IFT 6758 Final Project

Final Project Presentation

3rd December, 2019

Team: User13 (Tempête De Données)

Team Members:

Akshay Singh Rana
Harmanpreet Singh
Himanshu Arora
Nitarshan Rajkumar
Sreya Francis

Scoreboard

How we beat all the baselines!

AGE	GENDER	OPN	NEU	EXT	AGR	CON
0.621	0.827	0.639	0.790	0.780	0.651	0.713

Problem Statement

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

Prediction Task Overview

Classification Tasks

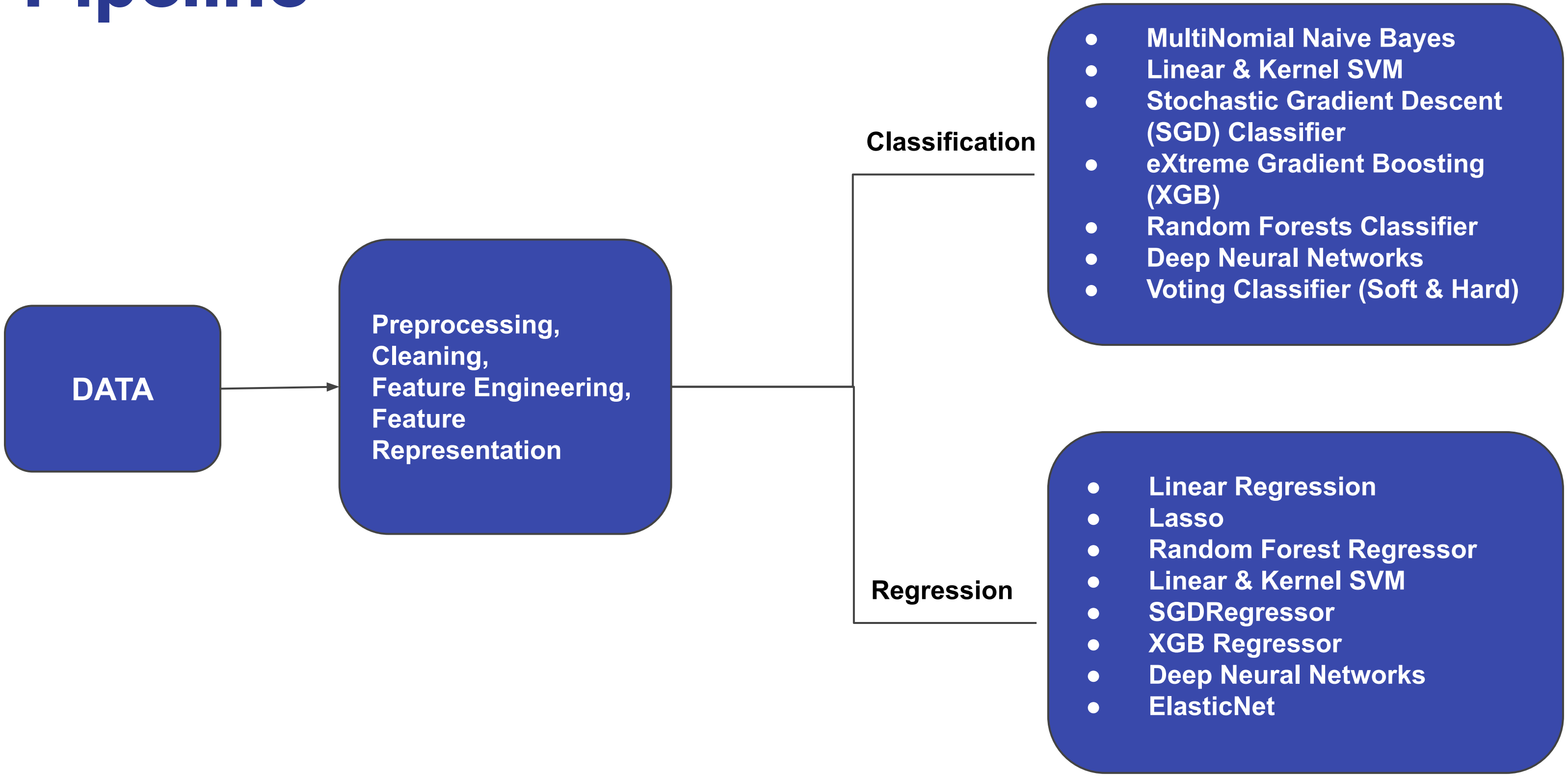
- Categorical age
- Gender

Regression Tasks

Régression Task:

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

Pipeline


DATA
Preprocessing,
Cleaning,
Feature Engineering,
Feature
Representation
Classification
MultiNomial Naive Bayes
Linear & Kernel SVM
Stochastic Gradient Descent (SGD) Classifier
eXtreme Gradient Boosting (XGB)
Random Forests Classifier
Deep Neural Networks
Voting Classifier (Soft & Hard)
Regression
Linear Regression
Lasso
Random Forest Regressor
Linear & Kernel SVM
SGDRegressor
XGB Regressor
Deep Neural Networks
ElasticNet

