Sanders-Twitter Sentiment Corpus

Niek J. Sanders
njs@sananalytics.com
October 24, 2011

Overview

This corpus is designed for training and testing Twitter sentiment analysis algorithms.

It consists of 5513 hand-classified tweets. These tweets were classified with respect to one of 4 different topics.

Each entry contains:

- Tweet id
- Tweet text
- Tweet creation date
- Topic used for sentiment
- Sentiment label: 'positive', 'neutral', 'negative', or 'irrelevant'

The break down of topics and data:

Topic	# Positive	# Neutral	# Negative	# Irrelevant	Twitter Search Term
Apple	191	581	377	164	@apple
Google	218	604	61	498	#google
Microsoft	93	671	138	513	#microsoft
Twitter	68	647	78	611	#twitter

Note: the @apple handle does not actually belong to Apple Inc. However, it is still extremely well focused on the topic of the Apple company and its products.

These searches were done without using a language specifier. As noted in the "Classifications" section, foreign language tweets were marked 'irrelevant'.

Installation

Because of restrictions in Twitter's Terms of Service, the actual tweets can not be distributed with the sentiment corpus. A small Python script is included to download all of the tweets. **Due to limitations in Twitter's API, the download process takes about 43 hours.**

Just three easy steps:

1. Start the tweet downloader script: `python install.py`
2. Hit enter three times to except the defaults.
3. Wait till the script indicates that it's done.

Note: the script is smart enough to resume where it left off if downloading is interrupted.

The completed corpus will be in `full-corpus.csv`. A copy of all the raw data downloaded from Twitter is kept in `/rawdata`.

Classifications

Four classifications are used in this corpus:

Positive	• Positive indicator on topic
Neutral	• Neither positive nor negative indicators • Mixed positive and negative indicators • On topic, but indicator undeterminable • Simple factual statements • Questions with no strong emotions indicated
Negative	• Negative indicator on topic
Irrelevant	• Not English language • Not on-topic (e.g. spam)

Sentiment assignment is an extremely subjective exercise.

For this corpus, "Positive" and "Negative" labels were reserved for tweets which clearly express an emotion or where the implications were unambiguous. As a rule of thumb, "neutral" was the preferred label for border line cases.

Examples:

> There are huge lines at the @apple store.
> Labeled **neutral**. From a shoppers perspective this could be bad, or it could be a sign of excitement about the product launch. From an investor's perspective this could be good, since it indicates a strong new product launch.
>
> I had to wait for six friggin' hours in line at the @apple store.
> Labeled **negative**. The tweeter is clearly unhappy with the situation and is referring to Apple in the negative sense.

Siri is down.
Labeled **negative**. This is the border line between simple factual statements and ones with stronger negative implications.

@apple can you get your phone to stop vibrating when you receive a text?
Labeled **neutral**. This is a simple question with the user expressing no strong emotions on the matter.

Having major battery drain issue since updating iPhone 4 to iOS 5. Anyone else?
Labeled **negative**. Again it's a simple question, but with negative connotations.

I found out on #google that my scumbag boyfriend cheated on me.
Labeled **neutral**. While the sentiment is clearly negative, it is not negative with respect to the classification topic (Google).

All classification was done by an American male who is fluent in English.

Contact

Niek J. Sanders
njs@sananalytics.com
http://linkedin.com/in/niekjsanders

Sanders Analytics LLC
http://www.sananalytics.com

Licensing

Revisions

October 24, 2011 (v0.2)
Python tweet downloader script made more robust. Fixed error in writing out unicode characters to full corpus csv.

October 19, 2011 (v0.1)
First public release