverted and neurotic while being less conscientious. Further, although deleted tweets contain more negative sentiment and swear words, they also show significant signs of being thoughtfully constructed.

Introduction

Online social networks have allowed their users to convey their thoughts and ideas, quickly and easily, to hundreds and thousands of others. However, in doing so they have also made their users susceptible to inadvertently exposing potentially private and embarrassing information. Once a user realizes that she regrets a post that she has made, the most common mending strategy is to delete the offending post (Sleeper et al. 2013).

In Twitter, deletion of tweets is quite widespread. According to our analysis, over 11% of tweets created, are deleted either by Twitter or the user posting them. This widespread deletion of tweets raises two interesting questions: First, is tweet deletion equally prevalent among all Twitter users or is it common only among users of a certain predisposition? Second, do tweets that are deleted later, have distinctive features, relative to tweets that are not deleted?

Understanding the above questions is fundamentally important for researchers and developers building systems, that help users manage potentially regrettable posts. What a user finds regrettable, depends heavily on her personality, her desired public image, and her general writing style. Further, the choice of deleting one of her own posts is a very personal one. Thus, it is essential for privacy concious systems to understand and take these into account when helping users take actions.

This work presents a large scale empirical study of all tweets posted and deleted by over 200 thousand users during a four week period in August 2015. Further, we undertook systematic data cleaning procedure, where we removed automated tweets, superficial deletion, and deleted retweets from our dataset. This thorough cleaning of data allowed us to discover interesting personality based characterizations of users who delete their tweets. We were also able to discover detailed linguistic differences between deleted and non-deleted tweets from the same users.

We found that users who delete their tweets are more likely to be extroverted and neurotic while also being less conscientious. Interestingly, deleted tweets are simultaneously less informative and less conversational. Surprisingly, we found that while use of profanity and swear words is significantly higher in deleted tweets, they also show significant signs of being thoughtfully constructed.

Related Work

Almuhimedi et al. (2013) presented the first large scale study of deleted tweets. They collected deleted and non-deleted tweets posted by 292 thousand users over a period of one week. The authors compared them along various dimensions and were able to find differences in attributes such as location of origin of the tweet and the Twitter client used for posting the tweet. They noted that superficial reasons (typos and rephrasings) were one of the major causes of tweet deletions.

Sleeper et al. (2013) surveyed 1,221 users via Amazon Mechanical Turk asking them to describe "one thing they had said and then later regretted" on Twitter and in the offline world. The authors categorized the stated regrets in Twitter and the offline world into eleven categories. They found that blunders, direct attacks/criticism, group reference, and revealing too much accounted for 83% of the stated causes of regret.

Few recent works (Petrovic, Osborne, and Lavrenko 2013; Bagdouri and Oard 2015; Zhou, Wang, and Chen 2016) have tried to create classifiers to predict whether a tweet will be deleted or not. Zhou et al. (2016) built a classifier for "content-identifiable regrettable tweets", which they defined as those tweets, which third party human annotators think are regrettable. They devised a set of 10 closed vocabulary features which represented words related to sensitive topics such as alcohol use, drug use, violence, etc. Using a decision tree classifier they were able to achieve a F1-score of 0.714.

# Tweets posted	17,147,771
# Tweets deleted	1,210,434 (7.05%)
# Users who posted at-least 1 tweet	194,495
# Users who deleted at-least 1 tweet	91,785 (47.19%)

Table 1: Total number of tweets posted and deleted by users in our dataset, after cleanup procedure.

Dataset

To select a representative sample of real and active Twitter users, we started by randomly selecting a set of 250,000 users whose tweets had been included in the Twitter random sample during the month of October 2014 and who also met the following criteria: their majority tweets were in English, they had posted at-least 10 tweets in their lifetime, and had at least 10 followers and 10 followees. Using the Twitter streaming API we followed these users and collected all tweets made by them as well as replies and retweets of their tweets, during the four week period of August 3, 2015 to August 30, 2015.

Since Twitter sends tweets via its streaming APIs in realtime, it has to send out status deletion notices for tweets that are deleted later.[1] As it is possible for a user to delete her tweet much later than when it was posted, we collected all status deletion notices for an additional week, that is during the five week period of August 3, 2015 to September 6, 2015. Out of the 43 million tweets thus collected, we found that 11.11% of them were later deleted.