Feature Analysis

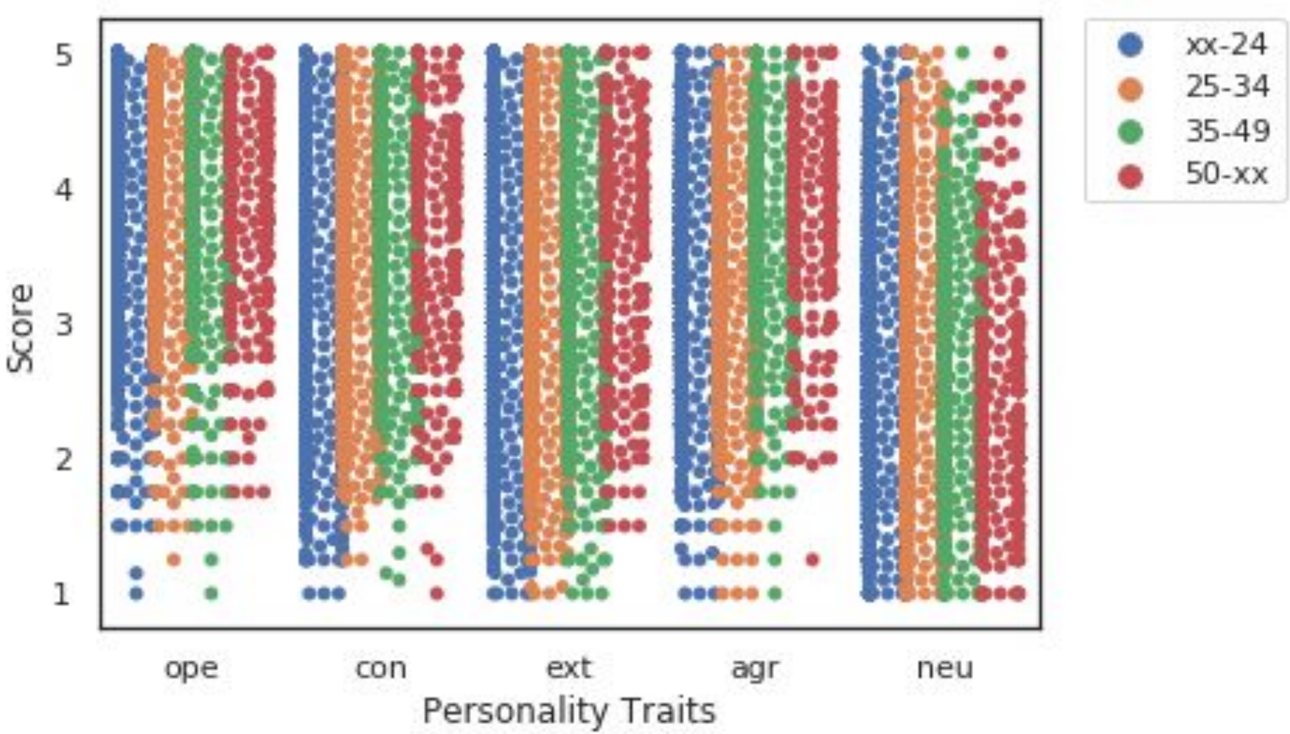
Features

1. Text: LIWC + NRC
2. Image: Oxford features
3. Graph: Users' page likes

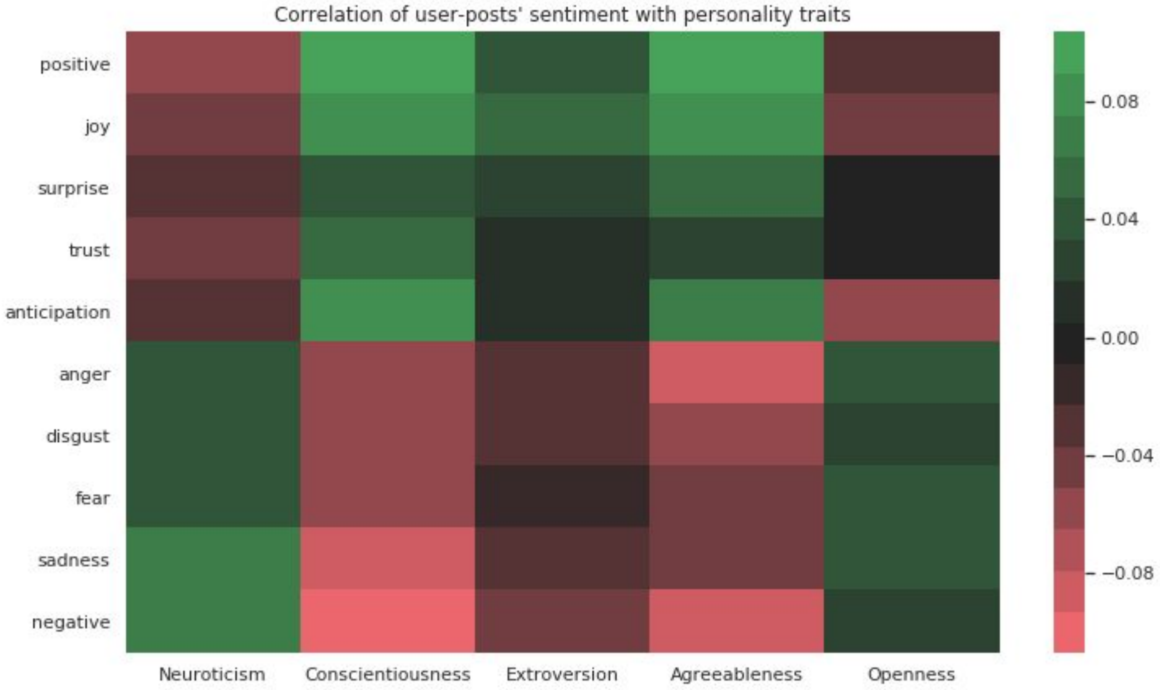
Feature Statistics

- 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

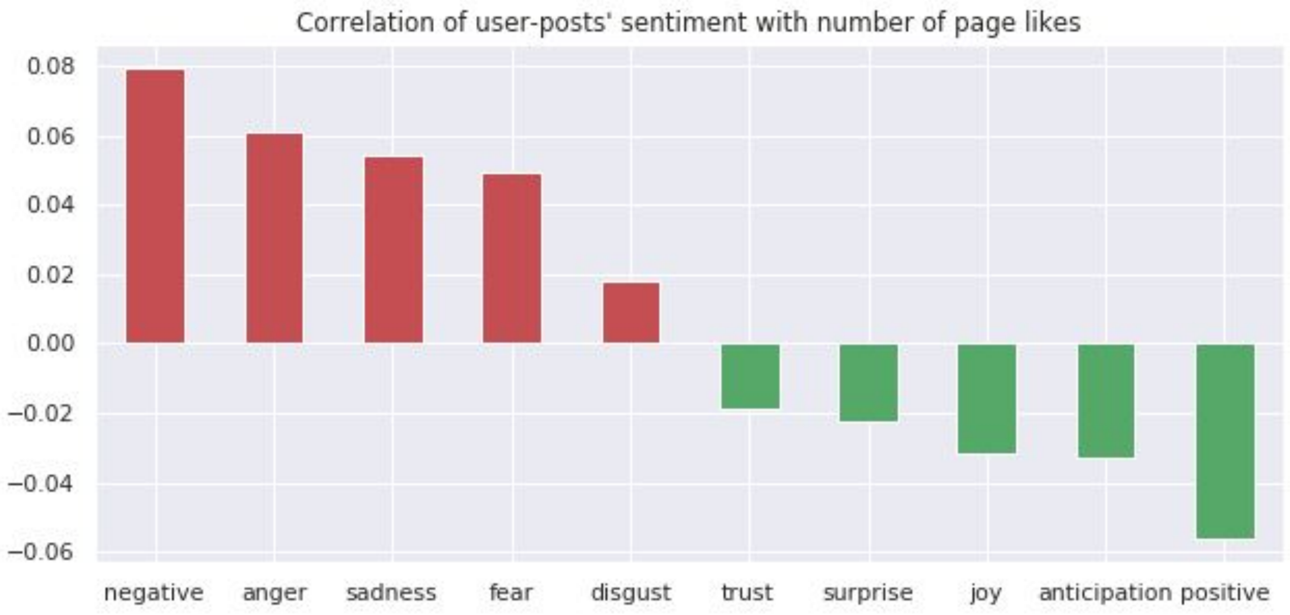
Personality vs Age



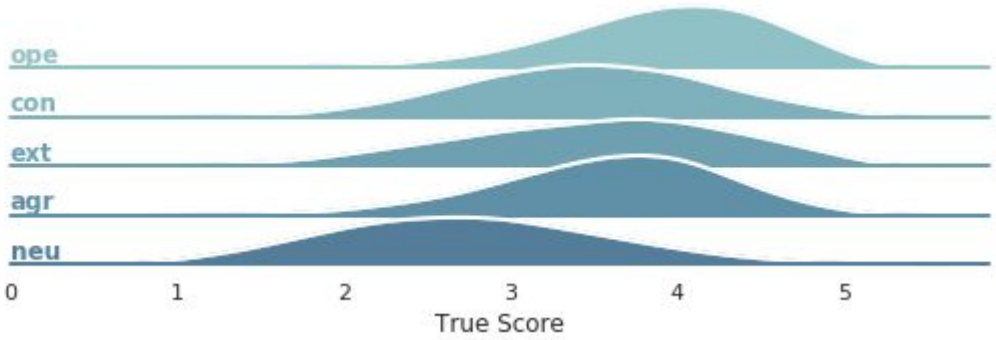
Personality vs User Posts

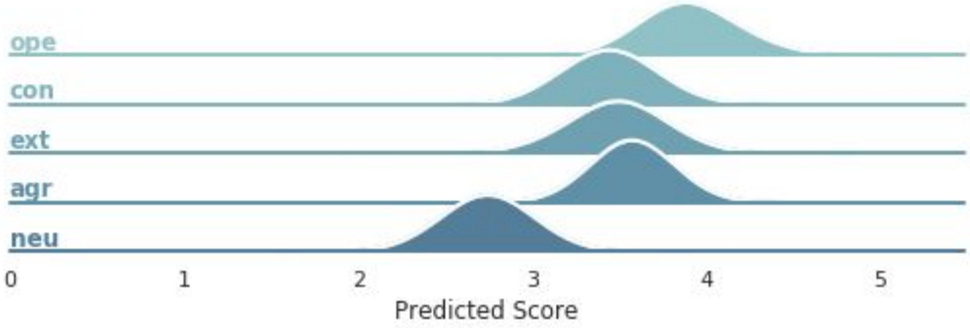


Page Likes vs User Posts

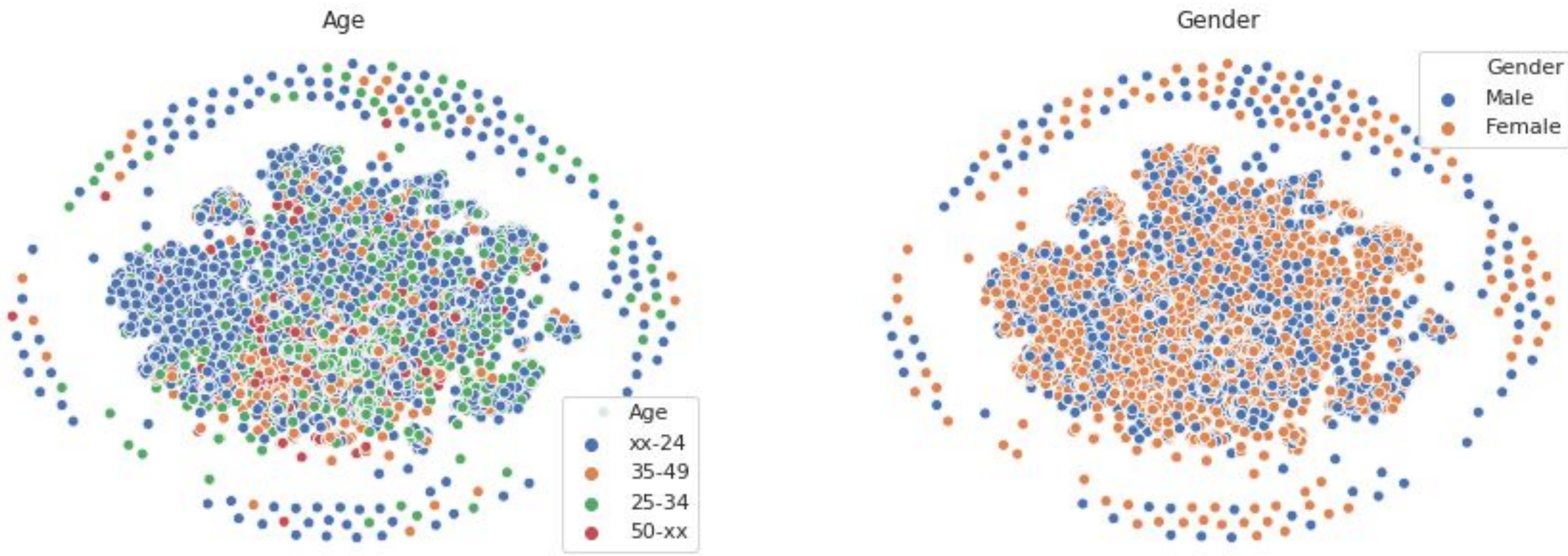


Personality Prediction Analysis

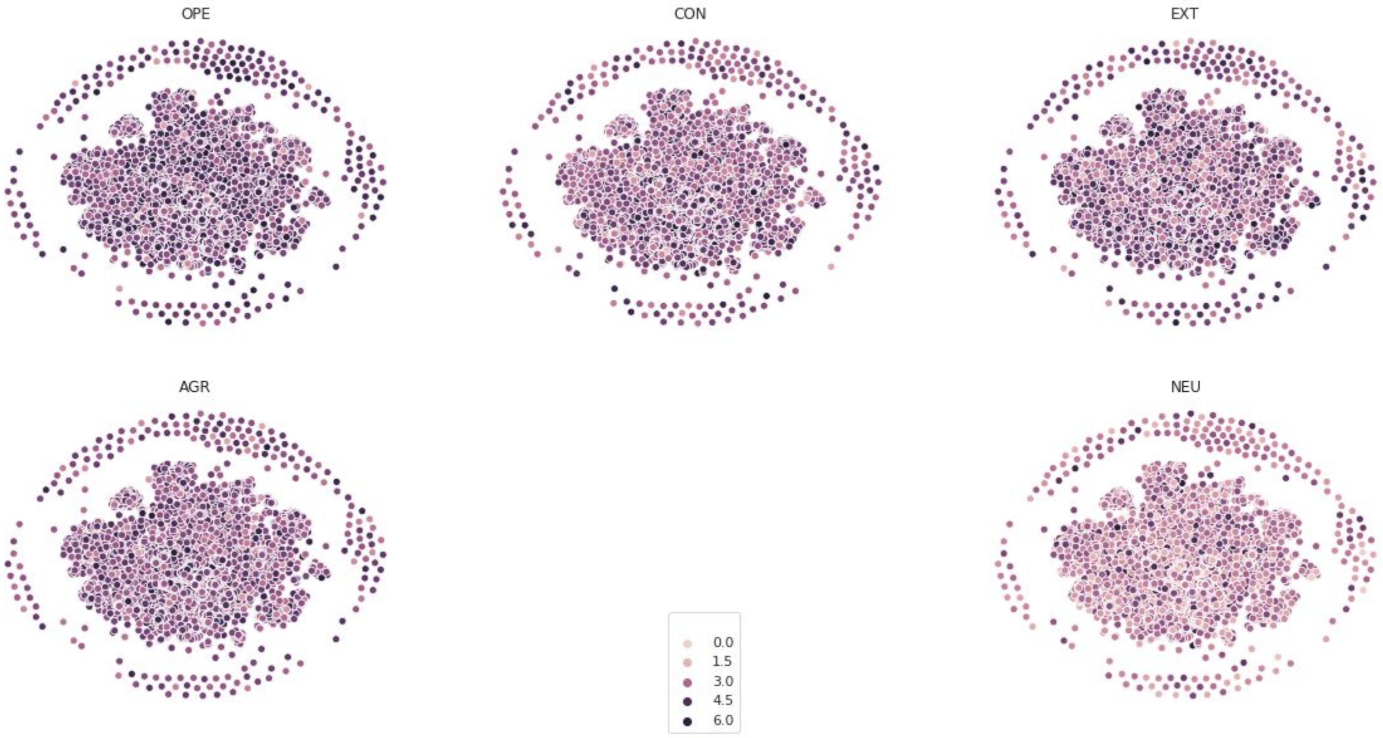




