1 - What Dataset are we using

The Sentiment140 dataset. This dataset contains data from 1,600,000 tweets, classified as either positive (4) or negative (0). It additionally contains some extra information about each tweet. Such information includes the username of the tweeter, the date the tweet was made, and the query used to scrape the tweet for the dataset.

2 - Source of the Dataset

We can download the dataset from kaggle at the following link:
https://www.kaggle.com/datasets/kazanova/sentiment140

3 - Size, Preprocessing Information, and the Format of the Dataset

3.1 - Format

The dataset is in CSV format with the following column labels: ‘target’,’id’,’date’,’flag’,’user’,’text’

target	The sentiment classification target; An integer value of either 0(Negative), 2(Neutral), or 4(Positive)
id	An integer identifier value that uniquely identifies every tweet
date	The date the tweet was posted (i.e.‘Mon Apr 06 22:19:45 PDT 2009’)
flag	The query of the tweet
user	The username of the user who posted the tweet
text	The text content of the tweet

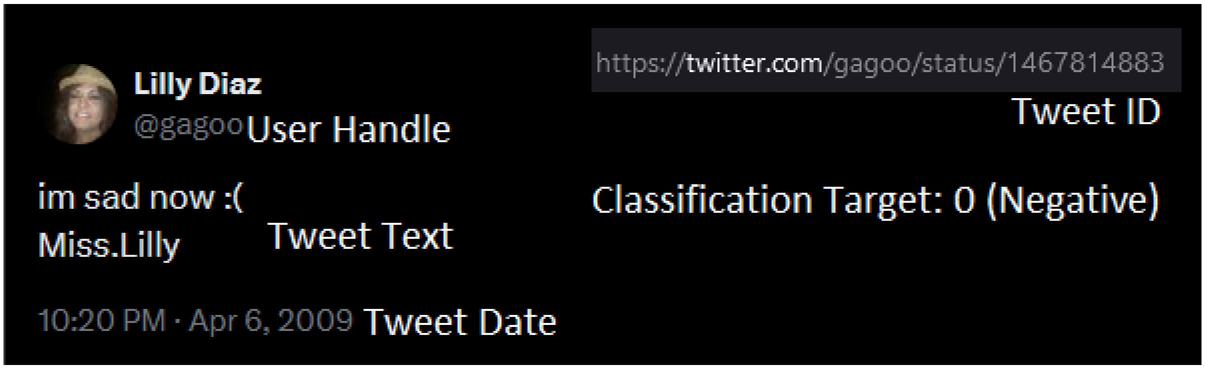
It is worth noting that even though the column ‘target’ has 3 possible values, the value 2 never appears in the dataset. It is additionally worth noting that the dataset appears to be made by another classification model, so some of the training data we use may have the wrong sentiment associated with it to begin with.

3.2 - Size

1,600,000 rows of data, each corresponding to a single tweet. The data contains an even split between targets 0 and 4, representing negative and positive, respectively.

3.2 - Preprocessing Information

The text of every tweet has emoticons removed. Here is an example tweet from the dataset where we can see each aspect of it. This particular tweet makes use of the :(emoticon which has been removed for its inclusion in the dataset.



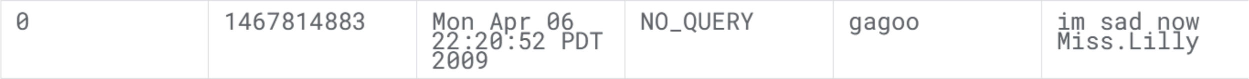
Lilly Diaz
@gagoo

im sad now :(
Miss.Lilly

10:20 PM · Apr 6, 2009

https://twitter.com/gagoo/status/1467814883

Here is the same tweet in the dataset where you can see that the emoticon used in it has been removed from the text field for the tweet.



0	1467814883	Mon Apr 06 22:20:52 PDT 2009	NO_QUERY	gagoo	im sad now Miss.Lilly

4 - References

[1] “Sentiment140 | TensorFlow Datasets.” TensorFlow https://www.tensorflow.org/datasets/catalog/sentiment140

[2] Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." CS224N project report, Stanford 1.12 (2009): 2009. https://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf