Web app for tweet emotion visualization on a map

http://twitter-emotion-visualizer.herokuapp.com*
Gaurav Ahuja
Uni: ga2371
ga2371@columbia.edu

Decmber 3, 2013

Abstract

Twitter is a micro blogging service where users publish and share their feelings in the form of tweets. Tweets are short and informal pieces of text. Twitter also provides geographical information from where the tweet originated. User generated content on Twitter (produced at an enormous rate of 340 million tweets per day) provides a rich source for gleaning people's emotions, interests and opinions, which is necessary for deeper understanding of people's behaviors and actions. Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan and Amit P. Sheth at Kno.e.sis Center, Wright State University [1] have studied ways to identify seven emotions(joy, sadness, anger, love, fear, thankfulness and surprise) from tweets and provide their dataset of approximately 2.5 million emotion labeled tweets. In this project I will use this dataset to train a classifier to predict emotions in tweets and create a web application that will let users query topics and extract emotions from popular tweets related to that topic. Using the geographical information provided by Twitter, emotion of these popular tweets will be visualized on a map.

1 Introduction

Twitter has become immensely popular because people like to express themselves on the fly. Twitter lets people express their emotions, opinions and information about current happenings in 140 characters.

With a large active user base, Twitter can become an important tool in analyzing sentiment of people towards a product, their emotional state etc. Twitter data is dynamic in nature. There have been instances where one can learn about a new happening before print or digital media broadcasts it. With this data companies have the opportunity to examine

*In order to access the app you will have to authorize the app using your twitter account

[1]Wang, Lu Chen, Krishnaprasad Thirunarayan, Amit P. Sheth. Harnessing Twitter Big Data for Automatic Emotion Identification. IEEE fourth conference on social computing, 2012 http://knoesis.org/library/resource.php?id=1749

what customers are saying about their products and services, stock market experts can analyze the market emotion related to a particular market event. It can also be used to analyze prevailing emotional state of people.

Tweets are short, vague and informal. No robust Natural Language Processing technique has been developed to describe the features of a tweet. Hence classifying emotions tweets using machine learning algorithms is a challenge. This projects aims at creating classifier for emotions in tweets and building a web application which allows users to query tweets by topics and display emotions in trending Tweets on a map.

2 Data and Tools

2.1 Annotated corpus

In order to build build a classifier using machine learning techniques an annotated corpus is required. In some sense Tweets are self annotated. Users specify there emotion using 'hashtags' in there Tweets, eg: #excited.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan and Amit P. Sheth at Kno.e.sis Center, Wright State University have studied ways to identify seven emotions(joy, sadness, anger, love, fear, thankfulness and surprise) from tweets and created a data set of about 2.5 million tweets by filtering emotion relevant tweets and annotating them automatically. This data set can be downloaded from `http://knoesis.org/projects/emotion`. The corpus classifies tweets into seven major emotion classes, joy, sadness, fear, love, anger, thankfulness and surprise.

Since Twitter regulations forbid sharing of Twitter data, the authors only provide Tweet ID and the emotion in the tweet. One of the major part of the project is to crawl Twitter and using its API interface download as many Tweets in the data set. This is explained the section on System Description.

2.2 Classification Tool

The classifier model used in this project is based on a Maximum entropy classifier also known as Log Liner model [2]. Maximum entropy models are very popular, especially in natural language processing. Maximum entropy classifiers are commonly used as alternatives to naive Bayes classifiers because they do not assume statistical independence of the features. However, learning in such a model is slower than for a naive Bayes classifier, and requires one to solve a complicated optimization problem. Hence a popular and robust package, Mega Model Optimization Package was used to get the parameters of the model from training data. The package can be downloaded from `http://www.umiacs.umd.edu/~hal/megam/`.

3 System Description

The system can be majorly divided into following sub division:

1. Corpus Down-loader and Pre-processor
2. Feature Extraction and Classification

[2]Note on Log Linear Model `http://www.cs.columbia.edu/~cs4705/notes/loglinear.pdf`

3. Web Application frontend and backend

This section will assume a working knowledge of web application development and basic understanding of Twitter API[3].

3.1 Corpus Down-loader and Pre-processor

Wenbo Wang et. al. provide a corpus of 2.5 million annotated tweets. Since only Tweet ids were provided in the data set, it was essential to download Tweets using the Tweet Ids.