Node2Vec Classification

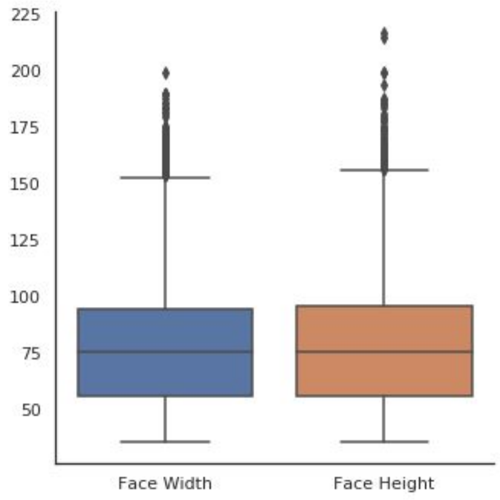


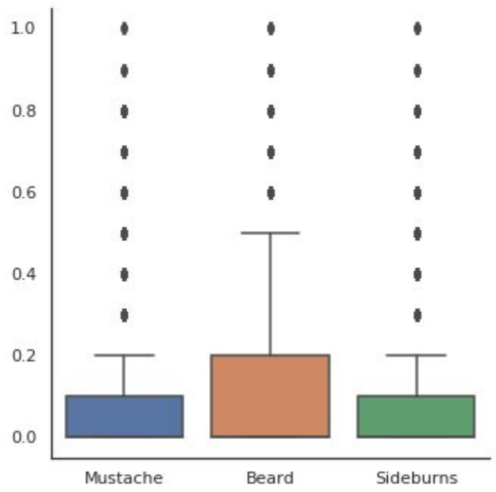
Node2Vec Regression



Feature Preprocessing

- Removed Outliers (using IQR)
- Scaling: Min-Max/Standard Scaling





Feature Selection

- Using Filter-based methods (SelectKBest)
- Using Embedded feature selection methods (Lasso LR and RandomForest)
- Important features identified:

FACIAL	LIWC	NRC
facialHair_mustache	ipron	negative
facialHair_beard	swear	anger
facialHair_sideburns	social	disgust
faceRectangle_width	negemo	fear
faceRectangle_height	feel	joy

