Predicting MBTI* Personality Types

Serdar BOZOGLAN

MS Operations Research

Data Science Career Track Capstone Project March-18 Cohort

Springboard

github.com/serdarbozoglan

Mentor

Tuhin Sharma

***MBTI:** The Myers Briggs Type Indicator (or MBTI for short) is a personality type system that divides everyone into 16 distinct personality types

Problem Definition

8675 posts are tagged/labelled by "PersonalityCafe forum". We will predict personalities from the posts issued in social media.

16 Different personality types

Who Might Care?

Relationship

Careers

Education

Workplace

And many more…

Data Information

Data is obtained from Kaggle

PersonalityCafe Forum

8675 posts

16 different personalties (classes)

Data Cleansing

Lowercase, Removing links, numbers, stop words, punctuations

	type	posts
0	INFJ	'http://www.youtube.com/watch?v=qsXHcwe3krw|||http://41.media.tumblr.com/tumblr_lfouy03PMA1qa1rooo1_500.jpg|||enfp and intj moments https://www.youtube.com/watch?v=iz7lE1g4XM4 sportscenter not top ten plays https://www.youtube.com/watch?v=uCdfze1etec pranks|||What has been the most life-changing experience in your life?|||http://www.youtube.com/watch?v=vXZeYwwRDw8 http://www.youtube.com/watch?v=u8ejam5DP3E On repeat for most of today.|||May the PerC Experience immerse you.|||The last thing my INFJ friend posted on his facebook before committing suicide the next day. Rest in peace~ http://vimeo.com/22842206|||Hello ENFJ7. Sorry to hear of your distress. It's only natural for a relationship to not be perfection all the time in every moment of existence. Try to figure the hard times as times of growth, as...|||84389 84390 http://wallpaperpassion.com/upload/23700/friendship-boy-and-girl-wallpaper.jpg http://assets.dornob.com/wp-content/uploads/2010/04/round-home-design.jpg ...|||Welcome and stuff.|||http://playeressence.com/wp-content/uploads/2013/08/RED-red-the-pokemon-master-32560474-450-338.jpg Game. Set. Match.|||Prozac, wellbrutin, at least thirty minutes of moving your legs (and I don't mean moving them while sitting in your same desk chair), weed in moderation (maybe try edibles as a healthier alternative...|||Basically come up with three items you've determined that each type (or whichever types you want to do) would more than likely use, given each types...

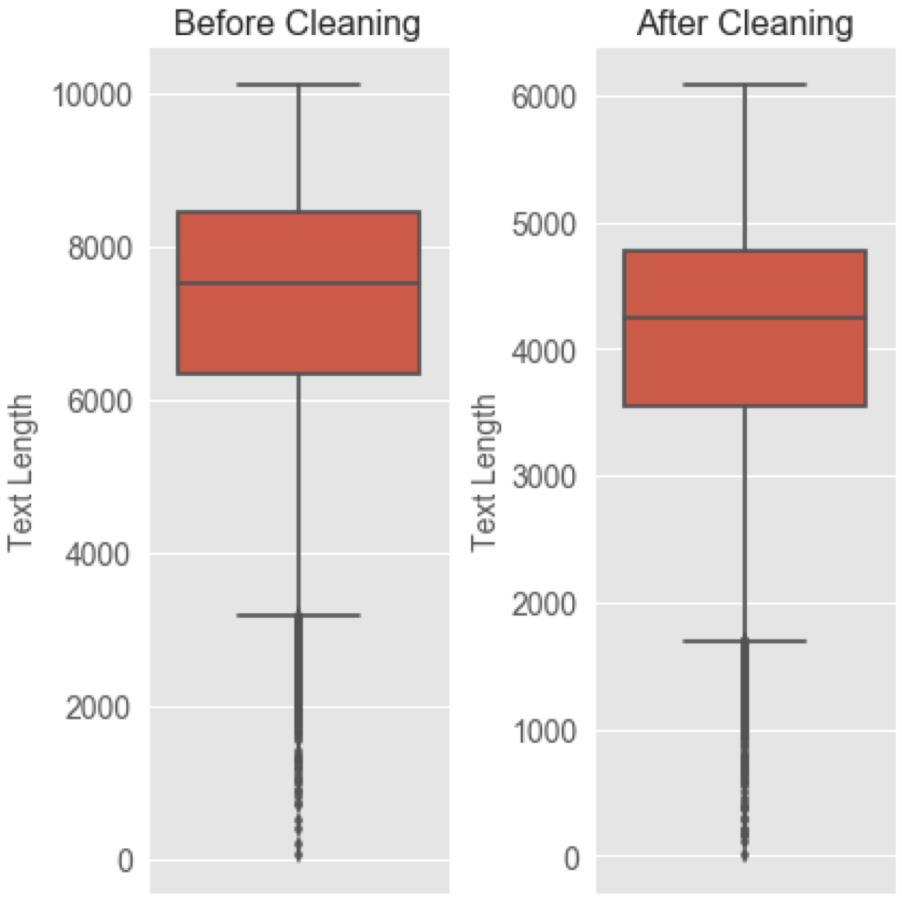
Before

```
0       enfp intj moment    sportscent not top ten play    prank life chang experi life       repeat today  may perc experi immers  last thing infj frien
d post facebook commit suicid next day  rest peac     hello enfj   sorri hear distress  natur relationship not perfect time everi moment exist  tri figur
hard time time growth                          welcom stuff    game  set  match  prozac  wellbrutin  least thirti minut move leg  not mean move sit desk
chair   weed moder  mayb tri edibl healthier altern    basic come three item determin type  whichev type want  would like use  given type  cognit functio
n whatnot  left    thing moder   sim inde video game  good one  note  good one somewhat subject not complet promot death given sim    dear enfp   favorit
video game grow  current favorit video game   cool   appear late   sad  someon everyon  wait    thought confid good thing  cherish time solitud b c revel
within inner world wherea time would workin    enjoy time  not worri  peopl alway around    yo entp ladi    complimentari person well  hey      main soci
al outlet xbox live convers even verbal fatigu quickli    realli dig part           ban thread requir  get high backyard  roast eat marshmellow backyard
convers someth intellectu  follow massag kiss     ban mani b sentenc  could  think b  ban watch movi corner dunc  ban health class clearli taught noth pe
er pressur  ban whole host reason      two babi deer left right munch beetl middl      use blood  two cavemen diari toda...
Name: clean_text, dtype: object
```