Twitter provides an interface where in a Twitter application can be authorized by a Twitter user and can download Tweets on behalf of that user. Twitter application is nothing but set of application specific keys and user specific keys used along with Twitter API to interact with Twitter.

Unfortunately Twitter limits the number of Tweets any user can download to 180 Tweets per 15 minutes. Hence a `twitter-dataset.herokuapp.com/` was created and registered as a Twitter application. Twitter users willing to assist in the project registered and authorized this application to download the tweets in the corpus.

The code submitted along with this report consists of a package **corpusHandler** that provides a class **tweetDownloader** that accepts the application keys and downloads Tweets from the corpus. Using this method and help of 5 twitter users who authorized the app, I was able to download about 0.47 million tweets over 6 days.

The same package also contains a class **tweetPreprocessor** that pre-processes the downloaded tweets. This involves removing unwanted twitter data and extracting the text. Also this text is processed to handle 'hashtags', punctuation's etc.

3.2 Feature Extraction and Classification

The classifier used is a Maximum entropy classifier which allow a very rich set of features to be used in a model, arguably much richer representations than the Naive Bayes. The model also allows features to have weights. Hence one can use tf-idf, word length, position as weights. In the current implementation I have used unigram features and selectively weighed unigrams based on its position. Since people tend to express emotions at the end of any message, unigrams occuring in the latter half of the tweet are weighed twice as much as unigrams in the first half.

The code submitted along with this report provides a package **tweetHandler** which contains a class **tweetFeatureExtractor** which extracts features and formats them in a way which is consistent with the MEGA M package used to find the parameters of the model.

Classifier Evaluation

The 0.47 million tweets were divided into 3 parts. Training data formed 75%, development data 10% and test data 15%. The parameters of the model were obtained by running the MEGA M optimization package on the training and test data. Using these parameters the following results for obtained for each of the emotion classes:

[3]Twitter Developer Documentations `http://dev.twitter.com`

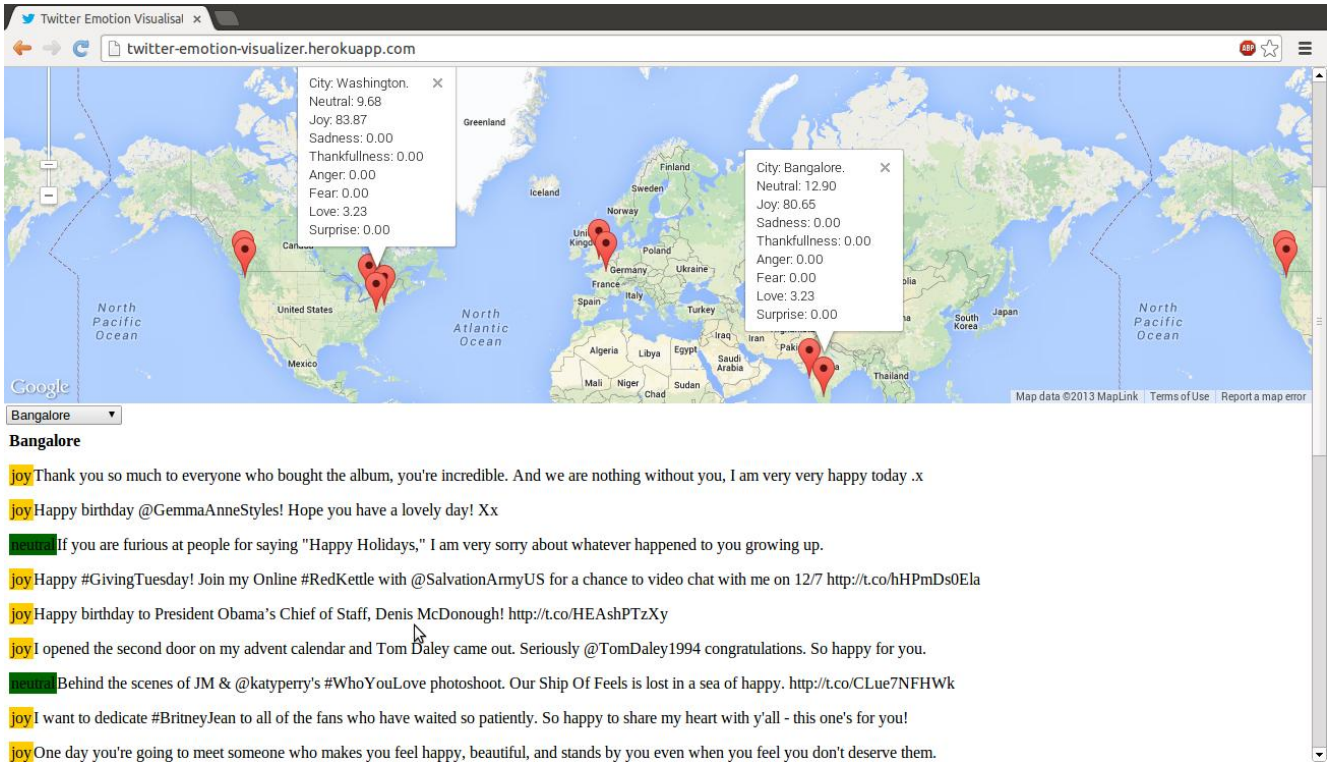
Emotion	Precision	Recall	F-measure
joy	0.96	0.96	0.96
sadness	1.00	0.92	0.96
anger	0.99	0.96	0.97
love	0.98	0.94	0.96
fear	1.00	0.78	0.88
thankfulness	1.00	0.92	0.96
surprise	1.00	0.00	0.00