Dataset cleanup

As we focus only on English tweets for this work, we filtered out any non-English tweets from our dataset. We used the tweet's language field for this purpose. Further, we also removed any tweets posted automatically via popular automated tweeting systems and account management tools like: RoundTeam, If this then that, Buffer, twittbot.net, fllwrs, Crowdfire App, etc. When a tweet is deleted, any retweet of the said tweet is also deleted by Twitter.[2] Since it is not possible for us to know whether a deleted retweet was deleted by the user who posted the retweet or the user who created the original tweet, we leave out analysis of deleted retweets from this work.

Twitter doesn't provide a method to edit tweets. Thus, to fix a typo or grammatical error, users have to delete the erroneous tweet and compose a new one. Since this work focuses on differences between deleted and non-deleted tweets, we ignore such superficial deletions from our analysis. To check if a tweet deletion is superficial, we followed the same methodology as (Almuhimedi et al. 2013).

Table 1 shows the final count of tweets and users in our dataset after removal of non-English tweets, automated tweets, retweets, and superficial deletions.

[1] https://dev.twitter.com/streaming/overview/messages-types#status_deletion_notices_delete

[2] https://support.twitter.com/articles/18906-deleting-a-tweet

	Deleters	Non-Deleters
# Users	91,785	102,710
Median Followers	508	375
Median Followees	403	394
Median Listed Count	2	2
Median Tweet Rate	8.85 tweets/day	4.74 tweets/day

Table 2: Differences in user attributes, between users in the deleter and non-deleter sets.

Characterizing user differences

To answer the question — does there exist any characterizing differences between users who delete tweets and those who do not — we divided users in our dataset into two subsets: (i) *non-deleters*: the set of 102 thousand users who had posted at least one tweet, but all their tweet deletions (if any) were classified as superficial deletions, and (ii) *deleters*: the set of 92 thousand users who had deleted at least one tweet, that was not a superficial deletion.

Big-Five personality traits

It is expected, that any significant differences in tweet deletion practices among the user groups (if they exist) would stem from their underlying personality differences. The Big-Five personality traits (Costa and McCrae 1992) is a system for modeling human personality along five dimensions: openness, conscientiousness, extraversion, agreeableness, and neuroticism. The five different traits refer to five non-overlapping personality traits related to human behavior. Here, we try to characterize the differences in personality traits between the deleter and non-deleter user sets along the Big-Five personality dimensions, using social and linguistic attributes.

Differences in social characteristics

Quercia et al. (2011) presented a study trying to predict the personality of Twitter users from their social features like: number of followers, number of followees, listed count,[3] etc. They were able to find strong correlations between these features and the Big-Five personality traits. To leverage the insights of this study, we computed the differences in social features between the deleter and non-deleter user sets.

Table 2 shows the median attribute scores for the two user sets. We find that, users in the deleter set have significantly higher follower count, with median follower counts for users in the deleter and non-deleter sets being 508 and 375 respectively. However, the difference between the distributions of followee count is not that prominent; we observe that median followee count for users in the deleter and non-deleter sets are 403 and 394 respectively. Interestingly, we find that the median tweeting rate of users in the deleter set is nearly twice the tweeting rate of users in the non-deleter set (8.85

[3] Lists are an organizational feature of Twitter, using which users can create and follow a named group of users separately. Listed count of a Twitter user indicates the number of Twitter lists the user is a member of. https://support.twitter.com/articles/76460

Deleter Trait		Predicting Features
Openness	(+4)	Quantifiers, Humans, Causation, Certainty
	(-3)	Biological Processes, Body, Work
Conscientiousness	(+1)	2nd Person Pronouns
	(-9)	Auxiliary Verbs, Future Tense, Negations, Negative Emotions, Sadness, Cognitive Mechanisms, Discrepancy, Feeling, Work
Extraversion	(+2)	Social Process, Family
	(-1)	Health
Agreeableness	(+3)	2nd Person Pronouns, Ingestion, Achievement
	(-2)	Causation, Money
Neuroticism	**(+3)**	Hearing, Feeling, Religion

Table 3: Personality traits for users in the deleter set, as predicted by relative use of words from different LIWC categories. For example, higher openness for users in deleter set is predicted by 4 features, while lower openness is predicted by 3 features.

tweets/day compared to 4.74 tweets/day). The median listed count for users in both the deleter and non-deleter sets were found to be 2.

When viewing the significant differences presented above and in light of the strong and significant correlations presented by Quercia et al. (2011), we find that these differences suggest, that *users in the deleter set are more likely to be extroverted and neurotic* compared to users in the non-deleter set.