After

Data Cleansing

Average Text Length Before Cleaning and After Cleaning

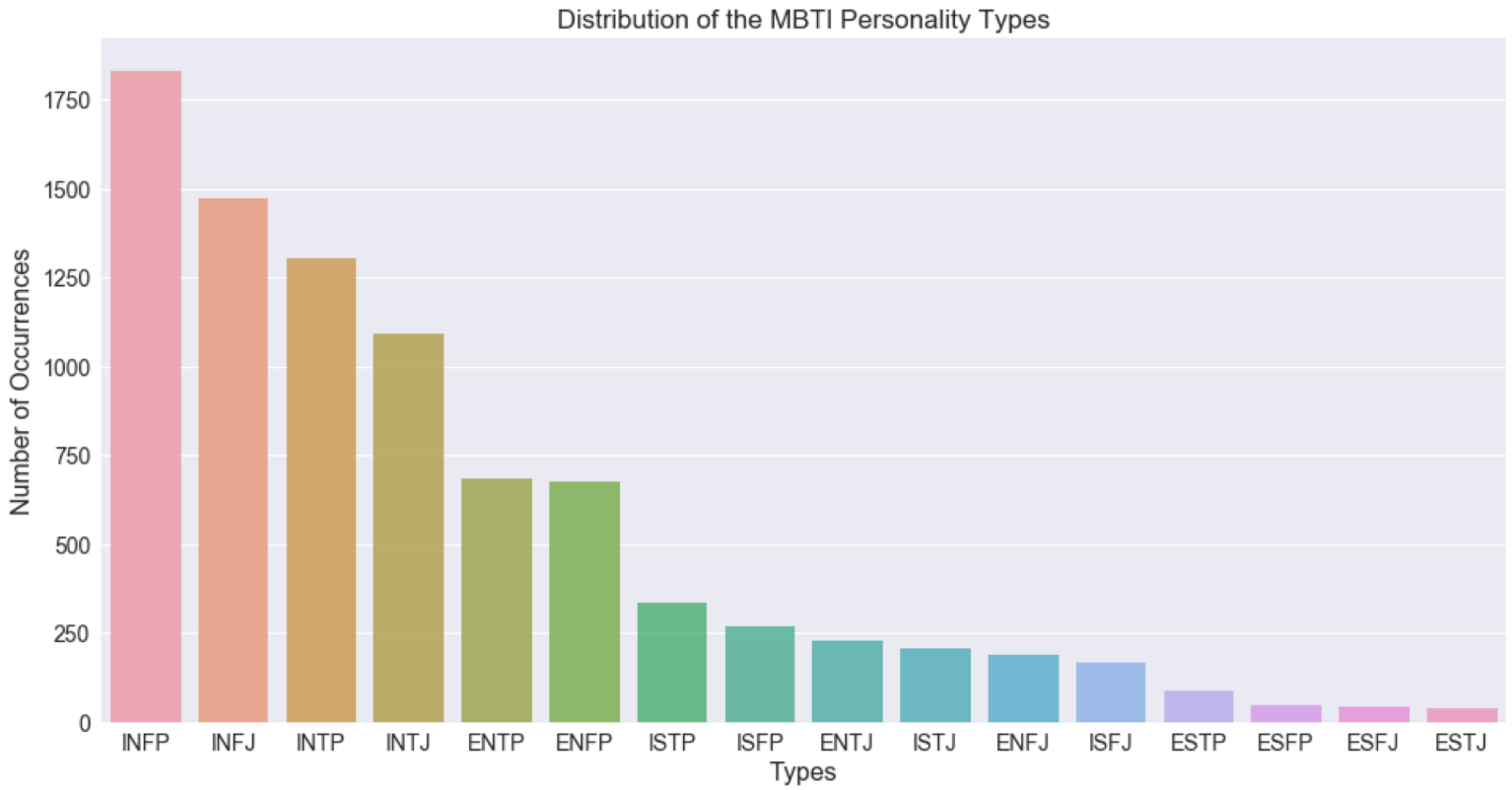


~40% reduction