Feature Engineering

- Merged facial features such as facial hair
- Converted raw facial coordinates into lengths and areas
- Created node embeddings for users based on page likes, text, and facial features

Feature Representation

Relation feature representation

- **Relations**
  - Multi-one hot encoding
  - Like-Frequency Inverse User Frequency (similar to tf-idf)
  - Weighted and Unweighted Node2vec

Multi-one hot encoding

- Creates sparse matrix containing user and likes
- Experimented with shortlisting pages with different thresholds: **0, 5, 10, 25**
- Converted 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

Like-Frequency Inverse User Frequency

- Creates sparse user-like matrix
- Experimented with shortlisting pages with different thresholds: **0, 5, 10, 25**
- Converted data into a multi hot encoding using approach similar to tfidf.

	Page 1	**Page 2**	**Page 3**	**Page 4**
User 1	0.25	0	0.2	0.4
User 2	0	0	0.4	0
User 3	0.25	0	0	0.2

Node2vec

- Optimize embeddings with biased random walks, using word2vec skip-gram model
- Learn low dimensional latent representations by projecting users and other entities as graph

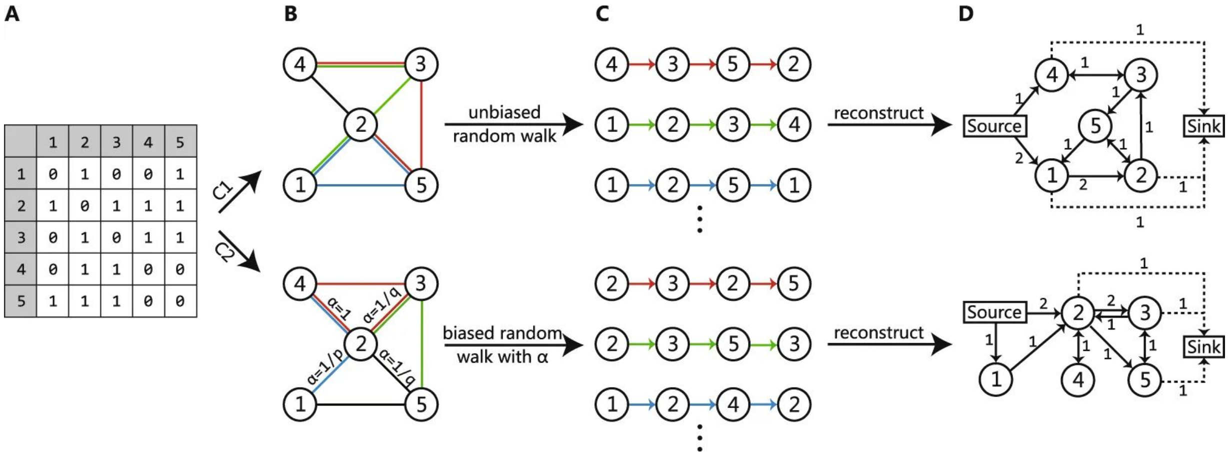
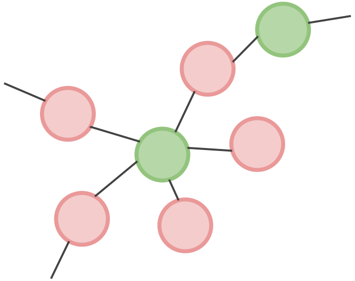
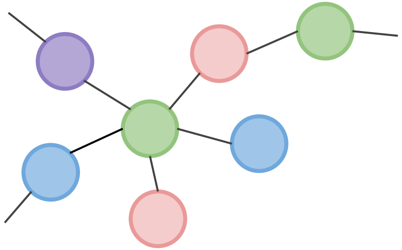


Image: https://www.nature.com/articles/s41598-017-12586-y

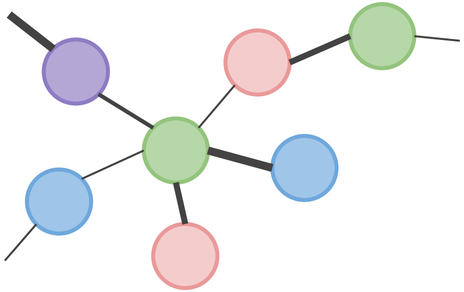
Node2vec Heterogeneous Graph Formation



(Un)-Weighted Graph
With only 2 types of nodes



(Un)-Weighted Graph
With multiple types of nodes



Weighted Graph

- With multiple types of nodes
- Page weights were assigned using tfidf kind of approach
- Improve weight to pages with less number of likes

Legend:

- Users
- Pages
- Selected Facial Features
- Selected Text Features

Embedding dimension - 128
Number of walks - 10, 20
Walk length - 50, 60, 80
Return hyperparameter - 0.75, 0.8, 0.9, 1
Inout hyperparameter - 0.9, 1.0
Context window size - 10
Number of iterations - 1, 2, 5

