Using Emotions to Predict User Interest Areas in Online Social Networks

Group 3

Ateev Goyal

Meghana Kiran

Nishigandha Kale

Tarun Bakshi

Venkata Hanish Chowdary Navuluri

Table of Contents

Objective

The main objective is to determine the relationship between emotions expressed by social network users and their perceived areas of interest. Also, we can predict the degree of interest in a topic that a social network user has from the emotions they express on the social network (this is based on sample from twitter users)

Data

Data Source is **Twitter tweets**. We have used **Apache Flume Agent** as Data Collection Tool and used **Cloudera** for data processing.

As first step, we have collected the tweets that contains keywords mentioned in the configuration file. Collected data is in JSON format and in key-value pairs.

- **Data Size: 146.8GB**
- **Period of collection: 2 months**

We have followed below steps to process the collected data so that we perform analytical analysis.

- **Step 1 – JSON Extractor for data cleansing**

We have used MapReduce code to run JSON extractor to removed unwanted data and get user-tweet pairs. Input to this program is collected data in JSON format and after processing we get output data i.e. user-tweet pairs.

- **Step 2 – Identify number of tweets per user**

We have used MapReduce job to calculate number of tweets per user. Input data for job is user-tweet pairs generated in Step 1.

- **Step 3 - User Tweet selection**

We have checked if the emotion expressed in the tweet exists in Keyword library or not. If emotion existed in library, we have the selected the user for further processing.

- **Step 4 – User – Emotion pair**

We have picked the user emotion expressed in tweet and mapped it with one of the Ekman's six level emotions. Output of this job is User-Emotion pair where Emotion is from Ekman's emotion framework: Joy, Sadness, Surprise, Anger, Fear and Disgust.

- **Step 5 – Interest Area Annotations**

Then we manually identified standard users from collected User-Emotion data. We have taken 50 standard users and 20 tweets per user for further analysis. We have ranked identified users' interest in possible areas (Sports, Movies, Technology Computing, Politics, News, Economics, Science, Arts, Health and Religion) on a scale of 5, where 5 means highly interested and 1 means least interested.

Methodology

We have performed below steps as part of our methodology to determine if it is possible to predict user area of interest based on emotions expressed in their tweets.

- **User Emotion Score**

We have analyzed all user-emotion pairs for each user and aggregated user emotions into a single value for each emotion, each user. For this we have used a MapReduce program to take User-Emotion pairs as input data and calculate mean for each emotion expressed by the user. Figure 1 pie chart is generated using R. Figure 1 shows that the emotion proportion scores for Ekman's six basic emotions.

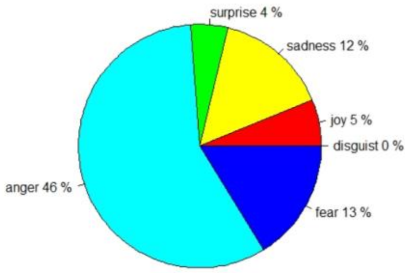
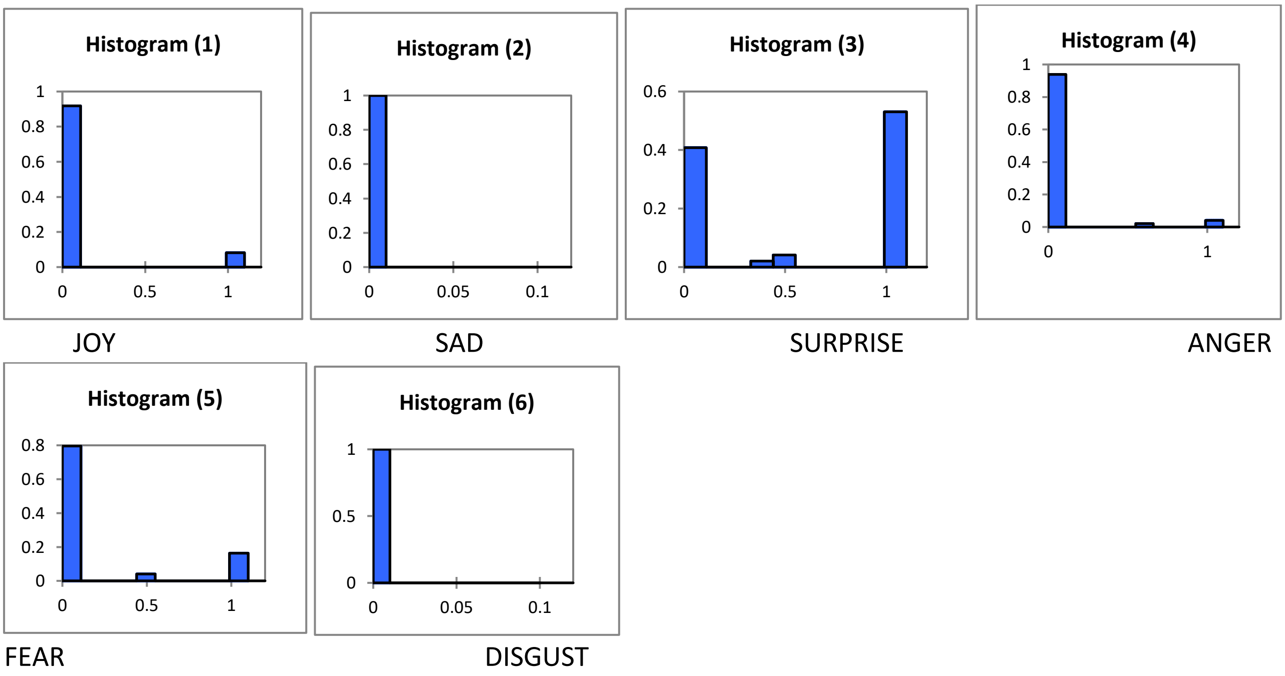