Data Exploration Analysis

Distribution of MBTI Personality Types

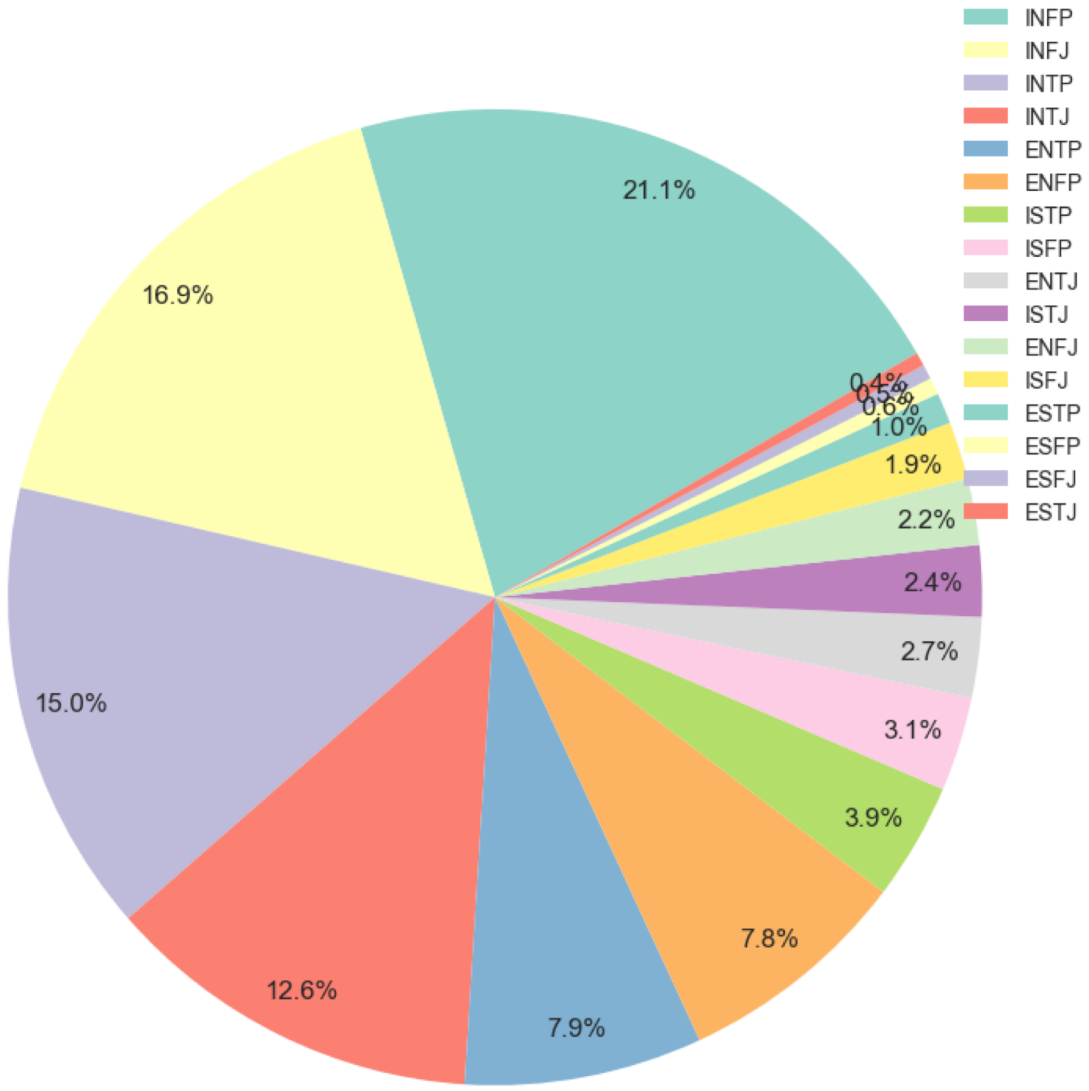
Most common type INFP, while least common one ESTJ



