

This is the sentiment140 dataset. It contains 1,600,000 tweets extracted using the twitter api . The tweets have been annotated (0 = negative, 4 = positive) and they can be used to detect sentiment.

Content

It contains the following 6 fields:

1. **target:** the polarity of the tweet (0 = negative, 4 = positive)
2. **ids:** The id of the tweet (2087)
3. **date:** the date of the tweet (Sat May 16 23:58:44 UTC 2009)
4. **flag:** The query (lyx). If there is no query, then this value is NO_QUERY.
5. **user:** the user that tweeted (robotickilldozr)
6. **text:** the text of the tweet (Lyx is cool)

800000 tweets are classified positive and 800000 are classified negative.

In order to perform machine learning on this dataset, we're only keeping the text.