

Myers-Briggs Personality Type Prediction with Text Classification

Raehyun Kim
rhkim79@berkeley.edu

Tianhao Wu
thwu@berkeley.edu

Shiping Xu
sxu2@berkeley.edu

Introduction

- Personality is predictive of many consequential outcomes
- **Problem:** Traditional personality assessments are time-consuming and inefficient
- **Solution:** Build a classifier to automatically predict personality types based on social media posts
 - Applications in various fields such as recruitment, online marketing, etc.
 - Allow more people gain access to their personality types in a faster and more reliable way

Dataset

- Kaggle dataset, crawled from Personality Cafe forum
- **Input:** last 50 posts made by 8,675 forum users
- **Output:** User's Myers-Briggs (MBTI)

Personality Type

- Extroverts (E) vs. Introverts (I)
- Sensors (S) vs. Intuitives (N)
- Thinkers (T) vs. Feelers (F)
- Judgers (J) vs. Perceivers (P)

PERSONALITY TYPES KEY



Extroverts

are energized by people, enjoy a variety of tasks, a quick pace, and are good at multitasking.



Introverts

often like working alone or in small groups, prefer a more deliberate pace, and like to focus on one task at a time.



Sensors

are realistic people who like to focus on the facts and details, and apply common sense and past experience to come up with practical solutions to problems.



Intuitives

prefer to focus on possibilities and the big picture, easily see patterns, value innovation, and seek creative solutions to problems.



Thinkers

tend to make decisions using logical analysis, objectively weigh pros and cons, and value honesty, consistency, and fairness.



Judgers

tend to be organized and prepared, like to make and stick to plans, and are comfortable following most rules.



Feelers

tend to be sensitive and cooperative, and decide based on their own personal values and how others will be affected by their actions.

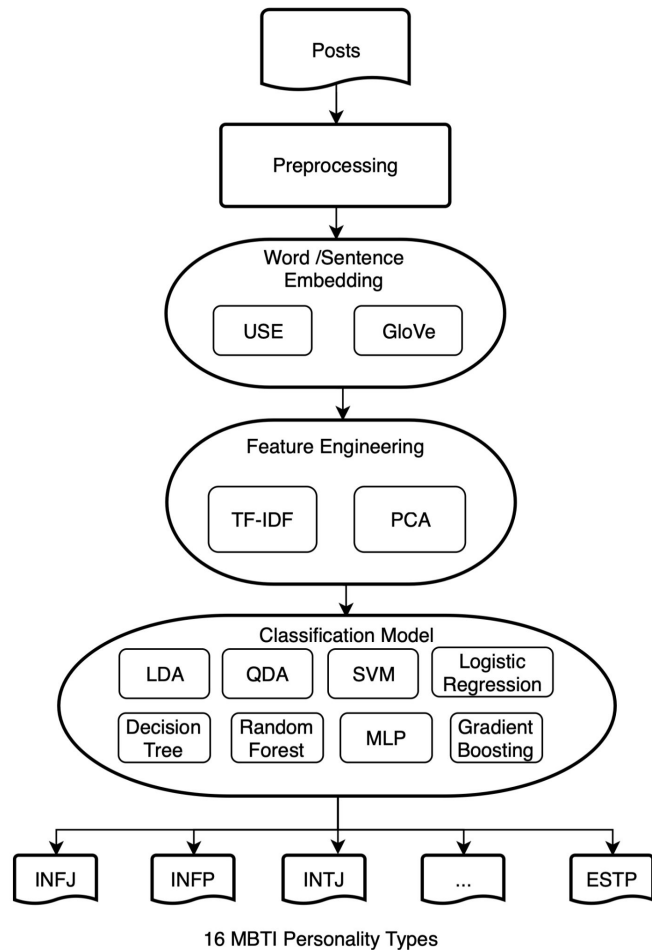


Perceivers

prefer to keep their options open, like to be able to act spontaneously, and like to be flexible with making plans.

Approach

- Preprocessing
- Word/Sentence embeddings
 - TF-IDF Vectorizer
 - Additional features
 - Number of words
 - Number of urls
 - Number of question marks
 - Whitening
 - Linear dimensionality reduction (PCA)



Models

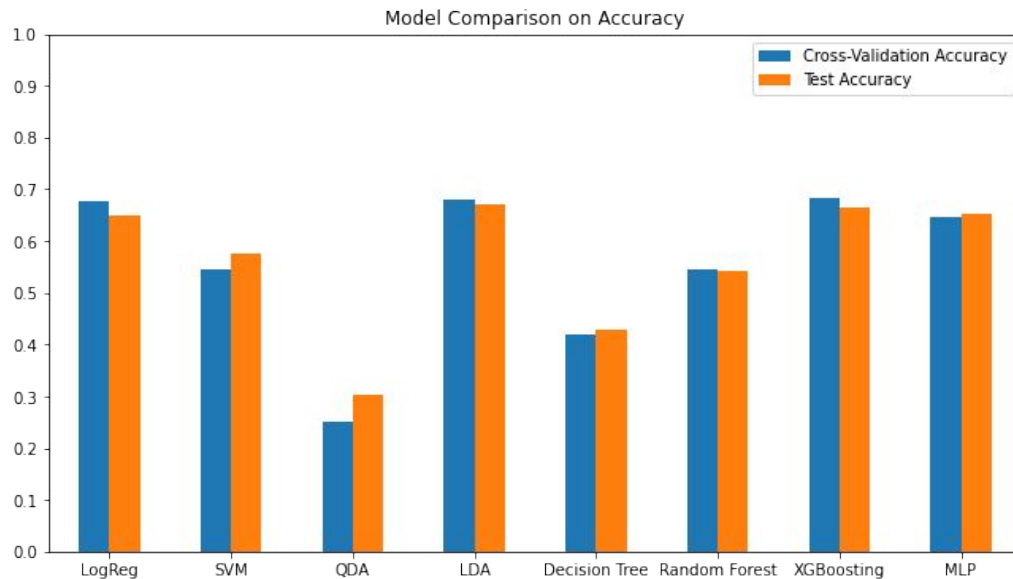
- Train 8 different models for the classification task
- Initial training step on the baseline datasets and further hyper-parameter tuning

Model	Baseline Dataset	#PC	Hyperparameter	Cross Validation Accuracy	Test Accuracy
Logistic Regression	3	300	C=0.8; penalty='l1', solver='liblinear'	0.678	0.65
SVM	3	200	C=0.14; kernel='rbf'	0.545	0.577
QDA	3	500	N/A	0.250	0.304
LDA	1	38	N/A	0.680	0.672
Decision Tree	3	500	max_depth=10	0.420	0.430
Random Forest	3	500	max_depth=20; n_estimators=150; min_sample_leaves=2	0.545	0.541
XGBoost	1	N/A	max_depth=2; n_estimators=200	0.684	0.666
MLP	3	500	alpha=0.05; hidden_layer_sizes=(20,)	0.647	0.651

Table 1: Best Models Parameters

Evaluation & Analysis

- Results Visualization
 - 5 folds cross-validation accuracy
 - Test accuracy
- Best model: LDA
 - Test Accuracy: 0.672
- Worst model: QDA



Conclusion

- High-level Takeaways
 - Myers-Briggs types prediction is feasible (0.67 test accuracy)
 - Unwhitened 2000 TF-IDF features + 3 additional features
- Future Directions
 - Feature selection
 - Include more additional features
 - backward elimination, forward selection
 - Predict based on different datasets
 - Reddit, Facebook, etc.