Data Exploration Analysis

Distribution of MBTI Personality Types

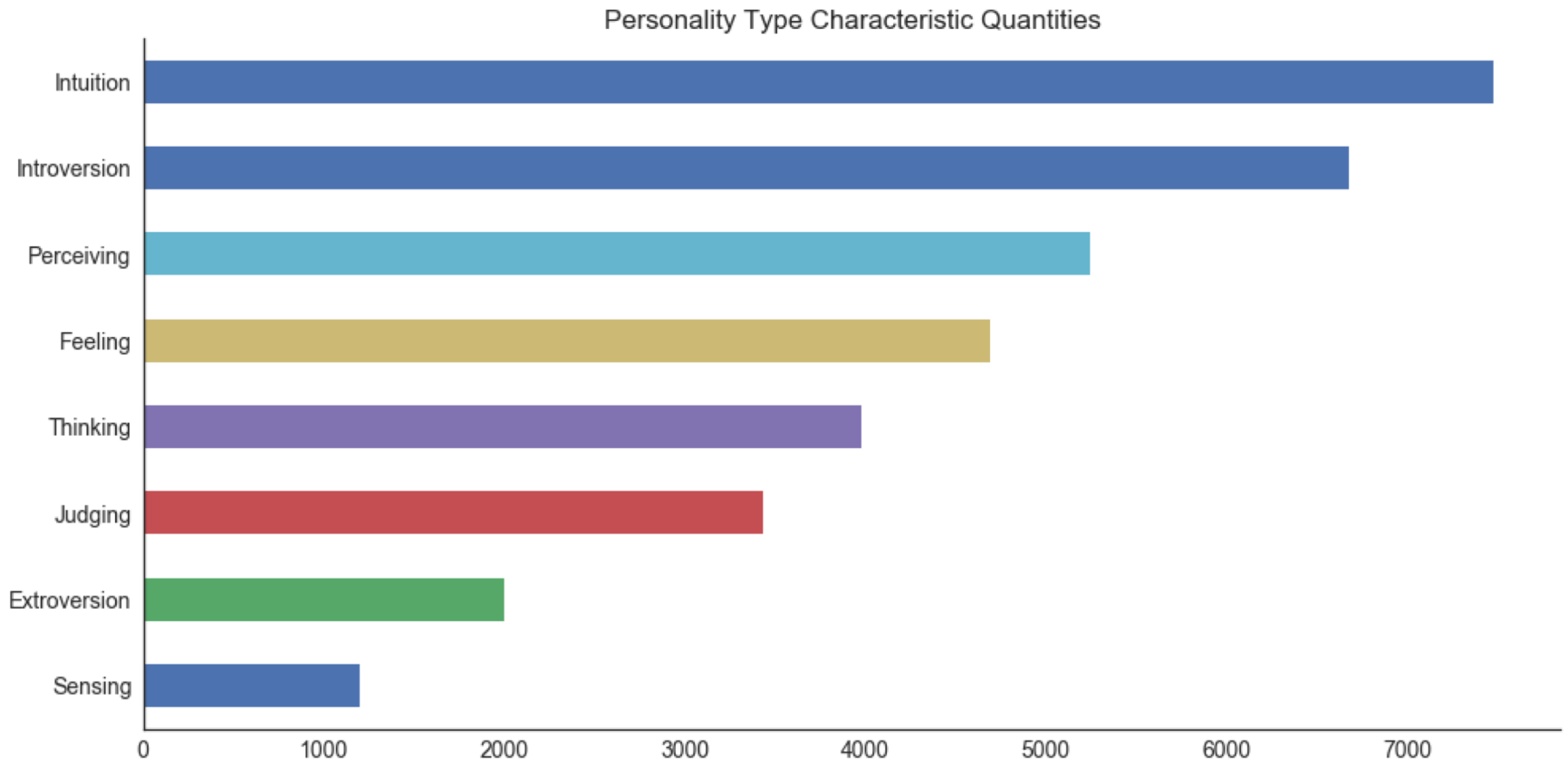
Most common type INFP, while least common one ESTJ



