Predicting Myers-Briggs Personality Types from Social Media Posts

Maria Yarolin

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)

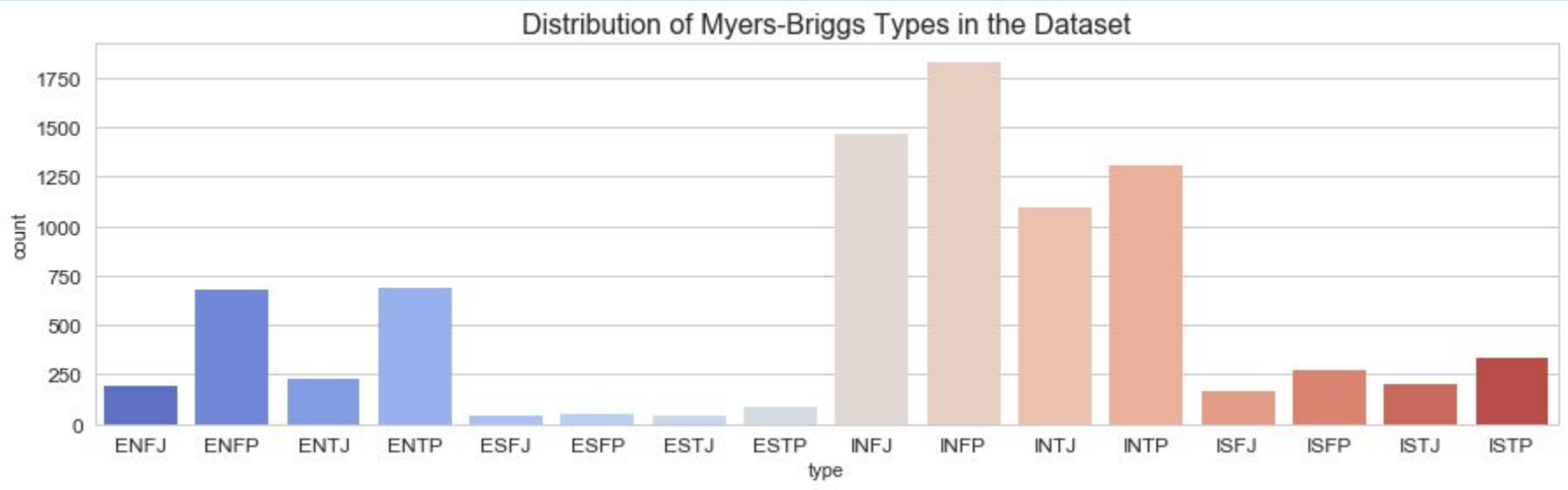
- Psychological assessment tool to classify people into one of 16 different personality types.
- Uses a four-letter code based on four axes, where each letter refers to the predominant trait on each axis continuum.
  - *Introversion (I) – Extroversion (E)*: preference for the “outer” or “inner” world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

