What we post, reveals who we are:

Relating Facebook posts to personalities

Akash Parvatikar

Mohammed Aldhoayan

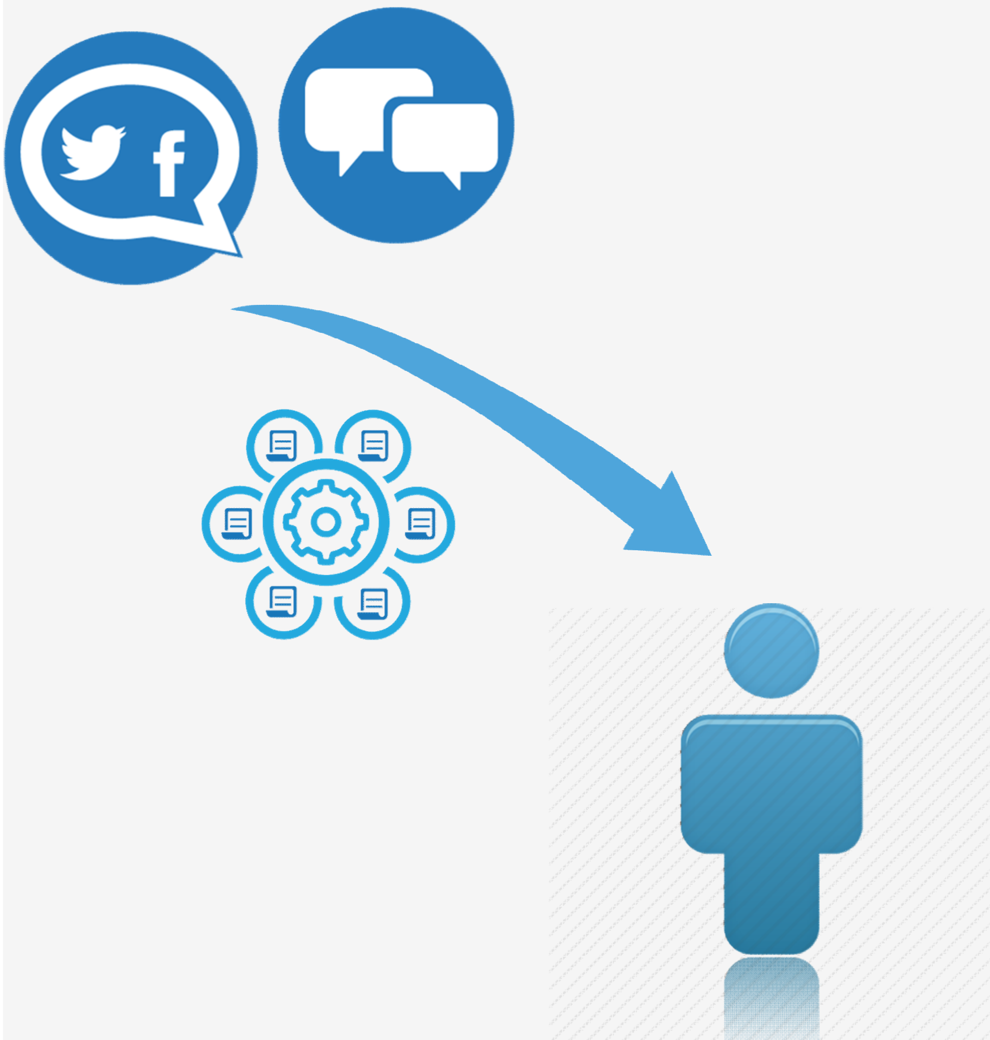
Yuqian Zhanfu

INFSCI 2160 | Data Mining | Group 16

CONTENT

01 PART ONE

PROJECT BACKGROUND - Motivation



- People's personality detection can be used for personalized marketing, recommendations to match individuals with similar personality and academia research.
- Social media data reflects people's traits through their posts at different times.
- Text mining can be used to unravel hidden topics in the individuals' post which may reflect their state of mind.

