COMP 576 Final Project Proposal

Twitter Sentiment Analysis

Wenxing Qiu (wq4) & Yuhong Cheng (yc96)

Introduction

Twitter is a popular microblog social media platform where users post short texts (tweets) to express their opinions about different topics. It is interesting to collect people's tweets and conduct sentiment analysis to find their opinions about products on the market, political topics, movies, and anything that you might be interested in.

Goal

Analyze people's emotion (positive, negative, or neutral) about any topic happening in the world based on their Twitter posts. The topics is decided by the user input.

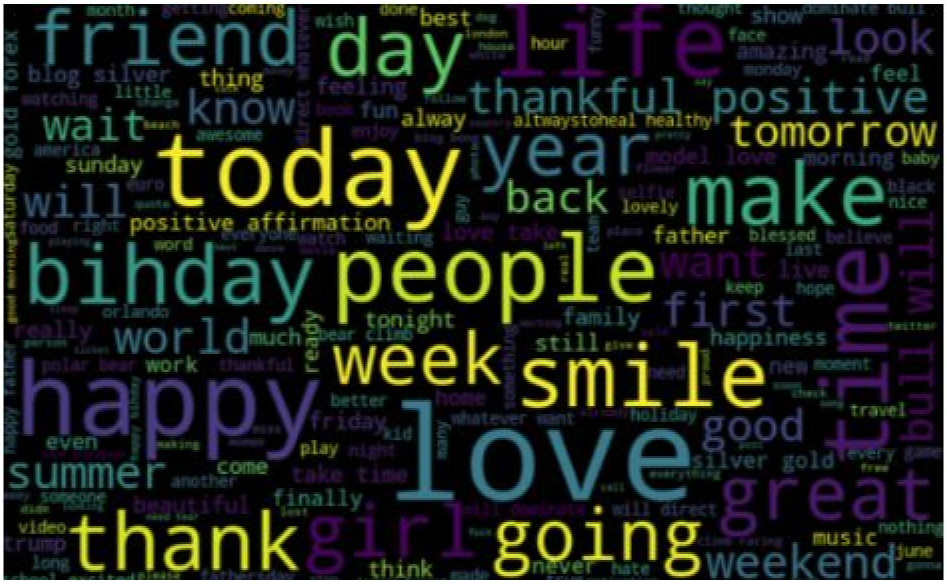
Dataset

We decided to use Sentiment140 dataset as the training dataset. Sentiment140 is a dataset used for sentiment analysis of a brand, product, or topic on Twitter (http://help.sentiment140.com/for-students). The data is in a CSV file with emoticons removed and it has 6 fields, which are:
0 - the polarity of the tweet (0 = negative, 2 = neutral, 4 = positive)
1 - the id of the tweet (2087)
2 - the date of the tweet (Sat May 16 23:58:44 UTC 2009)
3 - the query (lyx). If there is no query, then this value is NO_QUERY.
4 - the user that tweeted (robotickilldozr)
5 - the text of the tweet (Lyx is cool)

Approach

1. Data preprocessing
   We will need to extract tweets data first and we decided to use Tweepy, which is a powerful Python library that provides access to the entire Twitter RESTful API. Then we will need to do data cleaning and tokenization.
2. Visualize testing dataset

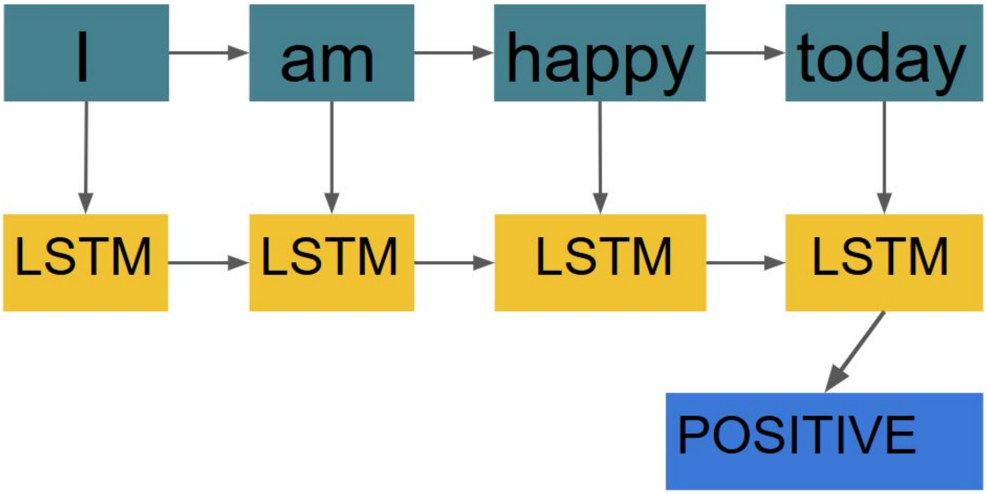
We are planning to use WordCloud to visualize the testing dataset. A WordCloud is a visualization where the most frequent words appear in large size and the less frequent words appear in smaller sizes. Below is an example of positive WordCloud visualization.



