Mood Meter

By:

George Mathew

CWID: A20352131

Vidhya

CWID: A20356005

Introduction

In this project we were trying to construct a classifier that classifies tweets into happy and sad and thus predict the mood of a tweet. Our major hypothesis was that the general mood of a region is reflected in the tweets made in that region. On basis of this hypothesis we believed we could guess the general mood of a region by analysing the moods of the tweets made in the region at any given point of time.

"**MONDAY morning** found **Tom Sawyer** miserable"

-Adventures of Tom Sawyer

By Charles Dickens

Everybody hates Mondays and everybody loves Friday evenings. So we concluded that if our hypothesis is true Mondays will have the lowest percentage of happy tweets and Friday evenings will have the highest percentage of happy tweets. So we decided to test our hypothesis by observing the tweets made in USA throughout a week and see how the mood of USA changes.

Data

Initially we collected around 800 tweets from the twitter stream and labelled them. Only the text fields of these tweets were used. All the tweets were classified into happy and sad tweets. These tweets were used to construct our classifier.

Then we collected 100,000 tweets each day from Monday to Sunday (11/23/2015 to 11/29/2015) from the twitter stream. The language of the tweets were specified as English ('en'). Location of the tweets were specified as '-124.637,24.548,-66.993,48.9974' as we wanted tweets from USA

The data collection code always started at 4pm and the code terminated once it collected 100,000 tweets. This took around 2 hours and the code usually terminated roughly around 6pm.

This 100,000 tweets collected each day were stored till data was collected for all 7 days. Once the data collection was completed, 100,000 tweets from each day was used to find the mood of that particular day. The percentage of happy tweets made each day was plotted and was used to observe the transition of the mood of USA during that week.

The code used to collect the tweets is available in the python notebook Data_clctr.ipynb.

Methods

The initial 800 tweets were labelled as happy and sad and were used to make two classifiers- OneVsRestClassifier and a LogisticRegression classifiers with and without TfidfTransformer. The accuracy of the classifiers were computed using k-fold and the classifier with the highest accuracy was used.

We had a tokenize function that removed all mentions, URLs and punctuations. So before training the classifier this tokenize function was used on the tweets.

Then we used the twitter stream API to collect 1000 tweets from USA and use our classifier to find out the most pre dominant mood. This will be declared as the mood of USA at that point of time.

The 700,000 tweets collected throughout the week is used to observe the transition in mood of USA. The 100,000 tweets belonging to each day were stored in separate folders. For example the tweets collected on Monday was stored in a folder called Monday. These 100,000 tweets of each day is stored in 10 text files each containing 10,000 tweets. 10,000 tweets are analysed at one time and the average of moods of these 10 files is used to find the mood of that particular day.

So in the end we got the percentage of sad and happy moods of each day .The percentage of happy mood denotes how happy USA was on that particular day and this was plotted to observe the transition.

Experiments

OneVsRestClassifier and a LogisticRegression classifiers with and without TfidfTransformer were constructed. The accuracy of each classifier was tested using k-fold and it was concluded that OneVsRestClassifier with TfidfTransformer had the highest accuracy:

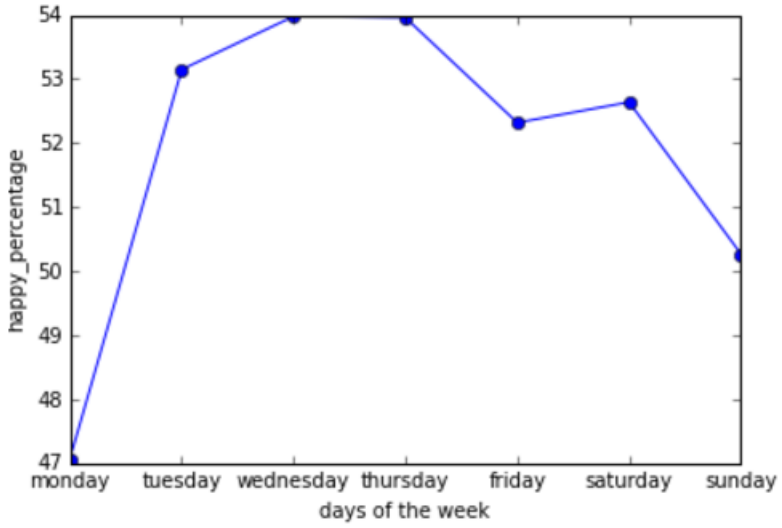
```
OneVsRestClassifier TfidfTransformer enabled
0.657830188679
OneVsRestClassifier TfidfTransformer disabled
0.626438679245
LogisticRegression TfidfTransformer enabled
0.451037735849
LogisticRegression TfidfTransformer disabled
0.578797169811
```

The parameters in countvectorizer used in OneVsRestClassifier with TfidfTransformer were also changed and different combinations were tried out to find the combination with the highest accuracy. The value of binary (True,False), min_df(1 to 11) and max_df(.1 to 1.) were tried and the highest accuracy was observed for binary=True,min_df=1 and max_df=0.2

The data collected throughout the week was used to find the mood of each day from Monday to Sunday and percentage of happy tweets made each day are as follows:

Monday=47.07
Tuesday=53.15
Wednesday=53.98
Thursday=53.96
Friday=52.32
Saturday=52.64
Sunday =50.27

It was observed that Monday had the lowest percentage of happy tweets. Wednesday and Thursday had the highest percentage of happy tweets. The data was collected during the thanksgiving weekend and Wednesday evening was when the holidays started. Thursday evening was the thanksgiving dinner time and thus it is justified that that time would have been the happiest mood in the week and that was reflected in the tweets made it USA.
The percentage of happy moods were plotted in a graph and is given below:



So during that week the saddest day was Monday and the happiness should increase as the week progresses till Wednesday. The happiness should be highest during Wednesday evening and Thursday and then the happiness should gradually go down. There was a slight unexpected dip in happiness during

Friday. We concluded that this could have been happened because of the pressure of getting best deals on the black Friday. Finding the best deals could be frustrating at times.
Thus the hypothesis that the general mood of a region is reflected in the tweets was proven.

Related Works

http://www.instructables.com/id/Mood-Lamp-with-Arduino/
The idea of a mood meter came from this project. The search for some beginner projects in arduino took us to this page where there was a mood lamp. The idea of this project was to detect moods from tweets and then lighting up different LED lights to show different colours for different moods. So the possibility of detecting the mood of the world using tweets came from this project.
Our project inspired from this project implemented its own classifier built from the scratch using the tweets we collected. The idea of tracking the mood of USA as the week progresses was our own.

Conclusions and Future Work

Thus it can be concluded that the mood of a region is indeed reflected in the tweets made from a region and this mood can be monitored from the tweets. And it can also be concluded that Mondays are usually the most depressing day of a week and holidays increase the happiness of the people
The percentage of happy moods were plotted in a graph and this is given below:

More data can be used to train the classifier and the accuracy of the classifier can be increased. The program can be used to track data for more weeks and a more assertive conclusions can be obtained. One month worth of data can be used to track the mood as it progresses from payday to the end of the month. The classifier can also be used to find state wise mood and find the happiest state.