

# The Edinburgh Twitter Corpus

Saša Petrović

School of Informatics  
University of Edinburgh  
sasa.petrovic@ed.ac.uk

Miles Osborne

School of Informatics  
University of Edinburgh  
miles@inf.ed.ac.uk

Victor Lavrenko

School of Informatics  
University of Edinburgh  
vlavrenk@inf.ed.ac.uk

## Abstract

We describe the first release of our corpus of 97 million Twitter posts. We believe that this data will prove valuable to researchers working in social media, natural language processing, large-scale data processing, and similar areas.

## 1 Introduction

In the recent years, the microblogging service Twitter has become a popular tool for expressing opinions, broadcasting news, and simply communicating with friends. People often comment on events in real time, with several hundred micro-blogs (*tweets*) posted each second for significant events. Despite this popularity, there still does not exist a publicly available corpus of Twitter posts. In this paper we describe the first such corpus collected over a period of two months using the Twitter streaming API.<sup>1</sup> Our corpus contains 97 million tweets, and takes up 14 GB of disk space uncompressed. The corpus is distributed under a Creative Commons Attribution-NonCommercial-ShareAlike license<sup>2</sup> and can be obtained at <http://demeter.inf.ed.ac.uk/>. Each tweet has the following information:

- timestamp – time (in GMT) when the tweet was written
- anonymized username – the author of the tweet, where the author’s original Twitter username is replaced with an id of type *userABC*. We anonymize the usernames in this way to avoid malicious use of the data (e.g., by spammers). Note that usernames are anonymized consistently, i.e., every time user *A* is mentioned in the stream, he is replaced with the same id.

<sup>1</sup><http://stream.twitter.com/>

<sup>2</sup><http://creativecommons.org/licenses/by-nc-sa/3.0/legalcode>

Table 1: N-gram statistics.

N-grams	tokens	unique
Unigrams	2,263,886,631	31,883,775
Bigrams	2,167,567,986	174,785,693
Trigrams	2,072,595,131	948,850,470
4-grams	1,980,386,036	1,095,417,876

- posting method – method used to publish the tweet (e.g., web, API, some Twitter client). Given that there are dozen of Twitter clients in use today, we believe this information could be very useful in determining, e.g., any differences in content that comes through different clients.

The format of our data is very simple. Each line has the following format:

```
timestamp \t username \t tweet \t client
```

where \t is the tab character, and client is the program used for posting the tweet. Note that the additional whitespaces seen above are only added for readability, and don’t exist in the corpus.

## 2 Corpus statistics

We collected the corpus from a period spanning November 11th 2009 until February 1st 2010. As was already mentioned, the data was collected through Twitter’s streaming API and is thus a representative sample of the entire stream. Table 1 shows the basic n-gram statistics – note that our corpus contains over 2 billion words. We made no attempt to distinguish between English and non-English tweets, as we believe that a multilingual stream might be of use for various machine translation experiments.

Table 2 shows some basic statistics specific to the Twitter stream. In particular, we give the number of users that posted the tweets, the number of links (URLs) in the corpus, the number of topics and the number of replies. From the first two rows of Table 2

Table 2: Twitter-specific statistics.

	Unique	Total
tweets	-	96,369,326
users	9,140,015	-
links	-	20,627,320
topics	1,416,967	12,588,431
replies	5,426,030	54,900,387
clients	33,860	-

Table 3: Most cited Twitter users

Username	number of replies
@justinbieber	279,622
@nickjonas	95,545
@addthis	56,761
@revrunwisdom	51,203
@	50,565
@luansantanaevc	49,106
@donniewahlberg	46,126
@eduardosurita	36,495
@fiuk	33,570
@ddlovato	32,327

we can see that the average number of tweets per user is 10.5. Topics are defined as single word preceded by a # symbol, and replies are single words preceded by a @ symbol. This is the standard way Twitter users add metadata to their posts. For topics and replies, we give both the number of unique tokens and the total number of tokens.

Table 3 shows a list of 10 users which received the most replies. The more replies a user receives, more influential we might consider him. We can see that the two top ranking users are Justin Bieber and Nick Jonas, two teenage pop-stars who apparently have a big fan base on Twitter. In fact, six out of ten users on the list are singers, suggesting that many artists have turned to Twitter as a means of communicating with their fans. Note also that one of the entries is an empty username – this is probably a consequence of mistakes people make when posting a reply.

Similarly to Table 3, Table 4 shows the ten most popular topics in our corpus. We can see that the most popular topics include music (#nowplaying, #mm – music monday), jobs ads, facebook updates (#fb), politics (#tcot – top conservatives on Twitter), and random chatter (#ff – follow friday, #tinychat, #fail, #formspringme). The topic #39;s is an error in interpreting the apostrophe sign, which has the ascii value 39 (decimal).

Table 4: Most popular topics on Twitter

Topic	number of occurrences
#nowplaying	255,715
#ff	220,607
#jobs	181,205
#fb	144,835
#39;s	110,150
#formspringme	85,775
#tcot	77,294
#fail	56,730
#tinychat	56,174
#mm	52,971

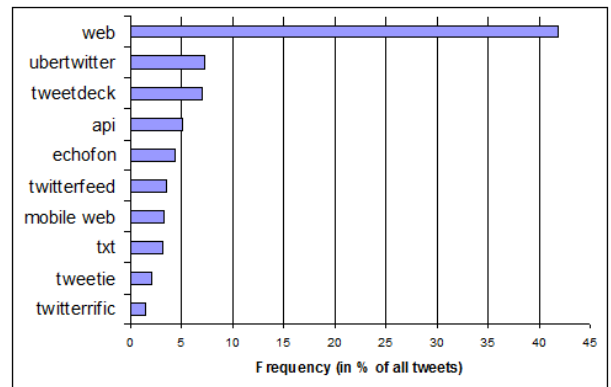


Figure 1: Different sources of tweets.

Figure 1 shows the ten most popular clients used for posting to Twitter. Despite the large amount of different Twitter clients used (over 33 thousand, cf. Table 2), figure 1 shows that almost 80% of all the tweets in our corpus were posted using one of the top ten most popular clients. We can see that traditional posting through the Twitter web site is still by far the most popular method, while UberTwitter and TweetDeck seem to be the next most popular choices.

### 3 Conclusion

In this paper we presented a corpus of almost 100 million tweets which we made available for public use. Basic properties of the corpus are given and a simple analysis of the most popular users and topics revealed that Twitter is in large part used to talk about music by communicating both with artists and other fans. We believe that this corpus could be a very useful resource for researchers dealing with social media, natural language processing, or large-scale data processing.