01 PART ONE PROJECT BACKGROUND – Related Works

Previous Works

1. Generate data analysis based on age, sex and demographic info. (Markovikj et al. 2013)
2. Directly using Facebook posts features to predict big five personality (Celli, et. al 2013).
3. Similar work was done on the Twitter data (Quercia et. al 2011)

Our Works

1. Acquire Facebook posts from the 'myPersonality' dataset.
2. Improve personality predictions based on temporal factors, emotions extracted from the posts and Network size of individuals.
3. Built emotions – words dictionary using Python2.7 using 'EmoLex' Lexicon to weigh each post with respect to the emotion.

? Do individuals' Facebook posts, network size and activity timing reveal their personalities

02 PART TWO

DATA PROCESSING

Previous Dataset

Python

+

'Emolex' lexicon

Processed Dataset

STATUS
#AUTHID
BETWEENNESS
NETWORKSIZE
DENSITY
BigFive personality predicted Labels
BigFive personality SelfLabels

STATUS
#AUTHID
NETWORKSIZE

10Emotions features

BigFive personality predicted Labels
BigFive personality Self Labels

02 PART TWO

DATA PROCESSING

?How do “emotion scores” come out

- For each post:
  - -- Remove stop words
  - -- Remove punctuations
  - -- Bag-of-words
  - -- For each word in the post:
    - i) Acquire word- emotion association score
  - -- Aggregate the scores of all words for each emotion

  End

PART TWO 02

DATA PROCESSING

!Let's try to extract emotions from some posts

Post 1: I enjoy reading news

Post After Cleaning: enjoy reading news

anger: 0
anticipation: 1.87
disgust: 0.06
fear: 0.01
joy: 0.95
negative: 0
positive: 2
sadness: 0.13
surprise: 0.51
trust: 1