Fig 1. Mean Proportion scores for Ekman's six basic emotions

As an alternative method, we have used a MapReduce program to use probability to calculate user emotion score. We calculated distribution score of an emotion for a user across all his/her tweets. Output of this program is a csv file that is used for further analysis.

Below Fig. 2 shows histogram of the emotion scores across users. Figure shows the large variability in emotions and histograms are not normally distributed, for example Surprise emotion is widely distributed.



JOY SAD SURPRISE ANGER

FEAR DISGUST

Fig. 2 Histogram of Emotion Scores Across Different Users

Y axis –proportion of users in dataset, X-axis- proportion of emotion score

- **Emotion and Interest Ares relationship**

We have used Mann-Whitney U-test to identify the relationship between user emotion and their area of interest.

We used this method to identify whether user is interested or disinterested in Areas of Interest or not. We have used XLSTAT tool to analyze all users' data for Mann-Whitney U-test and Table 1. Output.

	Joy	Sad	Surprise	Anger	Fear	Disgust
Sports	<= 0.0001	<= 0.0001	<=0.043	<= 0.0001	<= 0.0001	<= 0.0001
Movies and Television	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001
Technology, computing and internet	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001
Politics, Society and news	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001
Business, economics and finance	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001
Arts, Painting and dance	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001
Science and environment	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001
Health and Medician	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001
Religion and Spiritualism	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001
Travel and Leisure	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001
Fashion	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001	<= 0.0001

Table 1. Pearson correlations between user interest score and emotion distribution score

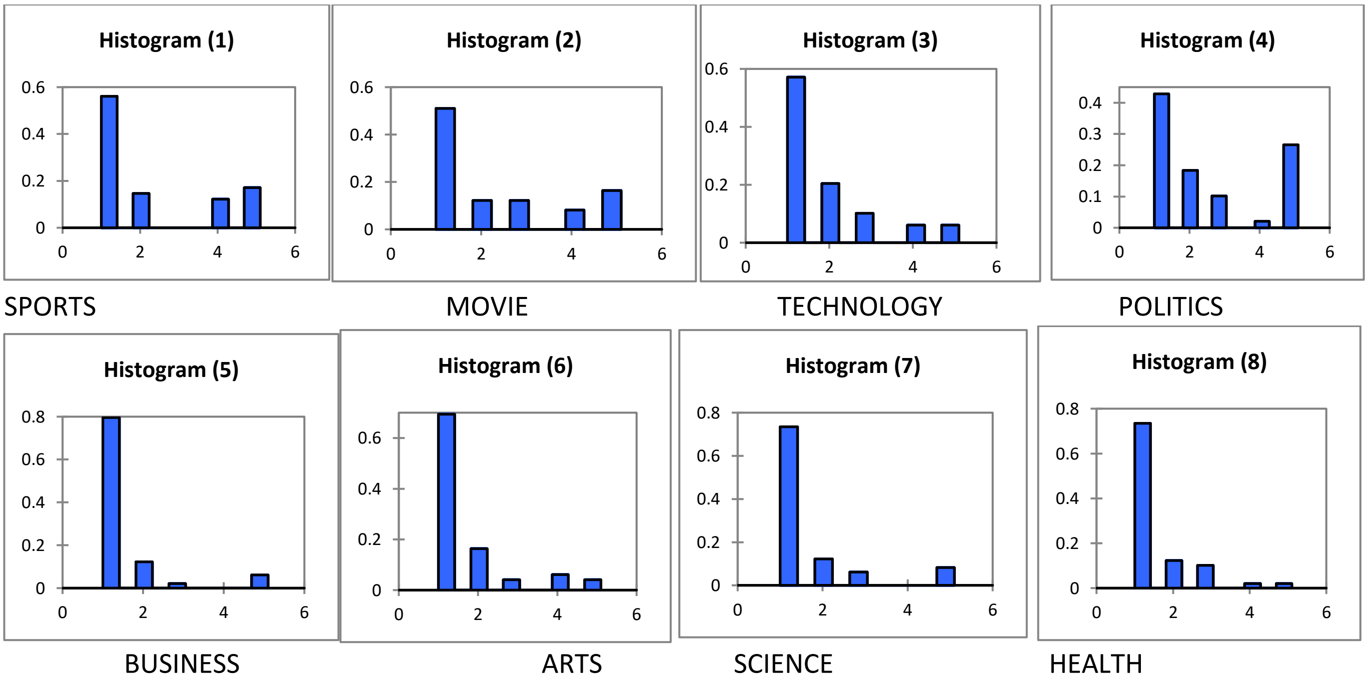
Also, we have used only Highly interested and Least interested users' data and performed Mann-Whitney U-test. Table 2. shows the output.

