Predicting Myers-Briggs Personality Types from Social Media Posts

Maria Yarolin

December 2017

Project Goals

Personality refers to individual differences in characteristic patterns of thinking, feeling and behaving. *Personality typologies* classify different types of individuals to reveal and enhance the understanding of their behaviors.[1]

This data science project has two main goals:

1. Use **natural language processing** techniques to analyze text postings on a social forum, and predict the writers' personality type using a **classification model**.
2. Compare personality types based on factors such as the **length** and **sentiment** (e.g., polarity, subjectivity) of their posts.

1) Source: American Psychological Association

Overview of Myers-Briggs Type Indicator (MBTI)

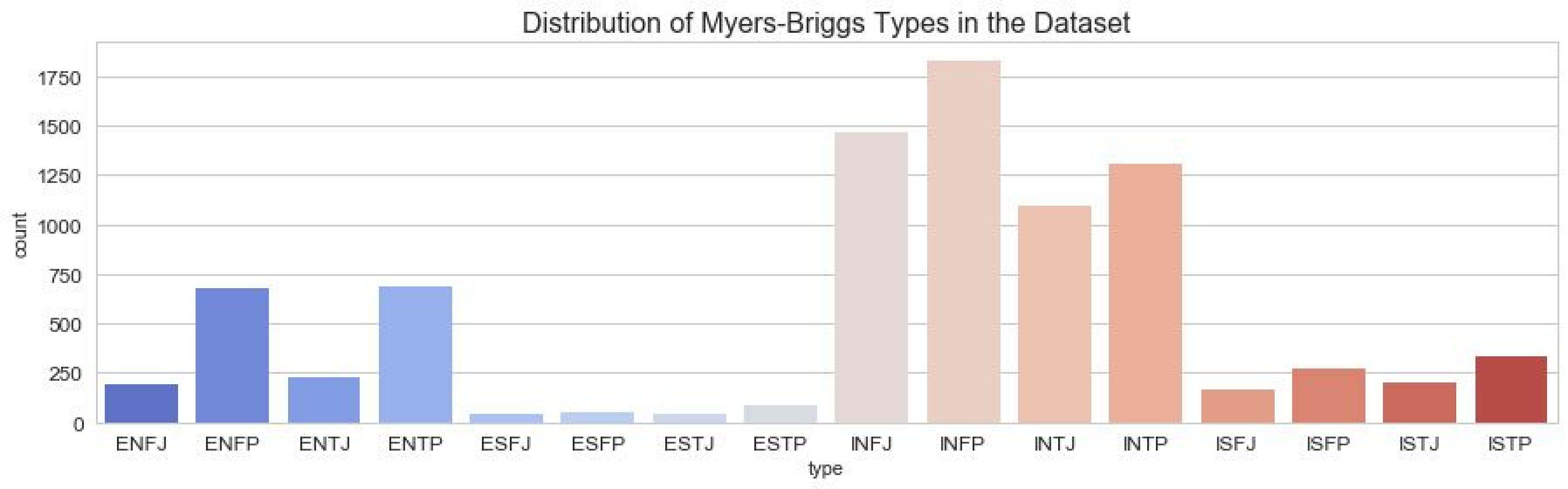
- Psychological assessment tool to classify people into one of 16 different personality types.
- Uses a four-letter code based on four axes, where each letter refers to the predominant trait on each axis continuum.
    - Introversion (I) – Extroversion (E): preference for the “outer” or “inner” world
    - Intuition (N) – Sensing (S): method of processing information
    - Thinking (T) – Feeling (F): method for making decisions
    - Judging (J) – Perceiving (P): orientation to the outer world

Data

- Dataset obtained on Kaggle
- Contains data from 8,675 subjects
- Consists of 2 columns:
  - ***type***: subjects’ MBTI code (16 codes in total)
  - ***posts***: subjects’ 50 most recent posts on *PersonalityCafe*, an online forum focusing on personality types

Data (cont’d)

- Skewed toward subjects from the Introversion-Intuition personality types (IN-)
- Low on Extroversion-Sensing types (ES-)



Project Pipeline

- Step 1: Exploratory data analysis
  - View first several records to see typical formatting of the posts and identify potential issues

	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 ...