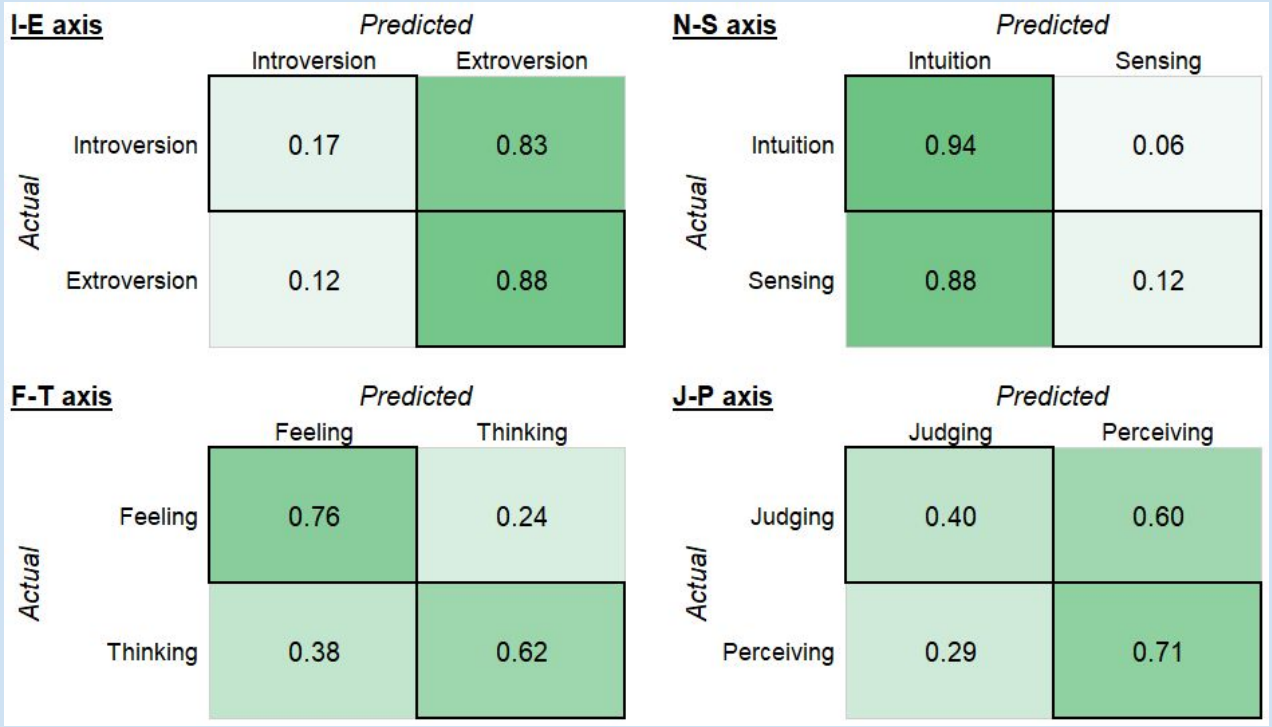
Prediction Percentages: One vs. the Rest Classifier

Actual \ Predicted	ENFJ	ENFP	ENTJ	ENTP	ESFJ	ESFP	ESTJ	ESTP	INFJ	INFP	INTJ	INTP	ISFJ	ISFP	ISTJ	ISTP
ENFJ	0.00	0.00	0.02	0.00	0.02	0.00	0.00	0.00	0.32	0.59	0.04	0.00	0.00	0.02	0.00	0.00
ENFP	0.00	0.01	0.01	0.11	0.01	0.01	0.00	0.02	0.24	0.59	0.02	0.00	0.00	0.00	0.00	0.00
ENTJ	0.00	0.00	0.02	0.24	0.03	0.00	0.03	0.00	0.26	0.28	0.07	0.05	0.00	0.00	0.00	0.02
ENTP	0.00	0.01	0.01	0.31	0.00	0.00	0.01	0.00	0.17	0.40	0.04	0.02	0.00	0.00	0.00	0.00
ESFJ	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.50	0.40	0.00	0.00	0.00	0.00	0.00	0.00
ESFP	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.22	0.44	0.00	0.00	0.00	0.00	0.00	0.00
ESTJ	0.00	0.06	0.12	0.12	0.06	0.00	0.00	0.00	0.24	0.29	0.06	0.00	0.00	0.00	0.00	0.06
ESTP	0.00	0.00	0.00	0.24	0.04	0.00	0.00	0.04	0.12	0.48	0.00	0.00	0.00	0.04	0.00	0.04
INFJ	0.00	0.00	0.00	0.07	0.01	0.00	0.00	0.01	0.34	0.54	0.03	0.00	0.00	0.00	0.00	0.00
INFP	0.00	0.00	0.00	0.06	0.01	0.00	0.00	0.00	0.18	0.72	0.02	0.00	0.00	0.00	0.00	0.00
INTJ	0.00	0.00	0.01	0.13	0.02	0.00	0.01	0.00	0.18	0.45	0.15	0.04	0.00	0.01	0.00	0.00
INTP	0.00	0.00	0.00	0.19	0.00	0.00	0.00	0.00	0.13	0.50	0.10	0.05	0.00	0.01	0.00	0.01
ISFJ	0.00	0.00	0.00	0.02	0.03	0.00	0.00	0.00	0.38	0.56	0.00	0.00	0.00	0.00	0.00	0.02
ISFP	0.00	0.00	0.00	0.04	0.00	0.00	0.01	0.01	0.21	0.68	0.02	0.00	0.00	0.01	0.00	0.01
ISTJ	0.00	0.01	0.00	0.21	0.00	0.01	0.01	0.00	0.23	0.43	0.06	0.03	0.00	0.00	0.00	0.00
ISTP	0.00	0.00	0.01	0.32	0.02	0.00	0.01	0.02	0.17	0.32	0.10	0.03	0.00	0.00	0.00	0.01