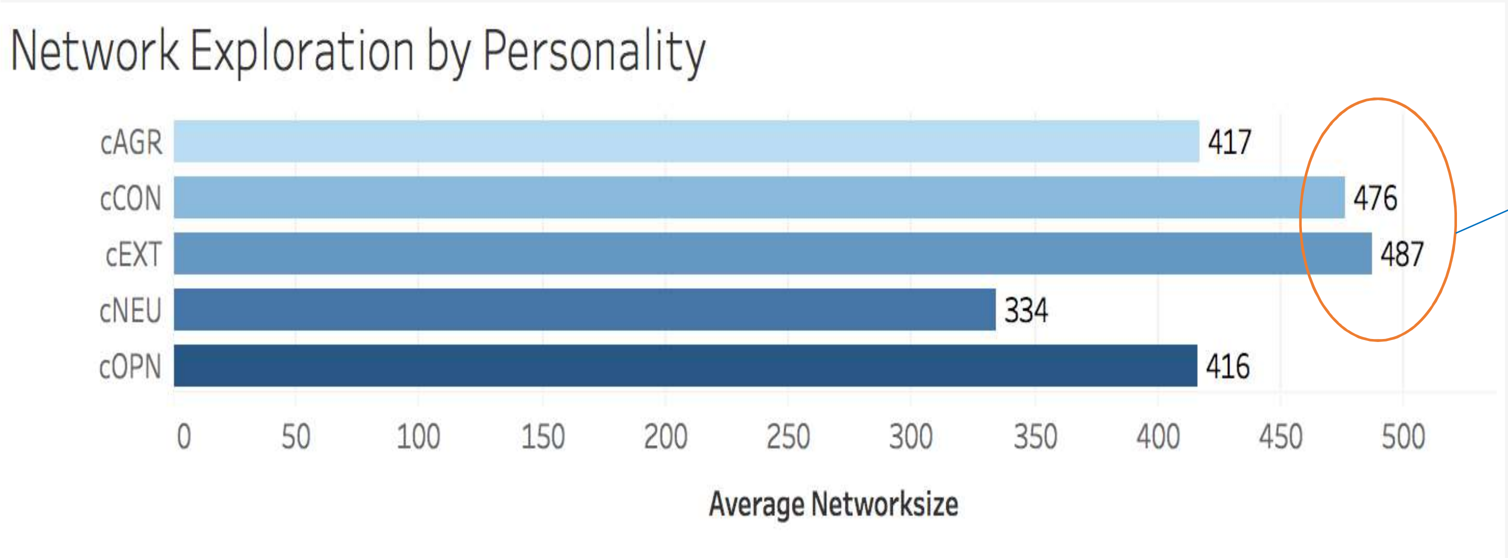
Post 2: I am hungry and sleepy

Post After Cleaning: hungry sleepy

anger: 0.868411622517
anticipation: 1
disgust: 0
fear: 0
joy: 0.0204051289119
negative: 1
positive: 0
sadness: 0.422351251604
surprise: 0
trust: 0

03 PART THREE DATA EXPLORATION

How is the Network size for each personality group?