Project Pipeline (cont'd)

- Step 2: Preprocessing the dataset
  - Create a new column for each of the four axes
  - Clean the text in the *post* column to prepare for analysis
    - Replace web links with “URL”
    - Remove ||| separators, punctuation, and digits
    - Convert all text to lower case
    - Use **PorterStemmer** to group words having a common stem
    - Define stopwords
    - Use **CountVectorizer** to encode the text
    - Use **TruncatedSVD** to reduce dimensionality

Project Pipeline (cont'd)

- Step 3: Modeling - Classification
  - Random Forest Classifier
  - K-Neighbors Classifier
  - One vs. the Rest Classifier
- Step 4: Sentiment analysis and Word Count
  - Calculate average number of words per post
  - Use **TextBlob** to analyze text and derive the overall sentiment
    - **Polarity** - whether the text is negative, neutral, or positive in tone
    - **Subjectivity** - whether the text is subjective or objective in tone

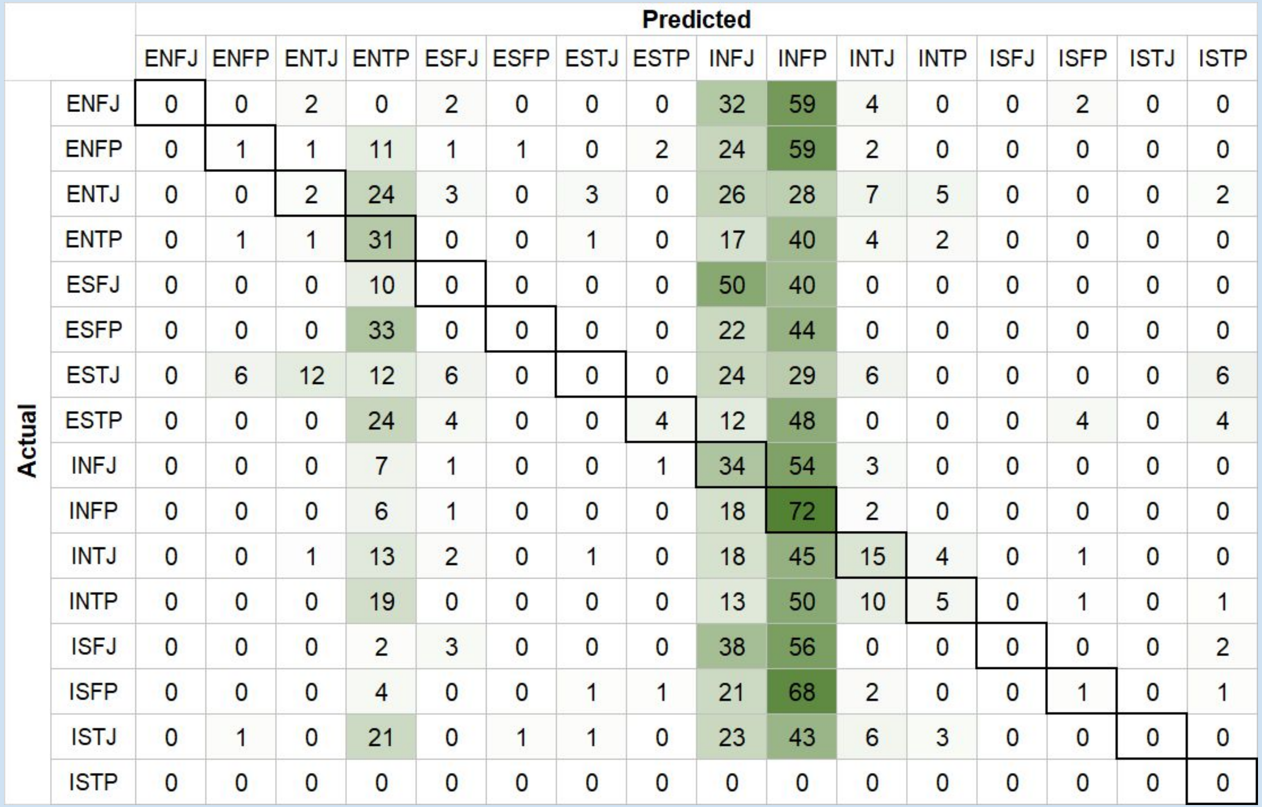
Model Evaluation

- The **baseline accuracy** is **0.211**.

Prediction accuracy of the three models:

- Random Forest Classifier: **0.269**
    - Highest accuracy score
    - All predictions were limited to Introversion-Intuition (IN-) personality types
- K-Neighbors Classifier using 11 neighbors: **0.217**
    - Lowest accuracy score
- One vs. the Rest Classifier: **0.265**
    - Not the highest accuracy score but still greater than the baseline
    - Predictions were better distributed across the classes

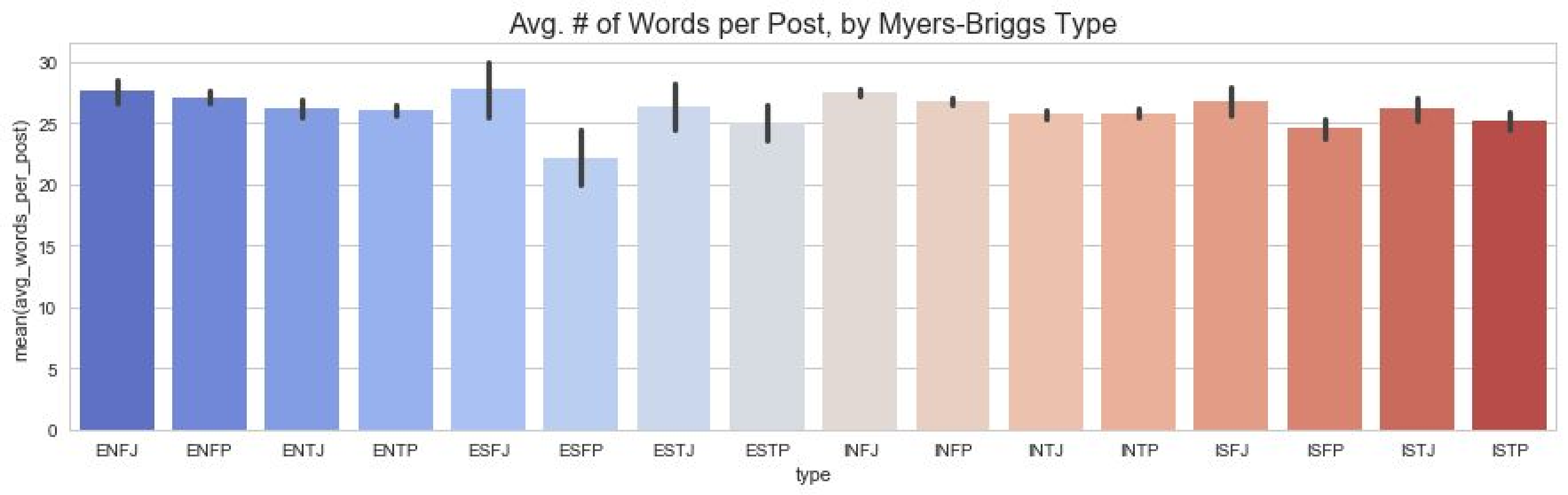
One vs. the Rest Classifier: Prediction Percentages



