Data Science Project

Mid-term Project Presentation

5th November, 2019

Team User13 (Tempête De Données)

Team Members:

Akshay Singh Rana
Harmanpreet Singh
Himanshu Arora
Nitarshan Rajkumar
Sreya Francis

Mila

Université de Montréal

Problem Statement

User modeling with multi-source user data such as text, images, and relations to arrive at accurate user profiles.

Mila

Université de Montréal

Prediction task overview

Classification Tasks:

- Categorical age
- Gender

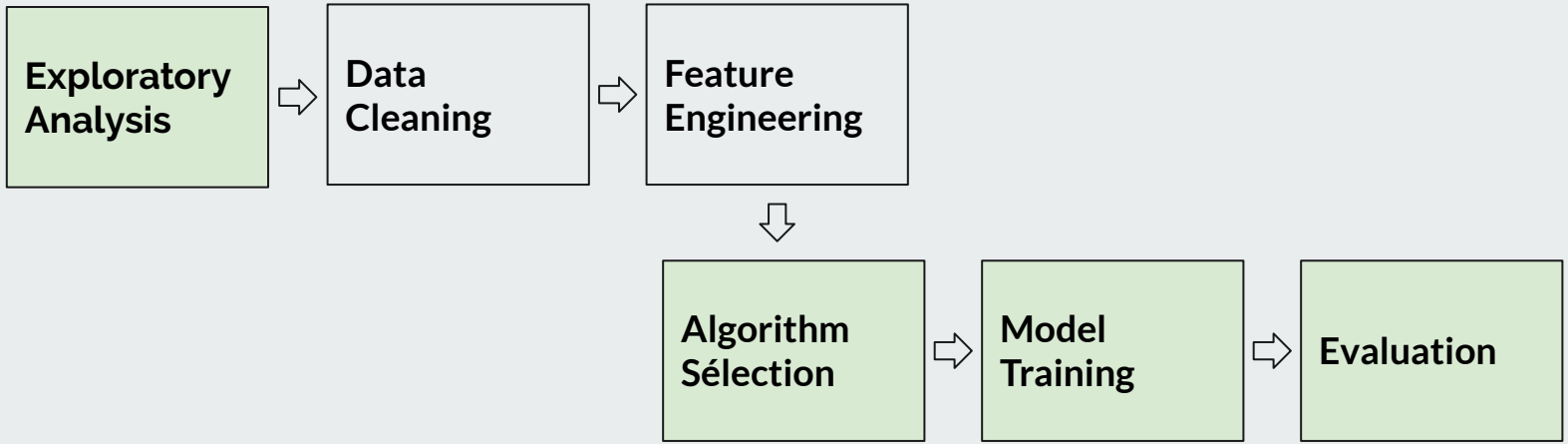
Régression Tasks:

- Personality Score Prediction
  - Openness
  - Conscientiousness
  - Extroversion
  - Agreeableness
  - Neuroticism

Mila

Université de Montréal

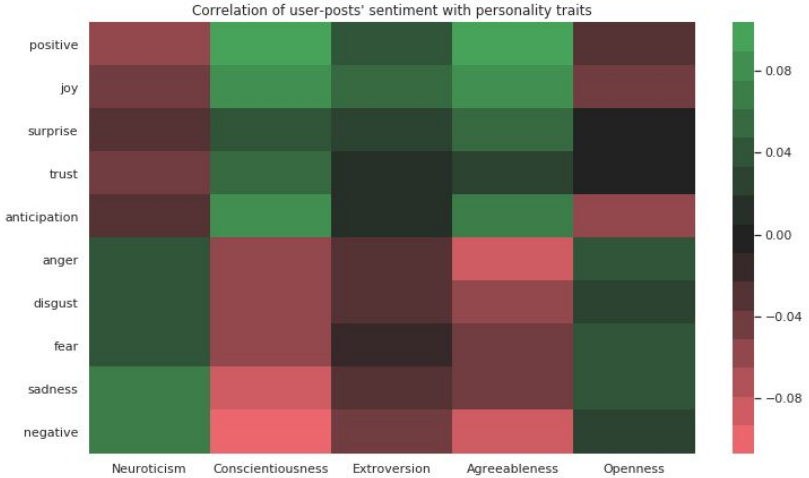
Pipeline

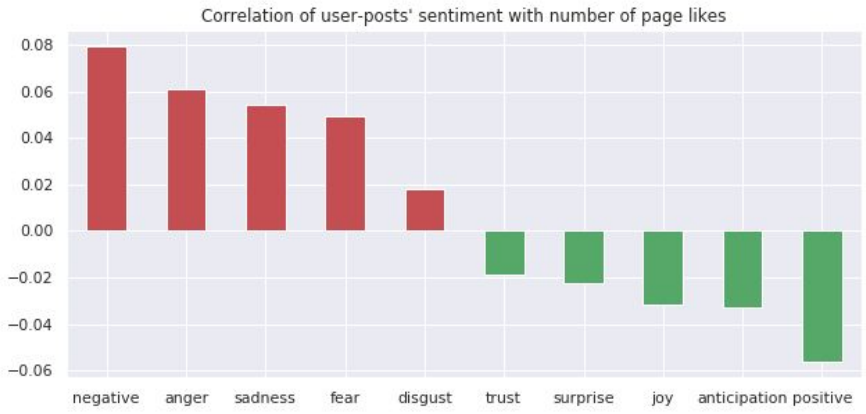

Exploratory Analysis
Data Cleaning
Feature Engineering
Algorithm Sélection
Model Training
Evaluation