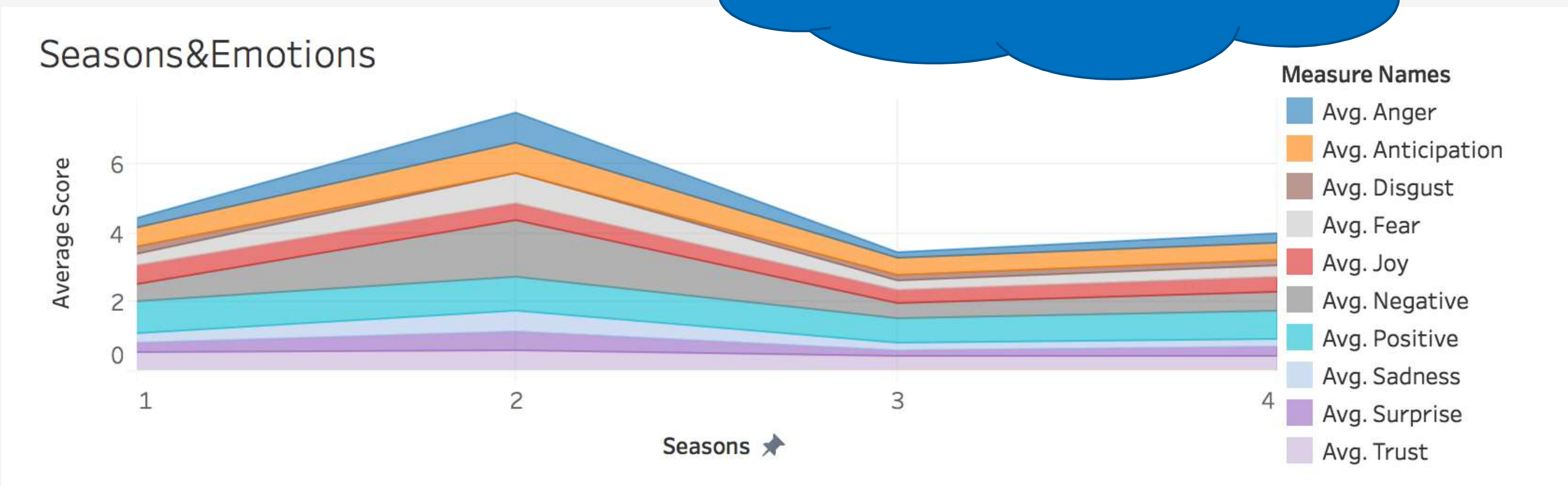


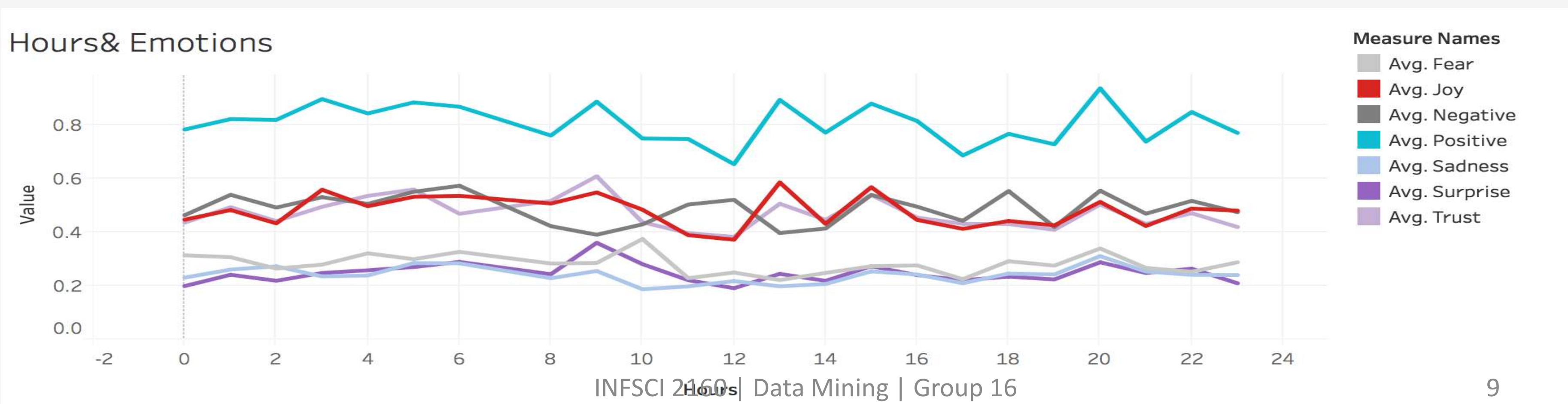
Conscientiousness & Extraversion

03 PART THREE

DATA EXPLORATION

How is the emotion pattern？





03 PART THREE

DATA EXPLORATION

How about the Emotions score appear in each personality?

04 PART FOUR

APPROACH 1

- Train the classifier using emotion scores from each Facebook post along with the personality of their author.
- From the data, each post is considered separately and we try to predict the personality of each post's author.
- When testing, for each individual or author, we take majority of the votes of his/her posts and conclude the personality.
- 9000+ posts were included and 10-fold algorithm was applied taking 10% of the posts for testing and remaining as the training data.

04 PART FOUR

APPROACH 2

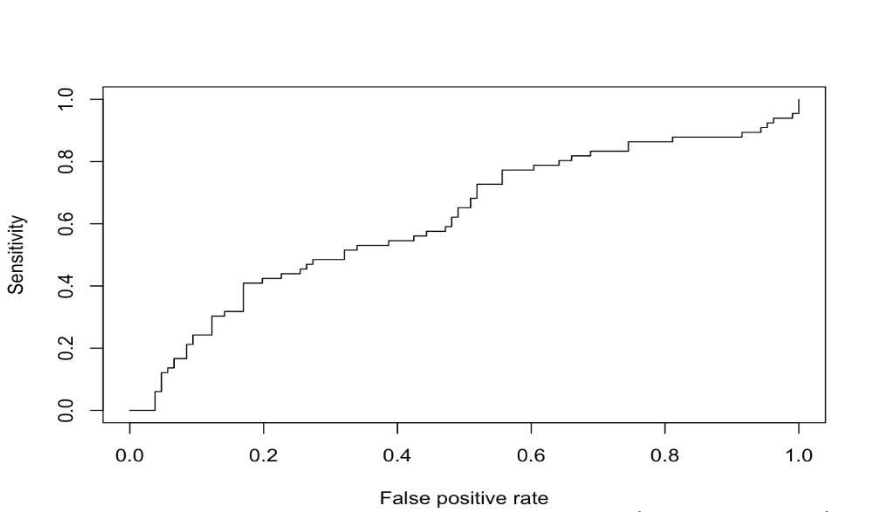
- We average emotions in all the posts for individuals who have more than 10 posts.
- Train the classifier using the average emotions scores from all Facebook post for each individual along with their personality.
- For testing, we predict the personality of each individual from the average of the emotions in their posts.
- In each fold, we use 10% of the users and not the posts for testing data and 90% for training.