		Predicted															
		ENFJ	ENFP	ENTJ	ENTP	ESFJ	ESFP	ESTJ	ESTP	INFJ	INFP	INTJ	INTP	ISFJ	ISFP	ISTJ	ISTP
Actual	ENFJ	0	0	2	0	2	0	0	0	32	59	4	0	0	2	0	0
	ENFP	0	1	1	11	1	1	0	2	24	59	2	0	0	0	0	0
	ENTJ	0	0	2	24	3	0	3	0	26	28	7	5	0	0	0	2
	ENTP	0	1	1	31	0	0	1	0	17	40	4	2	0	0	0	0
	ESFJ	0	0	0	10	0	0	0	0	50	40	0	0	0	0	0	0
	ESFP	0	0	0	33	0	0	0	0	22	44	0	0	0	0	0	0
	ESTJ	0	6	12	12	6	0	0	0	24	29	6	0	0	0	0	6
	ESTP	0	0	0	24	4	0	0	4	12	48	0	0	0	4	0	4
	INFJ	0	0	0	7	1	0	0	1	34	54	3	0	0	0	0	0
	INFP	0	0	0	6	1	0	0	0	18	72	2	0	0	0	0	0
	INTJ	0	0	1	13	2	0	1	0	18	45	15	4	0	1	0	0
	INTP	0	0	0	19	0	0	0	0	13	50	10	5	0	1	0	1
	ISFJ	0	0	0	2	3	0	0	0	38	56	0	0	0	0	0	2
	ISFP	0	0	0	4	0	0	1	1	21	68	2	0	0	1	0	1
	ISTJ	0	1	0	21	0	1	1	0	23	43	6	3	0	0	0	0
	ISTP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Word Count Comparison

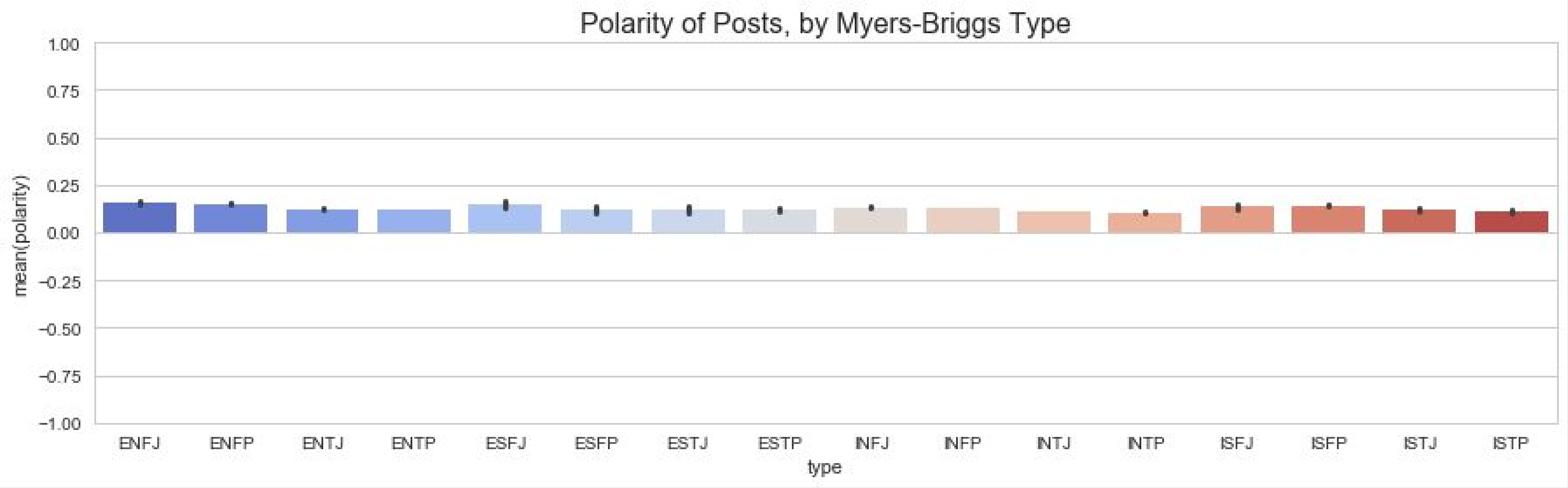
- Relatively little difference in post length among the personality types
- Overall average of 26.4 words per post