The model accurately predicted:
72% of INFP cases
34% of INFJ cases
31% of ENTP cases.

Yet, many other classes were misattributed to the two most abundant classes, INFJ and INFP.

Prediction Percentages: The Four Axes

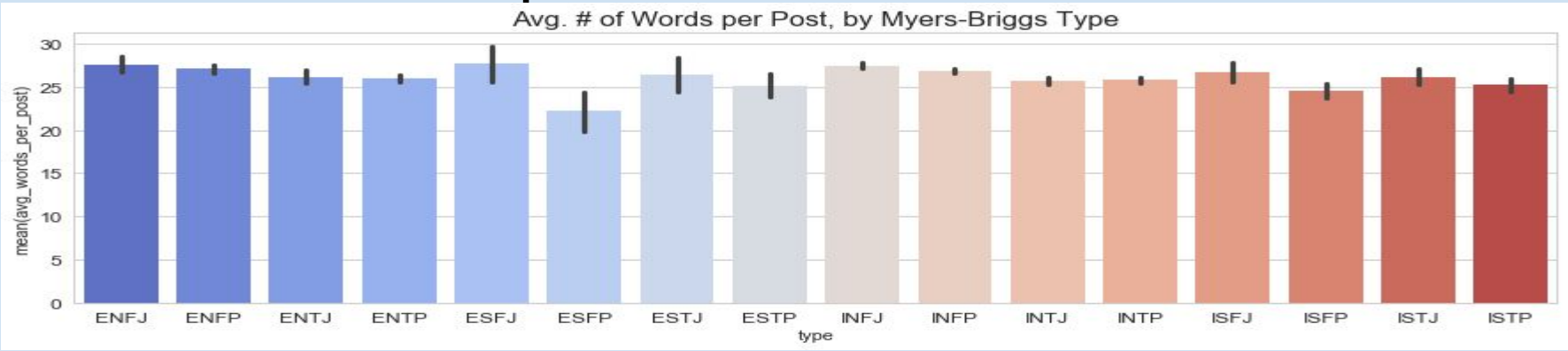


A Random Forest Classifier model was used to predict the four individual axes.

On the *F-T axis*, the model accurately predicted 76% of the *Feeling* cases and 62% of the *Thinking* cases.

Yet, predictions on each the other three axes tended to favor the more abundant case.

Word Count Comparison

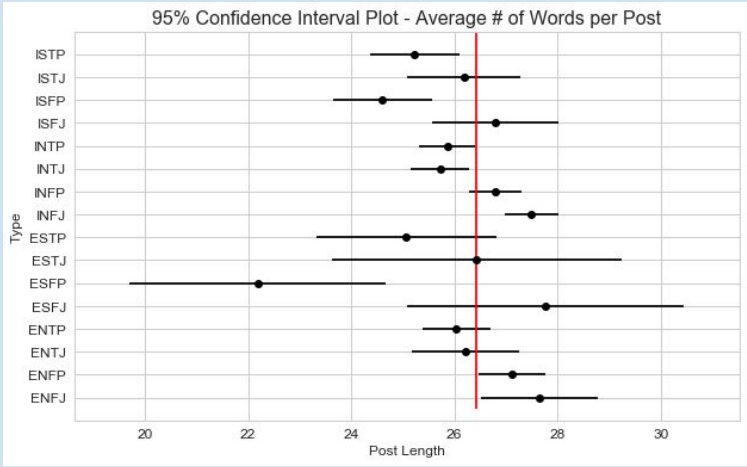


The average post length is **26.4** words per post.