Data Exploration Analysis

MBTI Personality Types Qunatities

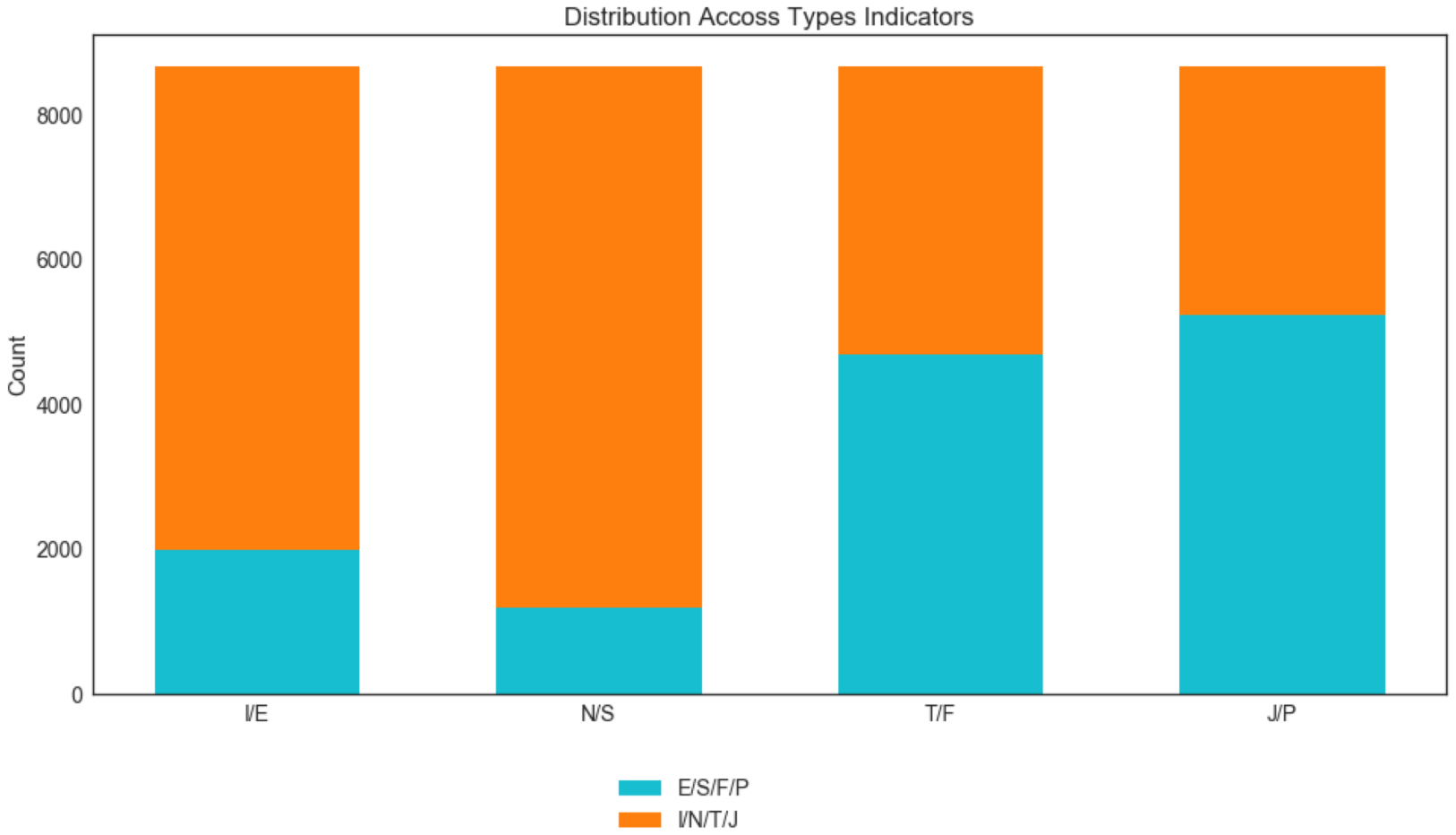
Most common type Intuition, while least common one Sensing