Note: Low base sizes (< 100) in the ESFJ, ESFP, ESTJ, and ESTP subgroups

Sentiment Analysis: Polarity

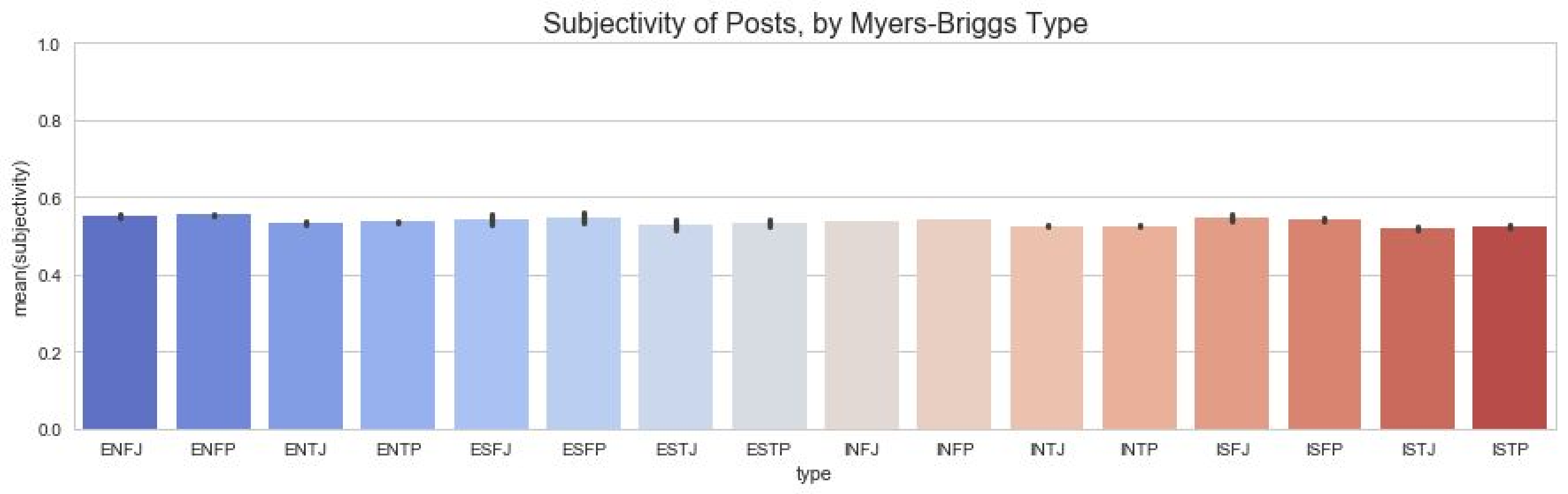
- The types are similar in polarity, with a neutral but positive-leaning tone
- Overall average score is 0.13



Polarity ranges from -1 to 1. Scores closer to -1 are more negative in tone, closer to 0 are more neutral, and closer to 1 are more positive in tone.

Sentiment Analysis: Subjectivity

- Consistent subjectivity among the types, neither too subjective nor objective
- Overall average score is 0.54



Subjectivity ranges from 0 to 1. Scores closer to 0 are more objective in tone, and scores closer to 1 are more subjective in tone.

Learnings

- The model performed best when predicting the most abundant classes in the dataset, but was less reliable on the rarer classes.
- The personality types do not show notable differences in terms of the length, polarity, or subjectivity of their posts.

Next Steps

- Analysis
  - Use modeling to predict classifications on each of the four axes
  - Use tf-idf vectorizer to highlight words with the most discrimination
- Other research approach
  - Collect text data from different social media sites, to analyze subjects’ writing style in different contexts. Personality type can be determined by administering surveys incorporating the MBTI assessment.