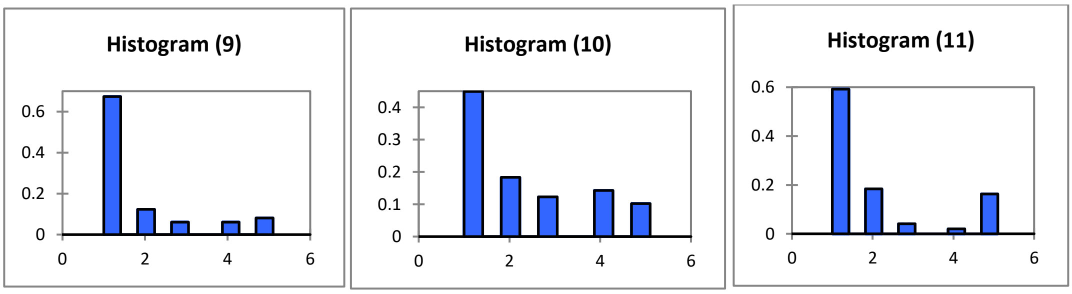
	Joy	Sad	Surprise	Anger	Fear	Disgust
Sports	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Movies and Television	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Technology, computing and internet	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Politics, Society and news	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Business, economics and finance	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Arts, Painting and dance	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Science and environment	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Health and Medician	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Religion and Spiritualism	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Travel and Leisure	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Fashion	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001

Table 2. Pearson correlations between user interest score and emotion distribution score

Because of limited data for our analysis and only one emotion expressed dominantly, our result shows same P-value for most of the Areas of Interest but from Table 1. And Table 2. we can say that there is a possible correlation between user emotion and their area of interest.

Fig. 3 represents histogram of user interest scores for each of the Area of Interest across all users.





RELIGION TRAVEL FASHION

Fig. 3 Histogram of User Interest Score

Y-axis –Proportion of their interest, X-axis –interest level (5- most interested, 0-Least Interested)

To analyze the prediction quality for Area of Interest, we have used linear regression model. Below Fig. 4 shows the prediction quality for the regression model measured in R2 values. It shows that for all the studied interest areas we studied, it is possible to determine whether a person is very interested or disinterested based on the emotions he/she has expressed. The best prediction was achieved for Health and Medicine, Sports and Movies and Television, and worst were achieved for Science and environment and Religion and Spiritualism.

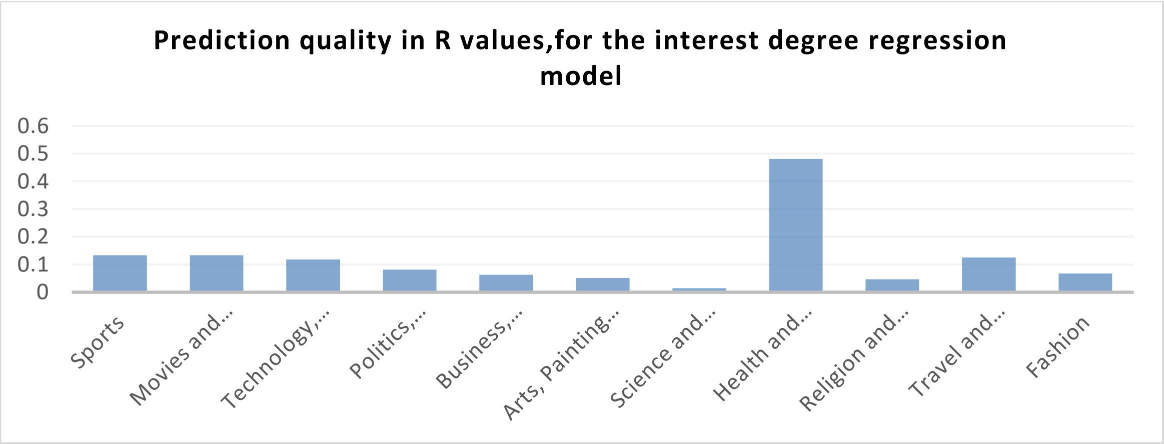


Fig. 4 - Prediction quality in R values, for the interest degree regression model

Conclusion

After above analysis of user emotions expressed on social media, we have observed that there is a high possibility of a correlation between users expressed emotions and their Interest areas. Our results have many applications in online advertising like it can be used to predict user interests to personalized advertising content. Another application of our application can be used in social and psychological logical research.