Data Exploration Analysis

Distribution Accross Personality Types

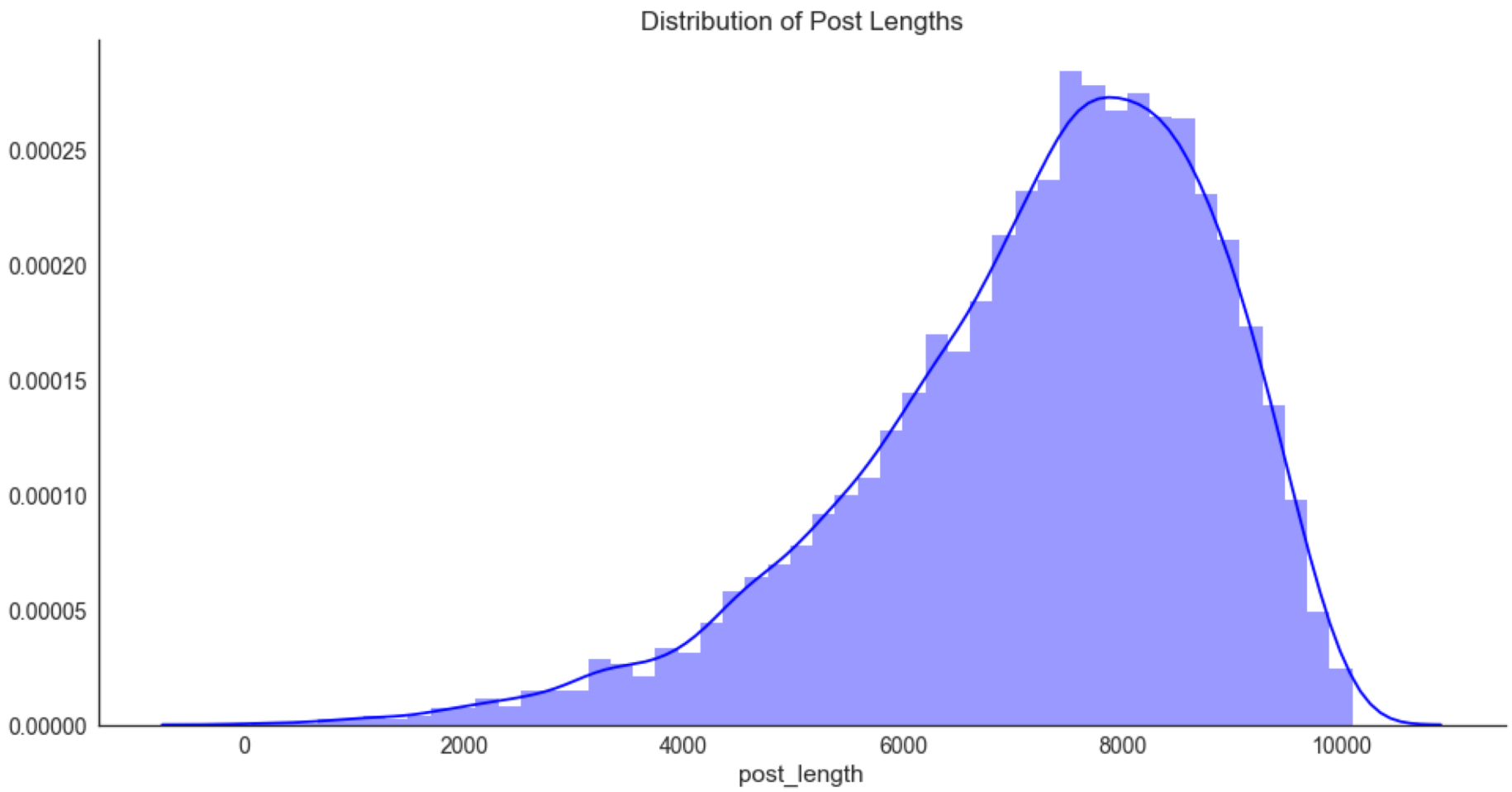
Intuition vs Sensing, Introversion vs Extroversion imbalanced