Experimented Methods

Approach 1

Using Individual Data Sources

Train all data sources individually for each task to find the best algorithm

Classification Tasks Evaluation

	Classification Accuracy							
	Gender				Âge			
Baseline	0.594				0.591			
Features Used	**Oxford**	**Text**	**N2v**	**Page Likes**	**Oxford**	**Text**	**N2V**	**Page Likes**
Random Forests	80.00	54.10	56.26	77.80	59.91	50.11	61.79	62.50
Linear SVM	68.11	55.10	56.05	78.50	61.47	61.47	61.79	**67.56**
MNB	71.9	44.15	-	54.61	-	53.17	-	60.15
lightGBM	80.17	55.67	55.18	76.83	60.78	61.43	61.17	62.79
XGB	**81.08**	55.10	55.89	77.36	61.50	62.60	61.32	63.40

Regression Tasks Evaluation (1/2)

		RMSE				
Algorithms	**Features Used**	**OPN**	**NEU**	**EXT**	**AGR**	**CON**
Linear Regression (Ridge/Elastic Net)	**Oxford**	0.634	**0.769**	0.819	0.669	0.726
	Text	0.626	0.776	0.778	0.651	0.707
	Page Likes	-	-	-	-	-
	N2v	0.596	0.7782	0.7697	0.6492	0.703
Linear Regression Lasso (L1)	**Oxford**	0.634	0.769	0.819	0.669	0.726
	Text	0.625	0.775	0.778	0.648	0.705
	Page Likes	-	-	-	-	-
	N2v	**0.5963**	0.7787	**0.7702**	**0.6479**	**0.7029**

Université de Montréal

Regression Tasks Evaluation (2/2)

		RMSE				
Algorithms	**Features Used**	**OPN**	**NEU**	**EXT**	**AGR**	**CON**
Random Forests	**Oxford**	0.638	0.782	0.831	0.673	0.736
	Text	0.627	0.780	0.779	0.653	0.710
	Page Likes	0.621	0.778	0.80	**0.64**	0.71
	N2v	0.6082	0.7866	0.7889	0.6527	0.713
XGB Regressor	**Oxford**	0.631	**0.768**	0.823	0.665	0.728
	Text	0.623	0.778	**0.778**	0.648	**0.701**
	Page Likes	0.621	0.773	0.791	0.65	0.712
	N2v	**0.609**	0.781	0.787	0.652	0.708

Approach 1

What worked - In a nutshell

Best Results

Gender Classification
Oxford Features using XGB

Age Classification
LIWC Features using XGB

Personality Prediction
LIWC Features using XGB

Approach 2

Multi Modal Fusion

Stack data sources for each task to find the best algorithm

Model Agnostic Approaches

Model-free approaches

- Early Fusion
- Late Fusion
- Hybrid Fusion

Multi Modal Fusion

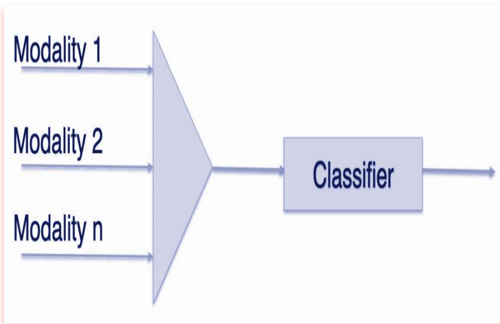
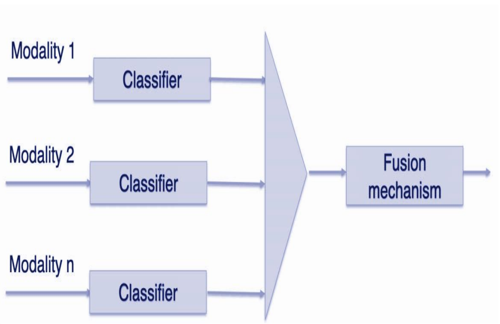
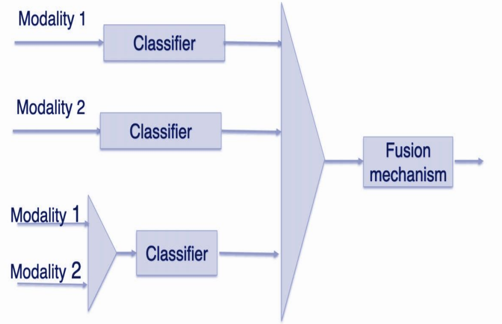
Early Fusion	Late Fusion	Hybrid Fusion
		
● In this approach, we implement feature concatenation. ● It helps exploit feature dependencies.	● In this approach, we train both unimodal and multimodal fusion predictors. ● Fusion mechanism tried was voting.	● In this approach, we combine both early and late fusion mechanisms.