PART FOUR

04 PREDICTION APPROACH

EMOTION: EXTRAVERSION

ALGORITHMS	PRECISION	RECALL	F-SCORE	AUC
GLM	0.915	0.642	0.75	0.615
D-TREE	0.701	0.626	0.65	0.524
ADA	0.746	0.668	0.699	0.616
NB	0.435	0.671	0.508	0.596



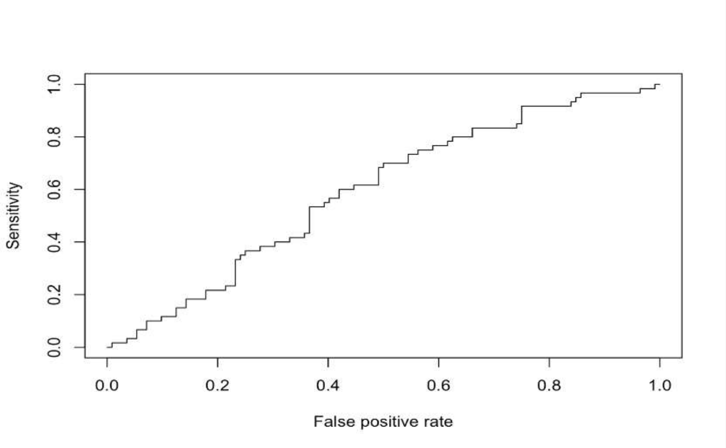
Best Algorithm: GLM for Extraversion emotion

PART FOUR

04 PREDICTION APPROACH

EMOTION: NEUROTICISM

ALGORITHMS	PRECISION	RECALL	F-SCORE	AUC
GLM	0.836	0.667	0.725	0.581
D-TREE	0.751	0.686	0.697	0.562
ADA	0.785	0.667	0.701	0.592
NB	0.808	0.677	0.715	0.553



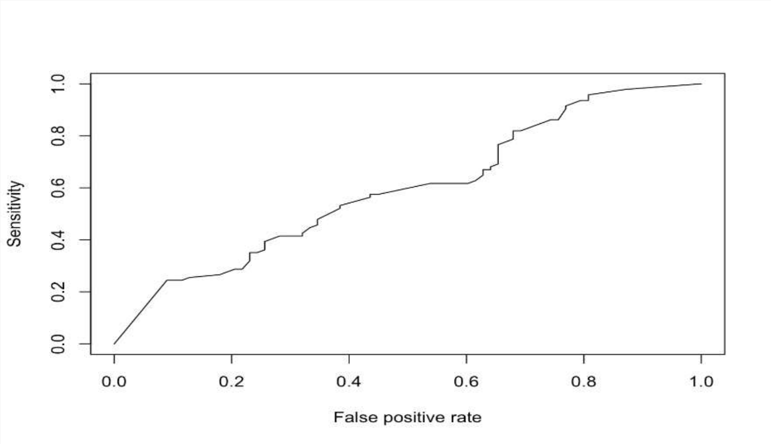
Best Algorithm:
ADA for Neuroticism emotion

PART FOUR

04 PREDICTION APPROACH

EMOTION: AGREEABLENESS

ALGORITHMS	PRECISION	RECALL	F-SCORE	AUC
GLM	0.502	0.595	0.506	0.642
D-TREE	0.569	0.549	0.532	0.597
ADA	0.423	0.502	0.431	0.542
NB	0.667	0.464	0.510	0.409