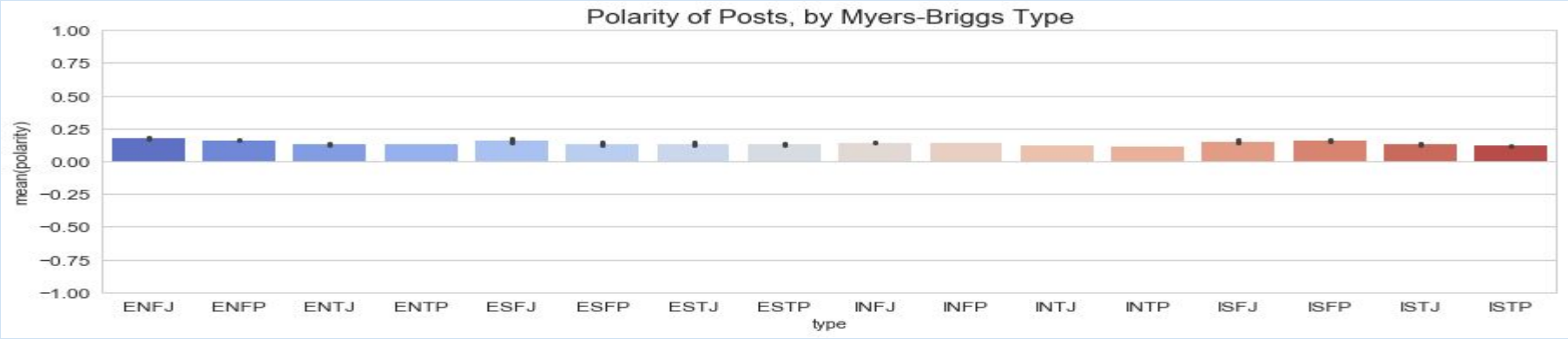
The **Intuitive-Feeling** types tend to use more words per post than those with an *Introversion-Intuitive-Thinking* combination, *Introversion-Sensing-Perception* combination, or type ESFP.

ESFPs tend to use the fewest words per post.

