Image Sentiment Analysis

# Introduction

Social media has proven to be a rich field for making sense of popular opinion on a number of different topics. From companies trying to understand how their brand is being received, to political actors striving to get a reading on popular opinion, Twitter mining has become an invaluable tool.

Much of this work has focused on text-based sentiment analysis. As a result, text-based sentiment analysis in social media is a fairly well-evolved area of machine learning.

But focusing exclusively on textual sentiment analysis misses the fact that social media is increasingly image-based. With image-centric platforms like Instagram and Snapchat becoming increasingly important, it is foolish to assume that text analysis alone can provide an accurate indicator of social media sentiment on any particular topic.

Unfortunately, much of the work in image analysis has focused on object identification, rather than sentiment determination. And much of the work that has been done on image sentiment has taken a hand-crafted, rather than machine-learning approach to sentiment ascription. Additionally, image sentiment analysis has been hampered by the fact that it is difficult to assemble sufficiently robust data image data sets that are labeled with reliable sentiments.

In this project, I will attempt to determine whether it is possible to build an effective predictor of image sentiment in by using text-based sentiment analysis can provide a ground-truth to effectively develop an image classifier. To do this, I propose to first classify Twitter-posted images based on textual clues. Using these ratings, I will then develop a Neural-Network based model to classify image sentiment.

The predictive ability of this model will be tested against images that have been had sentiment scores ascribed by crowd-sourcing. The model will be compared against benchmarks, and against a model derived solely from images whose sentiment has been crowd-sourced.

# Literature Review

There are a couple of different categories of literature that need to be reviewed to ensure this project is properly grounded in best practices. These are set out systematically below.

## Image Sentiment Analysis

Foo

## Text Sentiment Analysis

Foo

# Dataset

Give the description of the dataset that you are using along with the individual attributes you will or will not use in your analysis. Also mention the source of the dataset (where did you get it from). In case the data is curated and created by you please explain the details. Descriptive statistics of the attributes and datasets can also be provided here.

The data used for the bulk of this analysis is derived from captured Twitter stream data. I downloaded approximately 500,000 tweets, and subsetted these to English language tweets that contain an image and that were not classified as retweets (i.e., the tweet appeared to be an original post).

The text of the downloaded tweet is used to determine overall tweet sentiment. The image URL is used to build the image classifier.

The final image classifier will be tested against raw Twitter data, and more importantly against a set of images that have have sentiment ascribed to them by human scorers in a crowd-sourced methodology. This data set is publicly available from the firm Crowdflower (need URL).

# Approach

Create a block diagram for the steps of your approach to clearly provide an overview. For example, if you first scrapped twitter, second applied NLP techniques to extract keywords, third labelled the tweets as positive and negative using a set of keywords, and fourth build a classifier, then you should create a box for each of the steps with arrows connecting one step to the next one. A sample block diagram is shown below.

Once this is done, explain each of the steps in detail. What are you planning to do in each step or have already done. For example, in the above case you would create subheadings for each of the steps.

1. Pull Twitter Stream Data
2. Subset to English language, non-retweets that contain an image
3. Pull responses to tweets
4. Assess tweet sentiment
5. Normalize pictures
6. Train CNN on pictures
7. Assess CNN results on:
   1. New set of tweets
   2. Images with crowd-sourced sentiment
8. Build CNN on crowd-sourced sentiment and compare to self & to

## Step 1: Pull Raw Twitter Data

This is a fairly straight-forward pull of streaming Twitter data using the GET STATUSES/SAMPLE command from Twitter’s public API. Because of the amount of data needed, this was done in a number of samples over a 2-week period, creating a number of different files

Code used for this step is posted at <https://github.com/asterix135/CKME136/blob/master/Python_code/twitter_stream.py>. Please note that unlike the rest of the code in this project, this is written for Python2, and is designed to be called from the command line. The introductory docstring provides details.

A sample raw pull is posted at <https://github.com/asterix135/CKME136/blob/master/Data/output_jan24.txt>

## Step 2: Subset to Original Tweets with Images

In order to effectively ascribe sentiment, we first need to ensure three things at a minimum:

1. That the tweet is in English
2. That the tweet contains an image
3. That the tweet contains text

Additionally, it is useful to try and ensure that the tweet is original, so as to avoid inadvertently including multiple retweets of the same image and text.

To do this, the various files of pulled stream data were processed on the following basis:

1. the lang attribute for the tweet’s json representation = “en”
2. tweet’s json representation contains a “text” attribute
3. tweet’s json representation contains an “extended\_entities” attribute (indicative of the presence of an image)
4. tweet’s text does not begin with “RT” (standard Twitter syntax for identifying a retweet)

Write details of the step 2. If there is any source code that you’d like to share then provide the link of the Github.

………………………………………………………………………………………………………………………..

………………………………………………………………………………………………………………………..

## Step N: <Name of the step>

Write details of the step N. If there is any source code that you’d like to share then provide the link of the Github.

## Bibliography