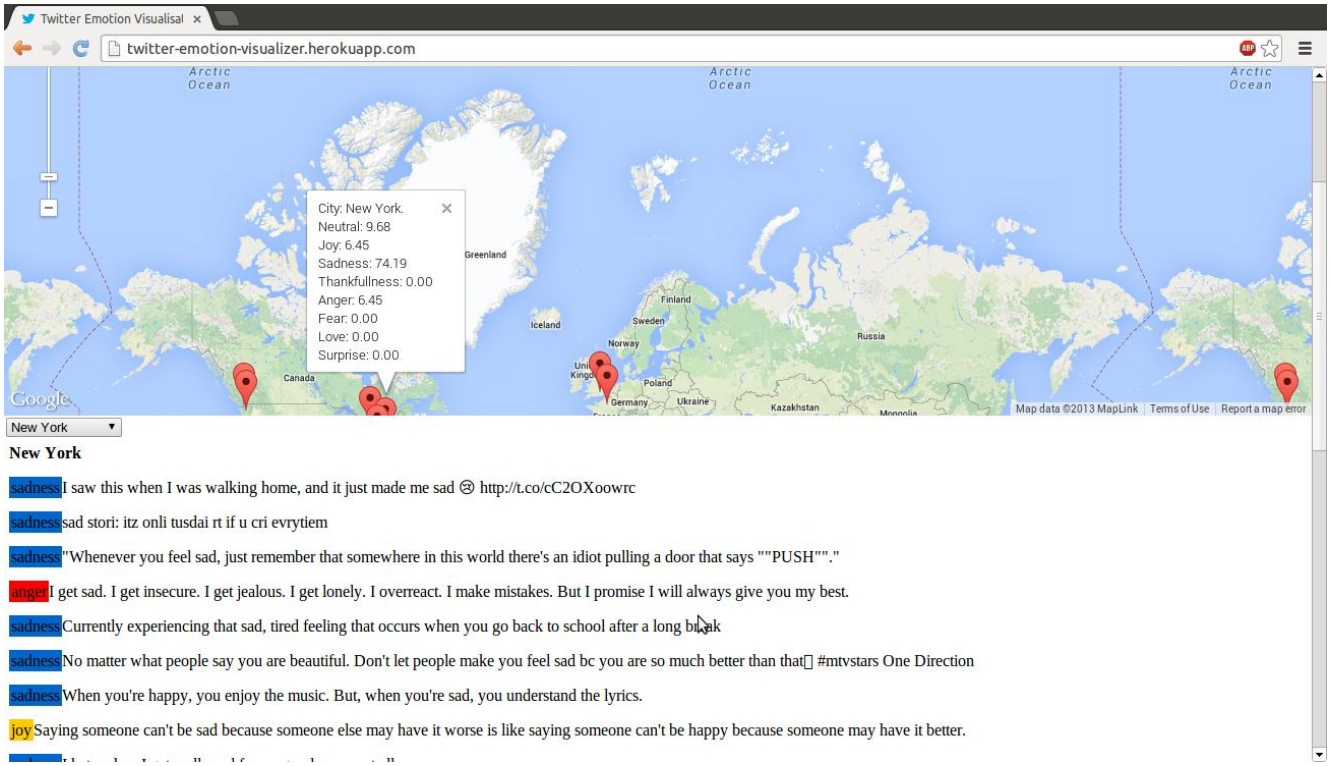
3.3 Web Application

Once the classifier gave good performance of the Test data, the parameters from the model were used to classify Tweets in realtime. Hence a web application `http://twitter-emotion-visualizer.herokuapp.com` was created which lets a user to query a particular topic. Using the Twitter API tweets relevant to that topic are extracted from major cities around the world. These tweets are then classified and emotion in the tweets are presented on the map. In order to use this application the user will have to authorize this application by using their twitter accounts.