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

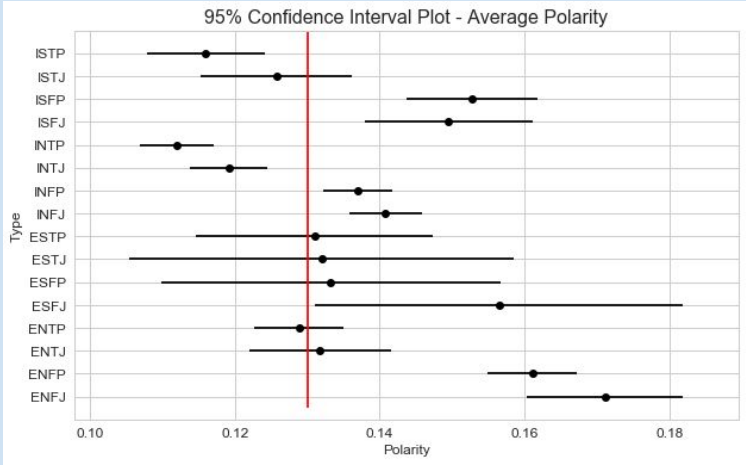


Sentiment Analysis: Polarity



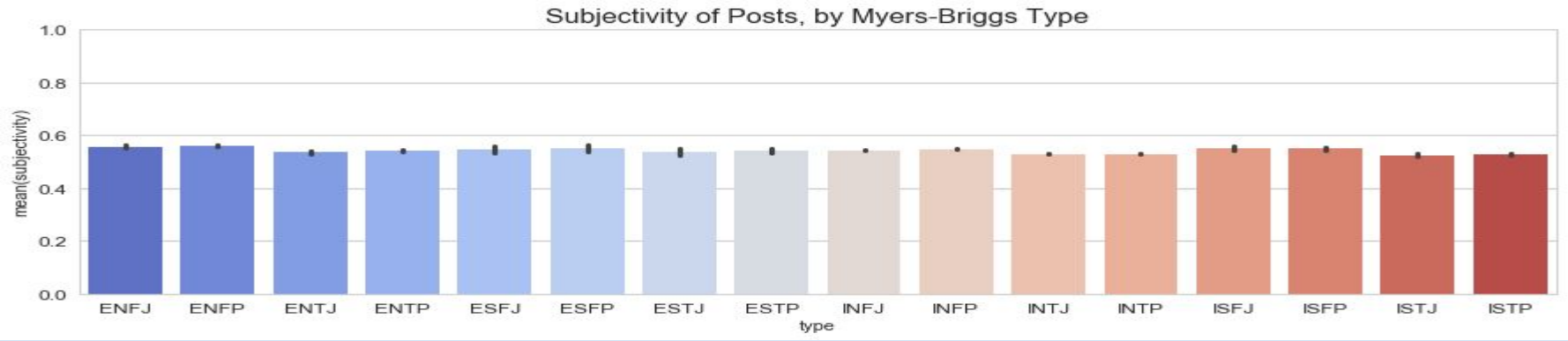
The average polarity of posts is **0.13**.

The posts from **Extroversion-Intuitive-Feeling** types tend to have higher polarity, while posts from the **Introversion-Intuitive-Thinking** types tend to have lower polarity, compared to most other types. Still, the polarity is relatively neutral.



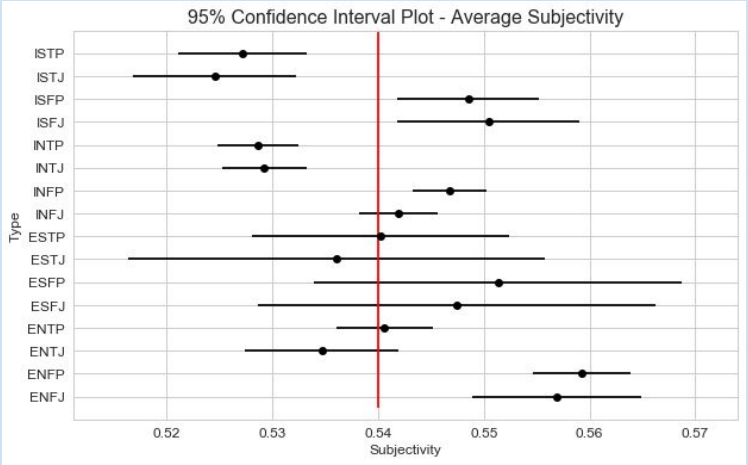
Polarity ranges from -1 to 1. Scores closer to -1 are more negative in tone, closer to 0 are more neutral, and closer to 1 are more positive in tone.

Sentiment Analysis: Subjectivity



The average subjectivity of posts is **0.54**.

The posts from **Extroversion-Intuitive-Feeling** types tend to have higher subjectivity than those with *Extroversion-Intuitive-Thinking*, *Introversion-Intuitive-Thinking*, or *Introversion-Sensing-Thinking* traits. Still, the sentiment is neither overly objective nor subjective.



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 *One vs. the Rest* model performed best when predicting the most abundant classes in the dataset, but was less reliable on the rarer classes.
- Predictions on the *Feeling-Thinking* axis were the most accurate of the four individual axes.
- **Post length** tends to be higher among the *Intuitive-Feeling* personality types, while ESFPs tend to use the fewest words per post.
- The **polarity** of posts is highest among the *Extroversion-Intuitive-Feeling* types, although the posts tend to be neutral in sentiment overall.
- The **subjectivity** of posts is highest among the *Extroversion-Intuitive-Feeling* types, although overall the posts tend to neither overly subjective nor objective in sentiment.

Next Steps

- Analysis
  - 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.