Differences in linguistic style

Numerous authors have successfully shown correlations between a person's writing style and their personality. Golbeck et al. (2011) used the LIWC 2007 toolkit[4] to show that similar correlations exist between linguistic style of Twitter users and their Big-Five personality traits.

To utilize the insights presented by Golbeck et al. (2011), we computed for both the deleter and non-deleter sets, the median percentage of words belonging to the different LIWC categories. Using the significant differences thus obtained, we computed the predicted personality trait for users in the deleter set, compared to users in the non-deleter set, for every LIWC feature for which strong correlations were presented in (Golbeck et al. 2011). Table 3 shows the predicted personality trait for users in the deleter set, for the different LIWC categories.

For example we find that, higher openness for users in deleter set is predicted by 4 features, while lower openness is predicted by 3 features. The above results indicate that *users in the deleter set are likely to be less conscientious* as predicted by the nine features, while the converse is predicted by only one feature. Also, *users in the deleter set are likely to be more neurotic* as predicted by the three features, while the converse is predicted by none. However, the predictions for the other three personality traits are not clear from the results in Table 3 as they contain mixed signals.

[4] http://www.liwc.net

	Non-Del Tweets	Del Tweets	Difference
Tweets w/ hashtags	11.85%	5.89%	-50.29%
Tweets w/ urls	17.32%	8.43%	-51.32%
Tweets w/ mentions	41.10%	29.63%	-27.90%
Replies	30.37%	22.84%	-24.79%

Table 4: Percentage of tweets in the non-deleted and deleted tweet sets that contain hashtags, urls, and mentions along with the percentage of tweets that are replies.

Characterizing tweets differences

Here, we try to answer the question: does there exist any characteristic differences between tweets that are deleted and those that are not. To answer this question, we compared 1.2 million deleted and 15.9 million non-deleted tweets *posted by users in the deleter set*, across different dimensions.

Comparing tweet attributes

Ghosh et al. (2013) had noted that tweets containing hashtags and urls are generally considered to be more informative, as hashtags put the tweet in context of a bigger discussion, while urls provide references to additional sources. To understand the differences in information content, we studied the differences between the deleted and non-deleted tweets along those dimensions.

Table 4 shows the percentage of deleted and non-deleted tweets which are replies, as well as the percentage of tweets in both sets that contain mentions, hashtags, and urls. We find that, when compared to non-deleted tweets, the percentage of deleted tweets that contain hashtags and urls is nearly half (50.29% drop for hashtags and 51.32% drop for urls). This indicates that overall, *information content of tweets in the deleted set is significantly lower* when compared to tweets that are not deleted.

One may try to explain this lack of informative content in deleted tweets by postulating that, deleted tweets are more conversational in nature. However, we find that fewer deleted tweets (24.79% less) are replies and a lesser fraction of them (27.9% less) contain mentions. This lack of conversational markers in deleted tweets indicate that overall *conversational tweets are less likely to be deleted*.

Comparing linguistic features

We also compared deleted and non-deleted tweets using the LIWC toolkit. Table 5 shows the difference in tweet vocabulary usage corresponding to different LIWC categories, between deleted tweets and non-deleted tweets.

We find that use of first person singular pronouns and third person pronouns (both singular and plural) increases significantly in deleted tweets — for first person singular pronouns the increase is 5.72%, while the increase for third person singular and plural pronouns are 11.12% and 5.89% respectively. Interestingly, the use of first person plural pronouns show a significant decrease (11.45% drop). This may indicate that, subjects of deleted tweets are more likely to be

	Non-Del Tweets	Del Tweets	Difference
Pronouns			
1st Person (singular)	6.52%	6.90%	5.72%
1st Person (plural)	0.48%	0.43%	-11.45%
2nd Person	2.41%	2.45%	1.60%
3rd Person (singular)	0.86%	0.95%	11.12%
3rd Person (plural)	0.45%	0.48%	5.89%
Impersonal	4.15%	4.42%	6.60%
Tense			
Past	2.29%	2.37%	3.45%
Present	9.95%	10.19%	2.43%
Future	0.87%	0.86%	-1.14%
Emotion			
Positive Emotion	4.45%	3.87%	-13.03%
Negative Emotion	2.71%	3.15%	16.23%
Anxiety	0.27%	0.28%	3.70%
Anger	1.39%	1.72%	23.74%
Sadness	0.48%	0.51%	6.25%
Swear Words			
Swear Words	0.93%	1.21%	30.10%
Sexual References	0.94%	1.01%	7.44%
Cognitive Process			
Insight	1.47%	1.55%	5.44%
Causation	1.22%	1.30%	6.55%
Discrepancy	1.57%	1.65%	5.09%
Tentative	1.78%	1.92%	7.86%
Certainty	1.34%	1.36%	1.49%
Inhibition	0.46%	0.48%	4.34%
Inclusive	2.49%	2.50%	0.40%
Exclusive	2.24%	2.43%	8.48%

Table 5: Percentage of tweet vocabulary consisting of words from different LIWC categories in deleted and non-deleted tweets.

about the author herself or about people with whom the author is unlikely to identify with.