4 Results and Screen-shots

Visit `http://twitter-emotion-visualizer.herokuapp.com`

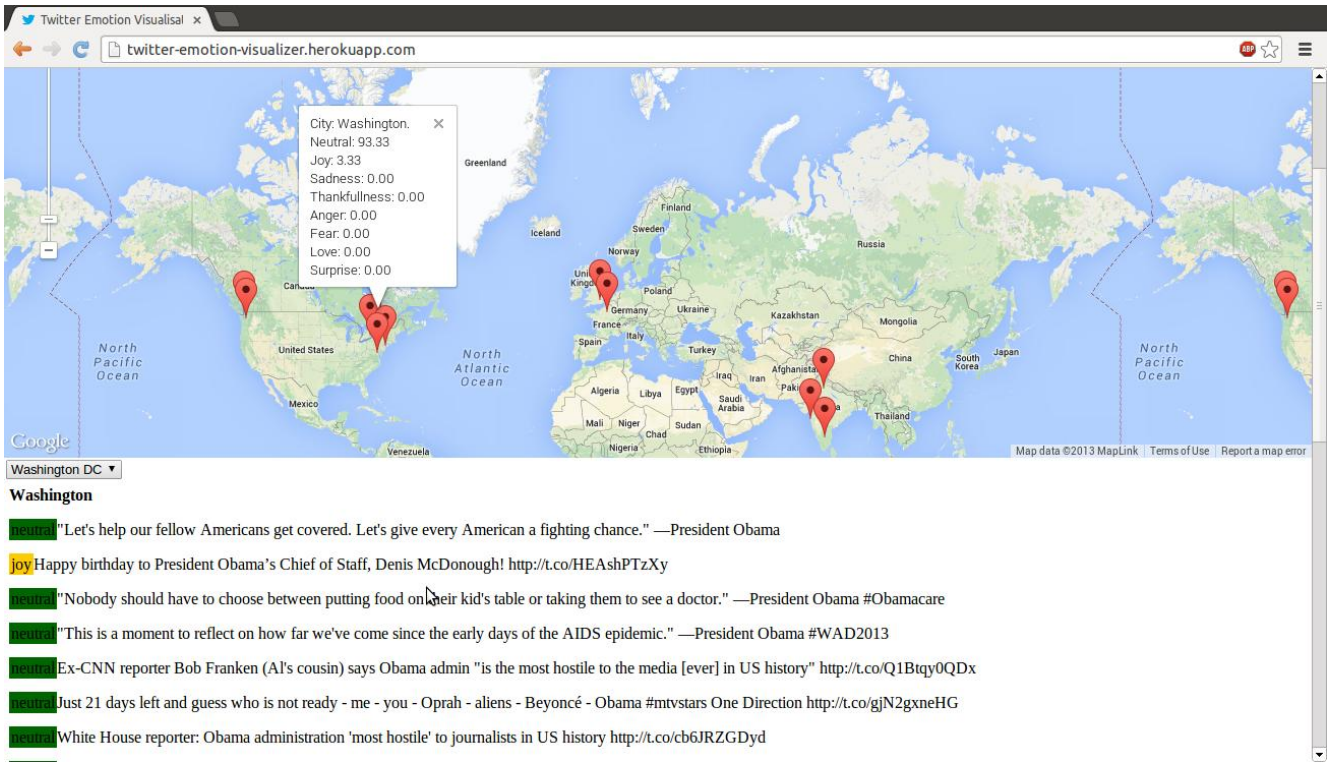


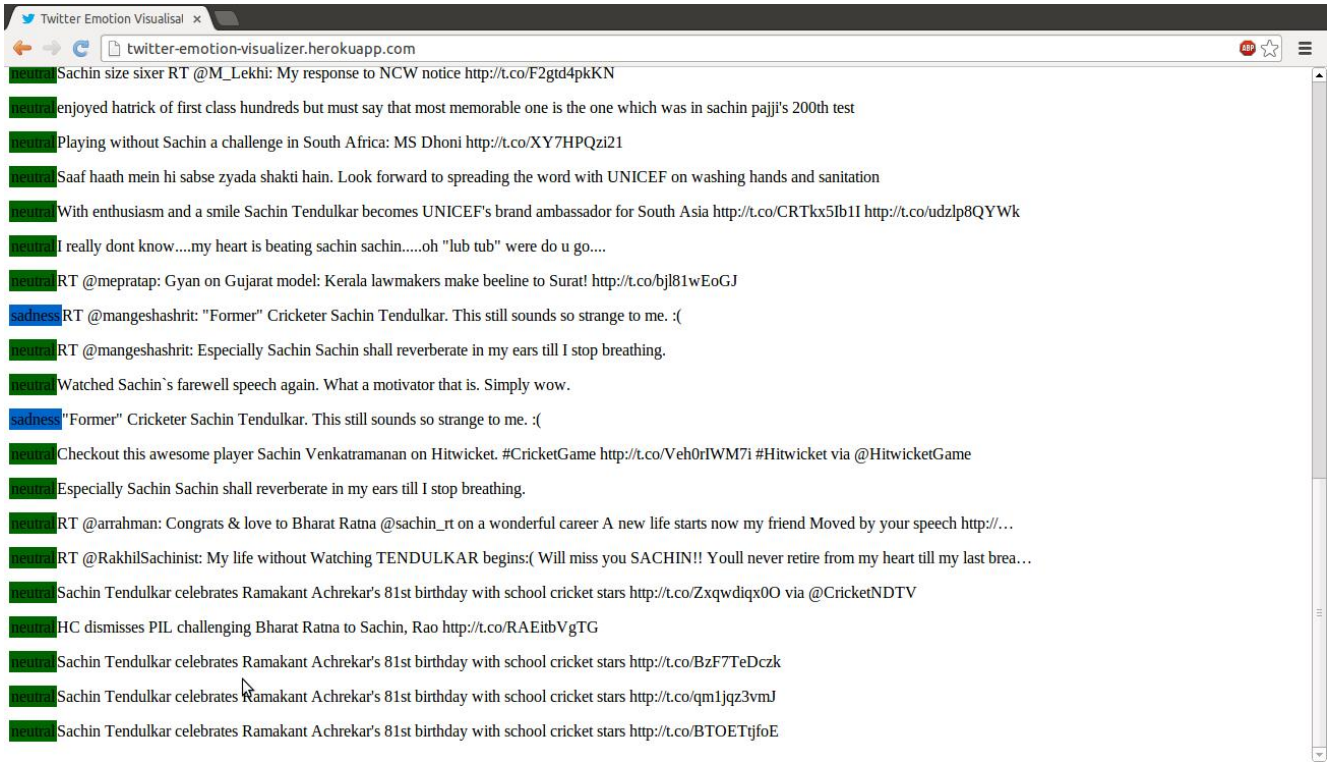


5 Further Improvements

Tweets are short, vague and informal. Contextual information is mostly not present. Also style of writing tweets varies across countries and cities. Hence following major areas of focus are suggested:

1. Build classifier parameters for each country
2. Since a word stemmer is not available for the kind of language used in twitter, more focus should be given to the unigram itself.
3. Since most tweets can be annotated automatically, an active learning algorithm should be used to counter the dynamic nature of the tweet language