3. Build the supervised model for classification
   We want to use two ways to deal with this problem, which are LSTM and bidirectional-LSTM.

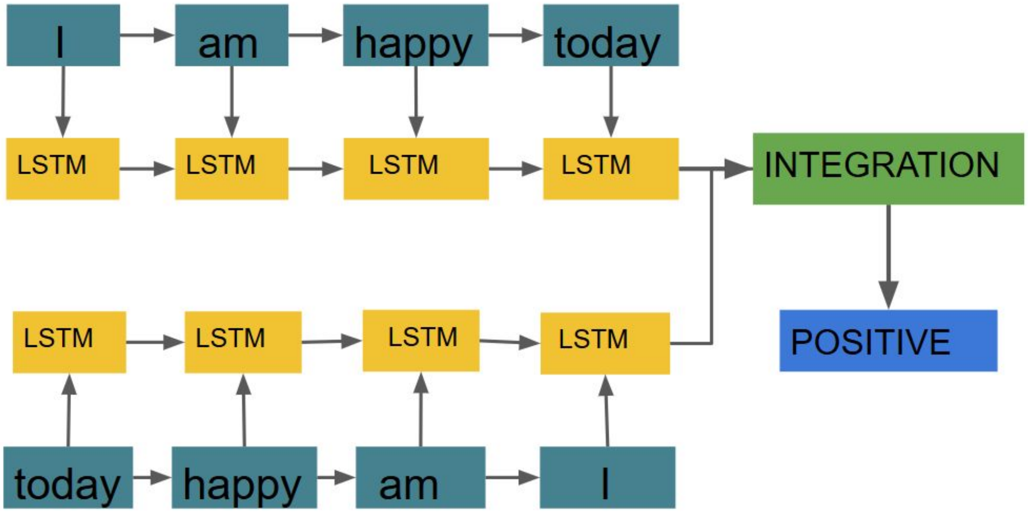
LSTM (long-short term memory)

Here is an example. One twittered:’I am happy today’. After participle, we will get several words: ‘I’ ‘am’, happy, ‘today’. We will apply LSTM in our project and the output we expect is ‘positive’. The model is shown below. LSTM will store all the information from all the layers and then come to a conclusion.



However when using LSTM to do the sentiment analytics, the network will learn the representations from the left to right and the order of the sentence will definitely influence our learning result, so we will also apply Bidirectional-LSTM to predict the sentiment.

Bidirectional-LSTM



This model has two networks, one access information in forward direction and another access in the reverse direction (as shown in figure above). These networks will access to the past and also the future information and then generate the output with integrating both result of the past and future context.

4. Generate results with custom input topic
   Run the user input topic-related testing dataset through the deep neural network and generate the sentiments about that topic.
5. Visualize the results
   We are considering using a Pie Chart to visualize the sentiment analysis results.

Feasibility

The two models we will use here are already well developed, and we can use Keras to implement them. However, there are still some challenges. During writing the proposal, we found that the accuracy of some models are around 50%, so how to modify the parameters to get better accuracy is still a big challenges. Also, if we have enough time, we will try to test the model using real time data from twitter. If we have some new ideas to improve the model during our study, we will update them.

References :

1. Understanding LSTM Networks. (n.d.). Retrieved November 6, 2018, from http://colah.github.io/posts/2015-08-Understanding-LSTMs/
2. Zhang L, Wang S, Liu B. Deep learning for sentiment analysis: A survey[J]. Wiley Interdisciplinary Reviews Data Mining & Knowledge Discovery, 2018.
3. Step-by-Step Twitter Sentiment Analysis: Visualizing United Airlines' PR Crisis. (2018, June 22). Retrieved from https://ipullrank.com/step-step-twitter-sentiment-analysis-visualizing-united-airlines-pr-crisis/
4. Ferro, R. (2018, April 07). Sentiment analysis on Trump's tweets using Python. Retrieved from https://dev.to/rodolfoferro/sentiment-analysis-on-trumpss-tweets-using-python-