Location Methodology

There are many ways to obtain geographic information from different social media platforms. We use two different methods to assign location to posts depending on the metadata available: posts that are geotagged (1% of posts) and posts that are not (99% of posts).

Geotagged Posts

Some posts are tagged with exact geographic coordinates when they are posted. This is possible only if users have posted from a mobile phone and have chosen to share their locations. As a result, about 1% of all posts can be located with this level of precision. We call these “Geotagged posts.” You can find these posts when you click on “Geotagged” in the Geography visualization, where you can zoom all the way into the street level!

Non-Geotagged Posts

For the remaining 99% of posts that don’t have the exact latitude-longitude coordinates, we estimate their locations based on various pieces of contextual information, for example, their profile information (*and other attributes listed in the next paragraph*). The estimation uses the large amounts of data at our disposal to identify the attributes of users whose locations are known. These attributes are then used to infer the location of users whose locations are not directly available but display similar attributes of Geotagged posts.

Using Geotagged Post to Infer the Location of Non-Geotagged posts

Based on the 1% of all posts that are geotagged, which is hundreds of millions of posts from every place on earth, we were able to build a statistical guesser. The most useful piece of information is the “location” field in user profiles, which is a free-form input space where users can describe their locations in their own words. In addition, other attributes such as users’ time zones and languages can be used to determine the countries, regions and cities from where they have tweeted.

If a match to a location is found based on user data such as profile location, time zones and post language, our statistical guesser is able to estimate and assign the location of the post based on this similarity. When no match is found, our guesser does not make a prediction; the locations of such posts are labeled as ‘unknown’ and excluded from the Geographic data.

Accuracy

For posts without geotags, in general we are able to match:

- About 90% of all posts to a country of origin
- About 70% of all posts to a specific state or province within that country
- About 50% of all posts to a city or urban area* within that state

*** Cities and Urban Areas**

People living in the same urban area generally share the same characteristics, even when those areas may be divided into different “cities”. In our maps, we always aggregate urban areas when we talk of cities. So for example:

- The area of Washington blends all the urban area that includes: Washington DC, as well as part of Virginia and Maryland
- The area of New York includes: New York City - Jersey City - Newark (USA);
- The area of Taipei includes: Taipei - New Taipei - Zhongli - Zhubei … (Taiwan);
- The area of Tokyo includes: Tokyo - Yokohama - Kawasaki (Japan)

Location Data per Content Source

We do not have location data available for all of our content sources. Below is a list of the current content sources and the date location data became available:

- Twitter: September 2009
  (*Updated location methodology available September 2014*)
- Blogs: May 2015
- Forums: September 2011
- Reddit: Not available
- Google+: Not available
- Tumblr: August 2017
- QQ: January 2018
- Facebook: Not available
- Instagram: September 2014
  (*We no longer receive location metadata on Instagram posts after Dec 11th 2018*)
- VK Keywords: March 2016
- Reviews: Not available
- News: May 2015
- YouTube: Not available