Further, we find significant increase in past and present references in deleted tweets (3.45% increase for past tense and 2.43% increase for present tense), with future references dropping slightly (1.14% decrease). This suggests, that in general, content of deleted tweets are more likely to be associated with past or present events.

Unsurprisingly, we find that a lesser percentage of deleted tweets (13.03% drop) have words related to positive emotions while words with negative emotions increase (16.23% increase) significantly in deleted tweets. Further, significantly more deleted tweets (23.74% more) contain words related to anger, while use of anxiety and sadness related words also show a relative increase. Also, we find that a larger fraction of tweets in the deleted tweet set contain swear words (30.10% increase) and sexual references (7.44% increase).

Interestingly, we note that for all categories of words related to cognitive processes, a larger fraction of deleted tweets contain words related to them. These demonstrate that in general *tweets that are deleted are more carefully constructed* than tweets in the non-deleted set.

Conclusion

This work presented a large scale empirical analysis of deleted tweets and their authors. We developed several insights during our study. In particular, we found that there exists significant differences in personality, between those who delete their tweets (even low numbers) and those who do not. Users who delete their tweets are more likely to be extroverted and neurotic while also being less conscientious. We also found that in deleted tweets, a significantly higher fraction of the vocabulary consists of swear words, and markers that indicate anger, anxiety, and sadness. Interestingly, a significant part of them show signs of being carefully constructed.

An obvious future work is to develop an online system that can utilize the above insights, to nudge users into making better judgments, of whether to post a tweet or not. However, the major challenge in building such a system would be to find a balance, such that while being useful and delivering appropriate nudges, the system refrains from actively nagging its users.

Acknowledgments

This research was supported in part by the Indo-German Max-Planck Center for Computer Science (IMPECS), through their grant for the project "Understanding, Leveraging and Deploying Online Social Networks".

References

Almuhimedi, H.; Wilson, S.; Liu, B.; Sadeh, N.; and Acquisti, A. 2013. Tweets Are Forever: A Large-scale Quantitative Analysis of Deleted Tweets. In *Proc. ACM CSCW*, 897–908.

Bagdouri, M., and Oard, D. W. 2015. On Predicting Deletions of Microblog Posts. In *Proc. ACM CIKM*, 1707–1710.

Costa, P. T., and McCrae, R. R. 1992. *Neo Personality Inventory-Revised (NEO PI-R)*. Psychological Assessment Resources.

Ghosh, S.; Zafar, M. B.; Bhattacharya, P.; Sharma, N.; Ganguly, N.; and Gummadi, K. 2013. On Sampling the Wisdom of Crowds: Random vs. Expert Sampling of the Twitter Stream. In *Proc. CIKM*, 1739–1744.

Golbeck, J.; Robles, C.; Edmondson, M.; and Turner, K. 2011. Predicting Personality from Twitter. In *Proc. IEEE PASSAT/SocialCom*, 149–156.

Petrovic, S.; Osborne, M.; and Lavrenko, V. 2013. I Wish I Didn't Say That! Analyzing and Predicting Deleted Messages in Twitter. `http://arxiv.org/abs/1305.3107`. Accessed On. March 10, 2016.

Quercia, D.; Kosinski, M.; Stillwell, D.; and Crowcroft, J. 2011. Our Twitter Profiles, Our Selves: Predicting Personality with Twitter. In *Proc. IEEE PASSAT/SocialCom*, 180–185.

Sleeper, M.; Cranshaw, J.; Kelley, P. G.; Ur, B.; Acquisti, A.; Cranor, L. F.; and Sadeh, N. 2013. "I Read My Twitter the Next Morning and Was Astonished": A Conversational Perspective on Twitter Regrets. In *Proc. ACM CHI*, 3277–3286.

Zhou, L.; Wang, W.; and Chen, K. 2016. Tweet Properly: Analyzing Deleted Tweets to Understand and Identify Regrettable Ones. In *Proc. WWW*.