Twitter Emotion Visualisal
twitter-emotion-visualizer.herokuapp.com
City: Washington.
Neutral: 93.33
Joy: 3.33
Sadness: 0.00
Thankfullness: 0.00
Anger: 0.00
Fear: 0.00
Love: 0.00
Surprise: 0.00
Greenland
Iceland
Finland
Sweden
Norway
Russia
Poland
Germany
Ukraine
France
Spain
Italy
Kazakhstan
Mongolia
China
South Korea
Japan
Turkey
Iraq
Iran
Afghanistan
Algeria
Libya
Egypt
Saudi Arabia
Thailand
Mali
Niger
Chad
Sudan
Nigeria
Ethiopia
Mexico
Venezuela
United States
North Pacific Ocean
North Atlantic Ocean
North Pacific Ocean
Google
Map data ©2013 MapLink
Terms of Use
Report a map error
Washington DC
Washington
neutral "Let's help our fellow Americans get covered. Let's give every American a fighting chance." —President Obama
joy Happy birthday to President Obama's Chief of Staff, Denis McDonough! http://t.co/HEAshPTzXy
neutral "Nobody should have to choose between putting food on their kid's table or taking them to see a doctor." —President Obama #Obamacare
neutral "This is a moment to reflect on how far we've come since the early days of the AIDS epidemic." —President Obama #WAD2013
neutral Ex-CNN reporter Bob Franken (Al's cousin) says Obama admin "is the most hostile to the media [ever] in US history" http://t.co/Q1Btqy0QDx
neutral Just 21 days left and guess who is not ready - me - you - Oprah - aliens - Beyoncé - Obama #mtvstars One Direction http://t.co/gjN2gxneHG
neutral White House reporter: Obama administration 'most hostile' to journalists in US history http://t.co/cb6JRZGDyd


Twitter Emotion Visualisal
twitter-emotion-visualizer.herokuapp.com
neutral Sachin size sixer RT @M_Lekhi: My response to NCW notice http://t.co/F2gtd4pkKN
neutral enjoyed hatrick of first class hundreds but must say that most memorable one is the one which was in sachin pajji's 200th test
neutral Playing without Sachin a challenge in South Africa: MS Dhoni http://t.co/XY7HPQzi21
neutral Saaf haath mein hi sabse zyada shakti hain. Look forward to spreading the word with UNICEF on washing hands and sanitation
neutral With enthusiasm and a smile Sachin Tendulkar becomes UNICEF's brand ambassador for South Asia http://t.co/CRTkx5Ib1I http://t.co/udzlp8QYWk
neutral I really dont know....my heart is beating sachin sachin.....oh "lub tub" were do u go....
neutral RT @mepratap: Gyan on Gujarat model: Kerala lawmakers make beeline to Surat! http://t.co/bjl81wEoGJ
sadness RT @mangeshashrit: "Former" Cricketer Sachin Tendulkar. This still sounds so strange to me. :(
neutral RT @mangeshashrit: Especially Sachin Sachin shall reverberate in my ears till I stop breathing.
neutral Watched Sachin`s farewell speech again. What a motivator that is. Simply wow.
sadness "Former" Cricketer Sachin Tendulkar. This still sounds so strange to me. :(
neutral Checkout this awesome player Sachin Venkatramanan on Hitwicket. #CricketGame http://t.co/Veh0rIWM7i #Hitwicket via @HitwicketGame
neutral Especially Sachin Sachin shall reverberate in my ears till I stop breathing.
neutral RT @arrahman: Congrats & love to Bharat Ratna @sachin_rt on a wonderful career A new life starts now my friend Moved by your speech http://…
neutral RT @RakhilSachinist: My life without Watching TENDULKAR begins:(Will miss you SACHIN!! Youll never retire from my heart till my last brea…
neutral Sachin Tendulkar celebrates Ramakant Achrekar's 81st birthday with school cricket stars http://t.co/Zxqwdiqx0O via @CricketNDTV
neutral HC dismisses PIL challenging Bharat Ratna to Sachin, Rao http://t.co/RAEitbVgTG
neutral Sachin Tendulkar celebrates Ramakant Achrekar's 81st birthday with school cricket stars http://t.co/BzF7TeDczk
neutral Sachin Tendulkar celebrates Ramakant Achrekar's 81st birthday with school cricket stars http://t.co/qm1jqz3vmJ
neutral Sachin Tendulkar celebrates Ramakant Achrekar's 81st birthday with school cricket stars http://t.co/BTOETtjfoE