Data Sources

- Text - LIWC + NRC
- Image - Oxford Features
- User-Page-likes
- Feature Stats
  - Number of users: 9500
  - Total number of features: 65 (oxford) + 1 (relationships) + 81 (liwc) + 10 (nrc)
  - Missing images for 2326 users
  - Multiple faces in images of ~700 users

Exploratory Analysis (1/3)

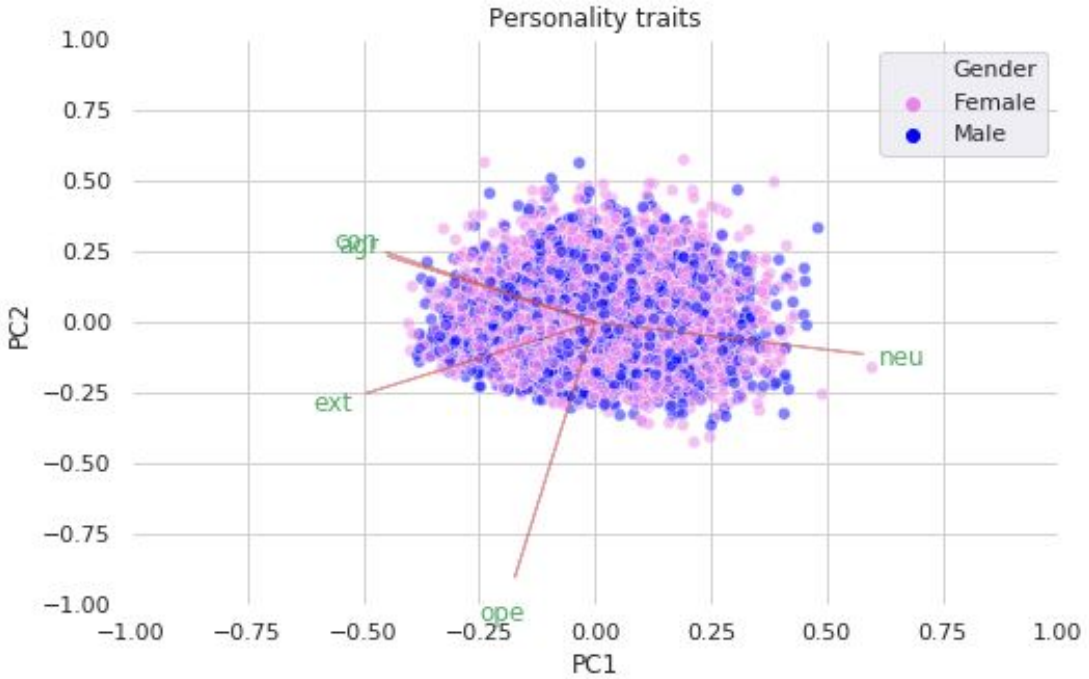




Mila

Université de Montréal

Exploratory Analysis (2/3)



Mila

Université de Montréal

Exploratory Analysis (3/3)

- Using Embedded feature selection methods (Lasso and Random Forest)
- Important features identified for age and gender prediction:
  - facialHair_mustache
  - facialHair_beard
  - facialHair_sideburns

Prediction Task (1/2)

Tasks	Classification (Acc, higher is better)				Régression (RMSE, lower is better)				
	Gender		Âge		OPN	NEU	EXT	AGR	CON
Baseline	0.594		0.591		0.652	0.798	0.788	0.665	0.734
SVM	0.613	0.583	0.591	0.583					
Random Forests	**0.871**	0.647	0.571	**0.621**	**0.605**	**0.785**	**0.771**	**0.632**	**0.707**
Features Used	**Oxford**	**LIWC + NRC**	**Oxford**	**LIWC + NRC**	**LIWC + NRC**	**LIWC + NRC**	**LIWC + NRC**	**LIWC + NRC**	**LIWC + NRC**

Mila

Université de Montréal

Third Source: User-Page Like

Shortlisted the pages with more than 10 likes.

Converted the data into a multi-one hot encoding.

	Page 1	Page 2	Page 3	Page 4
User 1	1	0	1	1
User 2	0	0	1	0
User 3	1	0	0	1

Age	Gender
24	Male
35	Male
58	Female

Prediction Task (2/2)

Tasks	Classification (Acc, higher is better)						Régression (RMSE, lower is better)					
	Gender			Âge			OPN	NEU	EXT	AGR	CON	
Baseline	0.594			0.591			0.652	0.798	0.788	0.665	0.734	
SVM	0.613	0.583	**0.819**	0.591	0.583	**0.670**	-	-	-	-	-	
Random Forests	**0.871**	0.647	0.788	0.571	0.621	0.660	**0.605**	**0.785**	**0.771**	**0.632**	**0.707**	😞
Features Used	**Oxford**	**LIWC + NRC**	**Page Likes**	**Oxford**	**LIWC + NRC**	**Page Likes**	**LIWC + NRC**	**LIWC + NRC**	**LIWC + NRC**	**LIWC + NRC**	**LIWC + NRC**	**Page Likes**

Mila

Université de Montréal

Learnings

- Improve the encodings in the user page-like data using Node2Vec, etc.
- All three data sources are important and we can leverage them by fusing them together.
- Endless possibilities of stacking models based on different features, algorithms and data sources and fusing them all together.
- We can also stack models with different tasks and combine all the task in the end.

Further Steps

- Investigate stacking features and models
- Feature Engineering
  - Forward search
  - Domain knowledge
  - Node2Vec
- Better Models
  - Gradient Boosted Trees (XGBoost)
  - Neural Networks
- Hyperparameter Search

Q & A