Image: https://ift6758.github.io/lectures/multimodal_learning.pdf

Classification Tasks Evaluation

	Classification Accuracy							
	Gender				Âge			
Features Used	Oxford + Text EARLY FUSION	Oxford + N2V EARLY FUSION	Oxford + Text + N2V EARLY FUSION	Oxford + Text + Occurrence Matrix LATE FUSION	Oxford + Text EARLY FUSION	Oxford + N2V EARLY FUSION	Oxford + Text + N2V EARLY FUSION	Oxford + Text + Occurrence Matrix LATE FUSION
Random Forests	83.11	84.368	84.526	82.13	61.947	61.947	62.684	61.11
Linear SVM	79.79	85.526	**85.684**	84.123	61.526	**71.368**	70.842	70.66
lightGBM	82.89	84.00	85.105	83.87	62.632	65.632	67.158	68.19
XGB	83.74	84.94	86.316	85.178	63.316	67.474	68.632	68.66

Regression Tasks Evaluation (1/2)

Algorithms	Features Used	Fusion Techniques	OPN	NEU	EXT	AGR	CON
Linear Regression	Oxford + Text	EARLY FUSION	0.633	0.790	0.814	0.659	0.708
	Oxford + N2V	EARLY FUSION	0.602	0.780	0.779	0.642	0.704
	Oxford + Co-Occurrence Matrix	LATE FUSION	0.612	0.782	0.789	0.652	0.709
	Oxford +Text +N2V	EARLY FUSION	0.613	0.780	0.780	0.645	0.707
	Oxford + Text + CO-Occurrence Matrix	LATE FUSION	0.607	0.788	0.781	0.648	0.705
Lasso (L1)	Oxford + Text	EARLY FUSION	0.631	0.791	0.804	0.656	0.716
	Oxford + N2V	EARLY FUSION	0.600	0.780	0.769	0.640	0.700
	Oxford +Text +N2V	EARLY FUSION	0.599	0.779	0.764	0.638	0.697
	Oxford + Text + CO-Occurrence Matrix	LATE FUSION	0.605	0.782	0.785	0.644	0.709

Regression Tasks Evaluation (2/2)

			RMSE				
Algorithms	Features Used	Fusion Technique	OPN	NEU	EXT	AGR	CON
Random Forests	Oxford + N2V	EARLY FUSION	0.611	0.785	0.783	0.649	0.710
	Oxford + Text + N2V	EARLY FUSION	0.609	0.689	0.779	0.644	0.703
XGB Regressor	Oxford + Text	EARLY FUSION	0.618	0.785	0.792	0.650	0.701
	Oxford + N2V	EARLY FUSION	0.611	0.786	0.779	0.646	0.707
	Oxford + Text + N2V	EARLY FUSION	0.609	0.777	0.7877	0.641	0.702

Best Results

Gender Classification
Oxford Features + Node2Vec Relation Features using SVM

Age Classification
Oxford Features + Text + Node2Vec Relation Features using SVM

Personality Prediction
Did not improve results with multiple data sources

Approach 3

Multi Modal Fusion

Stacking predictions as input to predict other tasks

Using predictions as features

Using previous predictions like gender, age as input to predict other tasks.

X	y1
x1	0
x2	1
x3	0

Classifier 1

X	y1	y2
x1	0	1
x2	1	0
x3	0	1

Classifier 2

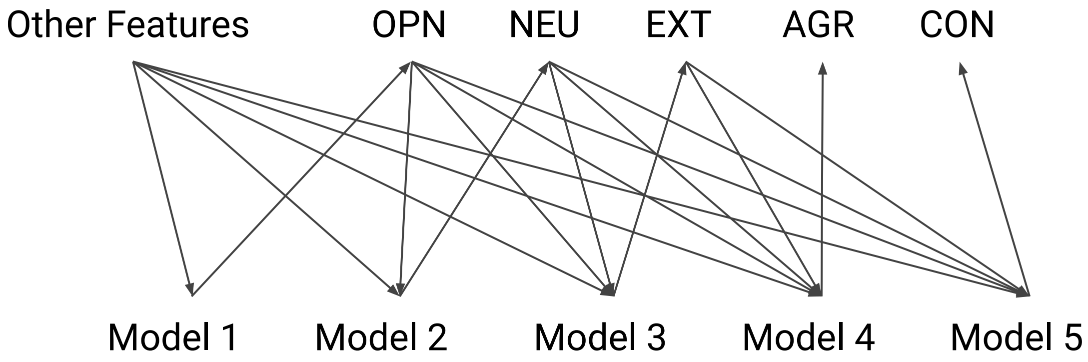
X	y1	y2	y3
x1	0	1	1
x2	1	0	0
x3	0	1	0

Classifier 3

X	y1	y2	y3	y4
x1	0	1	1	0
x2	1	0	0	0
x3	0	1	0	0

Classifier 4

Regressor Chaining



What Next? Approach 4

Combining all the approaches

Stack models built on individual data source, multiple data sources and train using classifier/regression chaining at all levels.