Data Exploration Analysis

Distribution of Post Lentghs

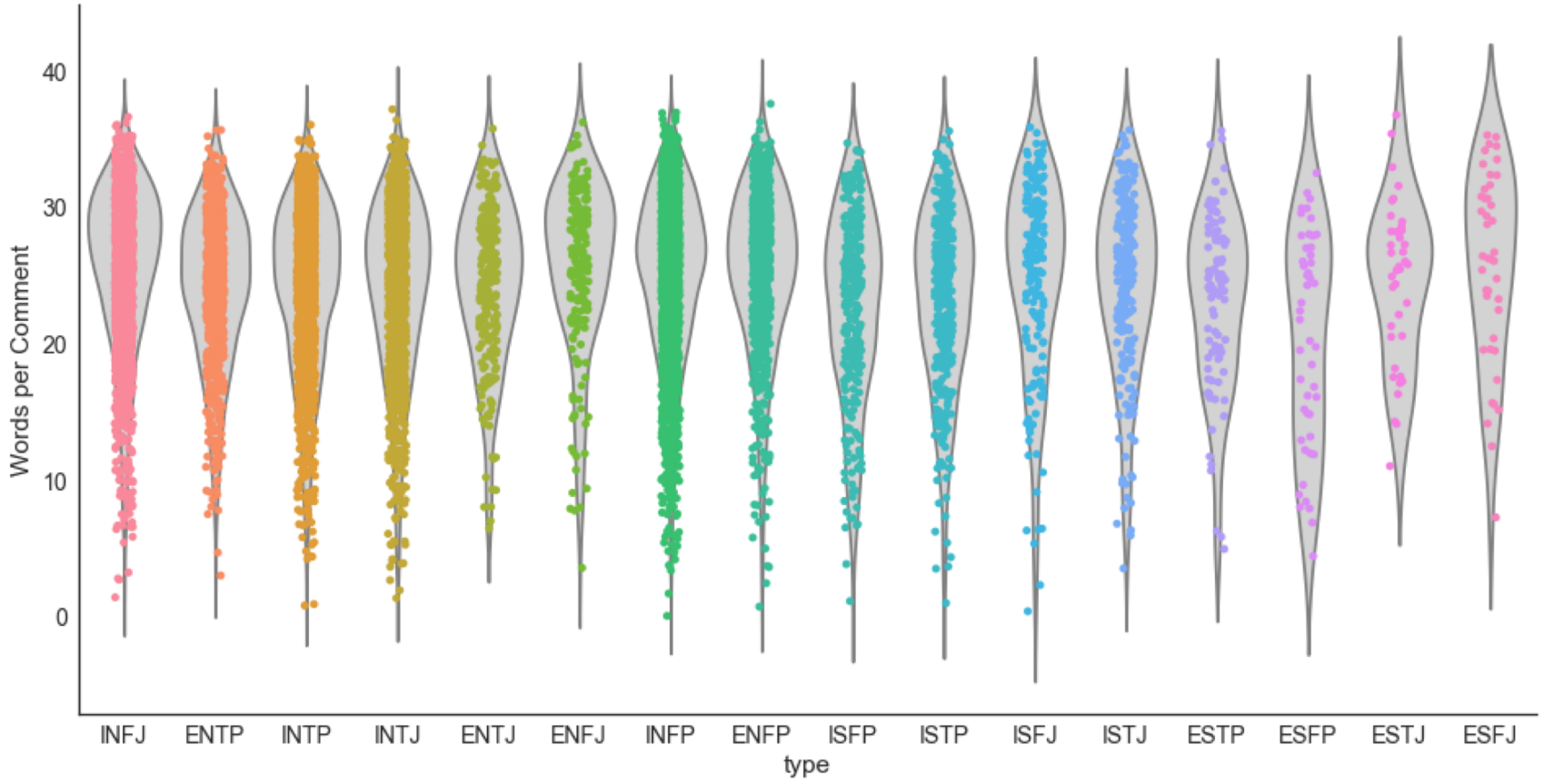
Seems slightly skewed Gaussian Distribution



Data Exploration Analysis

Words in per comment

Average word number in comment is around 20



Machine Learning Models

Overview

Type : Supervised Learning

Multiclass Classification

Tools : Scikit Learn

Models

Logistic Regression w Count Vectorizer

Logistic Regression w Tf-Idf Vectorizer

Linear SVM w Count Vectorizer

Linear SVM w Tf-Idf Vectorizer

XGBOOST w Count Vectorizer

XGBOOST w Tf-Idf Vectorizer

Part-1 (16 classes with MBTI labels)

Result

Linear SVM with Count Vectorizer

	precision	recall	f1-score	support
ENFJ	0.43	0.68	0.53	38
ENFP	0.58	0.63	0.60	148
ENTJ	0.49	0.57	0.53	61
ENTP	0.78	0.68	0.73	187
ESFJ	0.21	0.36	0.27	11
ESFP	0.21	0.31	0.25	13
ESTJ	0.35	0.60	0.44	10
ESTP	0.34	0.65	0.45	20
INFJ	0.82	0.62	0.71	357
INFP	0.73	0.78	0.76	464
INTJ	0.71	0.69	0.70	277
INTP	0.77	0.75	0.76	320
ISFJ	0.60	0.71	0.65	38
ISFP	0.74	0.54	0.62	74
ISTJ	0.58	0.66	0.62	56
ISTP	0.65	0.73	0.69	95
avg / total	0.71	0.69	0.70	2169

Accuracy : 0.691101890272

Part-2 (16 classes without MBTI labels)