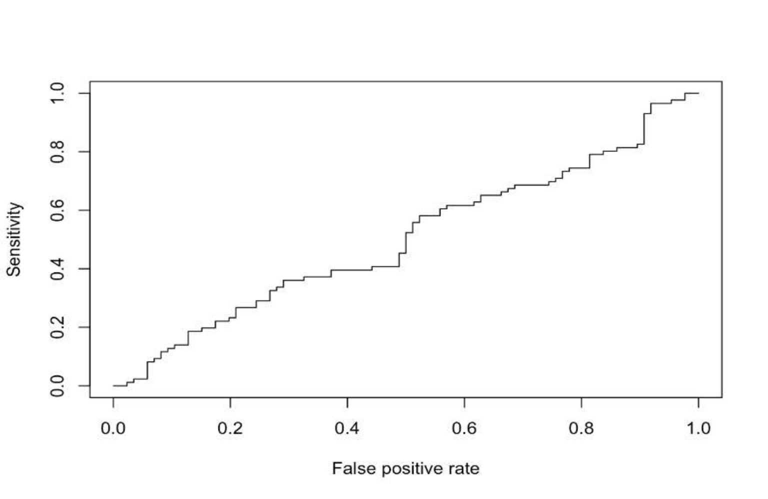
Best Algorithm: D-TREE for Agreeableness emotion

PART FOUR

PREDICTION APPROACH

EMOTION: CONSCIENTIOUSNESS

ALGORITHMS	PRECISION	RECALL	F-SCORE	AUC
GLM	0.502	0.48	0.51	0.501
D-TREE	0.46	0.5	0.447	0.511
ADA	0.438	0.487	0.429	0.477
NB	0.374	0.49	0.43	0.518



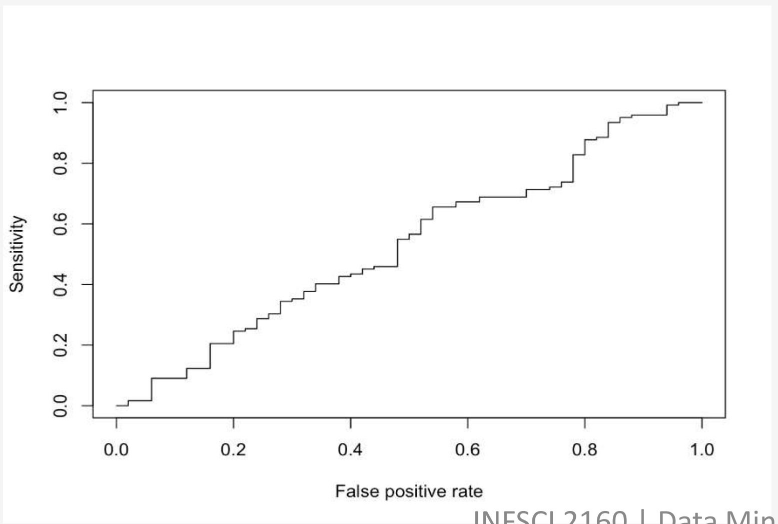
Best Algorithm:
GLM for Conscientiousness emotion

PART FOUR

04 PREDICTION APPROACH

EMOTION: OPENNESS

ALGORITHMS	PRECISION	RECALL	F-SCORE	AUC
GLM	0.846	0.306	0.43	0.53
D-TREE	0.472	0.303	0.41	0.511
ADA	0.655	0.249	0.35	0.404
NB	0.389	0.323	0.33	0.510



Best Algorithm:
GLM for Openness emotion

05 PART FIVE

DISCUSSION

- Predicting personality is difficult but achievable due to the variety of possibilities that the data has to offer
- Lack of feature space makes it more complicated to derive prediction models from the available data
- Sparse distribution of predicted emotions from posts poses a challenge to predict the personality
- Thus, careful selection of parameters and picking the number of users for certain number of posts is done to improve the algorithm

06 PART SIX

CONCLUSION

- Different algorithms were tested for the 5 personalities. Depending on the performance measures, appropriate algorithms were chosen.
- Data preprocessing is a critical phase for any data mining project.
- The more you post on social media, the more we know you!!
- The Network is a good feature to predict personalities.
- Due to sparseness in the dataset and no strong predictive features, prediction although seems difficult but an improvement in the approach is presented

PART SIX

CONCLUSION- FUTURE WORK

- Test the model on a larger dataset
- Detect the emotion pattern based on temporal scale
- Connecting people with similar pattern of emotions together
- Improve model's performance