Result

Logistic Regression with Tf-Idf Vectorizer

	precision	recall	f1-score	support
ENFJ	0.10	0.11	0.10	38
ENFP	0.22	0.30	0.26	148
ENTJ	0.16	0.11	0.13	61
ENTP	0.33	0.34	0.33	187
ESFJ	0.00	0.00	0.00	11
ESFP	0.00	0.00	0.00	13
ESTJ	0.00	0.00	0.00	10
ESTP	0.20	0.10	0.13	20
INFJ	0.42	0.42	0.42	357
INFP	0.51	0.51	0.51	464
INTJ	0.33	0.33	0.33	277
INTP	0.42	0.48	0.45	320
ISFJ	0.22	0.24	0.23	38
ISFP	0.11	0.08	0.09	74
ISTJ	0.39	0.21	0.28	56
ISTP	0.32	0.29	0.31	95
avg / total	0.36	0.37	0.37	2169

Accuracy : 0.371138773628

Part-3 (with MBTI labels)

Introversion vs Extroversion

XGBOOST with Tf-Idf Vectorizer

	precision	recall	f1-score	support
0	0.77	0.53	0.63	488
1	0.88	0.95	0.91	1681
avg / total	0.85	0.86	0.85	2169

Accuracy : 0.86030428769

Part-3 (with MBTI labels)

Intuition vs Sensing

XGBOOST with Tf-Idf Vectorizer

	precision	recall	f1-score	support
0	0.81	0.38	0.51	317
1	0.90	0.98	0.94	1852
avg / total	0.89	0.90	0.88	2169

Accuracy : 0.895804518211

Part-3 (with MBTI labels)

Thinking vs Feeling

Logistic Regression with Tf-Idf Vectorizer

	precision	recall	f1-score	support
0	0.89	0.83	0.86	1143
1	0.83	0.88	0.85	1026
avg / total	0.86	0.86	0.86	2169

Accuracy : 0.857538035961

Part-3 (with MBTI labels)

Judging vs Perceiving

XGBOOST with Count Vectorizer

	precision	recall	f1-score	support
0	0.83	0.88	0.85	1321
1	0.80	0.71	0.75	848
avg / total	0.81	0.82	0.81	2169

Accuracy : 0.815122176118

Part-4 (without MBTI labels)

Introversion vs Extroversion

XGBOOST with Count Vectorizer

	precision	recall	f1-score	support
0	0.57	0.06	0.11	488
1	0.78	0.99	0.87	1681
avg / total	0.74	0.78	0.70	2169

Accuracy : 0.778699861687

Part-4 (without MBTI labels)

Intuition vs Sensing

XGBOOST with Tf-Idf Vectorizer

	precision	recall	f1-score	support
0	0.80	0.01	0.02	317
1	0.86	1.00	0.92	1852
avg / total	0.85	0.86	0.79	2169

Accuracy : 0.855232826187

Part-4 (without MBTI labels)

Thinking vs Feeling

Linear SVM with Tf-Idf Vectorizer

	precision	recall	f1-score	support
0	0.60	0.93	0.73	1143
1	0.81	0.31	0.45	1026
avg / total	0.70	0.64	0.60	2169

Part-4 (without MBTI labels)

Judging vs Perceiving

XGBOOST with Count Vectorizer

	precision	recall	f1-score	support
0	0.68	0.85	0.75	1321
1	0.61	0.37	0.46	848
avg / total	0.65	0.66	0.64	2169

Accuracy : 0.662056247118

Conclusion

In Part-1, we predicted the personality type considering 16 classes while MBTI labels are in the posts. We obtained approximately 69% accuracy with MBTI labels through our best model. It seems a quite good result for 16 classes classification problem.

In Part-2, we removed the MBTI labels from the posts. As a result, we concluded that MBTI labels are very important features to get higher accuracy for personality type prediction. We can roughly say that those labels affects almost 30% accuracy increment.

Conclusion

In Part-3, we predicted only opposite personality types such as Intoversion vs Extroversion etc. We kept MBTI labels in the posts data set. In this part of the study , we improved the baseline accuracy 9% for Introversion-Extroversion, 4.5% for Intuition-Sensing, 33% for Thinking-Feeling and 21% for Judging-Perceiving.

In the last part, we removed the MBTI labels and predicted opposite personality types. As a result, we improved the baseline accuracy 8.5% for Introversion-Extroversion, 25% for Thinking-Feeling and 6% for Judging-Perceiving. Unfortunately our models for Intuition-Sensing could perform as good as baseline accuracy.

Future Study

1. Some personality types are imbalanced. We used class_weight to overcome this issue but we will use Upsampling techniques such as SMOTE to overcome the issue and get better accuracy results.

2. We will implement Deep Learning models to get better results.

3. We will use word2vec technique in NLP part.

4. We will implement Dask library for parallel processing to decrease run time.

5. After decreasing run time, we will focus on hyperparameter tuning more.

Thank You

- Serdar BOZOGLAN
- Email: zserdarb@hotmail.com
- https://www.linkedin.com/in/serdarbozoglan/
- https://github.com/serdarbozoglan