Introvert or Extrovert - Capstone Final Report

Jose Reyes – Jan. 2018
Springboard Data Science Career Track

I. Introduction

For ages scholars have studied and aimed to classify human personality; humourism in ancient times classified people's personalities based on the humors (phlegm, blood, and bile) and more recently the Myers-Briggs Personality Type Indicator (MBTI) assesses the way humans view the world around them. Although MBTIs can change for a single person over time, suggesting weak validity and poor reliability, it has grown in popularity with the business world often serving as the starting point for career choice and as a way to communicate with various types of employees.[1]

While it is considered good practice by career counselors and business professionals to take the Myers-Briggs for self-assessment, it is often costly and time-consuming. Also, if the need arises to target clients with a particular marketing campaign, it would be useful to know if the audience can be considered introverts or extroverts. Many internet users, leave comments on sites such as YouTube, Yelp, and Rotten Tomatoes to name a few, but what if these posts can be traced back to the user and their personality types, could this be predicted from a user's text comments? The ability to do so would provide firms with a cost-effective solution to quickly asses the personality type of potential clients without the need to formally take the MBTI assessment. This way the firm can better customize the way it presents its information to the client, make a lasting impression and improve the firm's bottom line. This study will specifically aim to deploy Natural Language Processing techniques and craft predictive algorithms to predict whether a user is an introvert or an extrovert using internet posts.

II. Description of the Dataset and Data Wrangling

The dataset utilized for this study is the Myers-Briggs Personality Type dataset and it was obtained from the Kaggle website.[2] It contains 2 variables and 8675 rows. The two variables include the user's MBTI type and a posts column which contains the user's text comments on internet forums and social media websites such as YouTube and Facebook. Initial inspection of the data showed that alongside each text post the URL was stored and had a triple-pipe delimeter (i.e. "|||") to mark the boundary of each unique post. A raw data sample post is shown below:

```
In [8]: # Original Post
        print(df.posts[0])

'http://www.youtube.com/watch?v=qsXHcwe3krw|||http://41.media.tumblr.com/tumblr_lfouy03PMA1qa1rooo1_500.jpg|||enfp an
d intj moments  https://www.youtube.com/watch?v=iz7lE1g4XM4  sportscenter not top ten plays  https://www.youtube.com/
watch?v=uCdfzeletec  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 Experie
nce 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 relation
ship to not be perfection all the time in every moment of existence. Try to figure the hard times as times of growth,
```

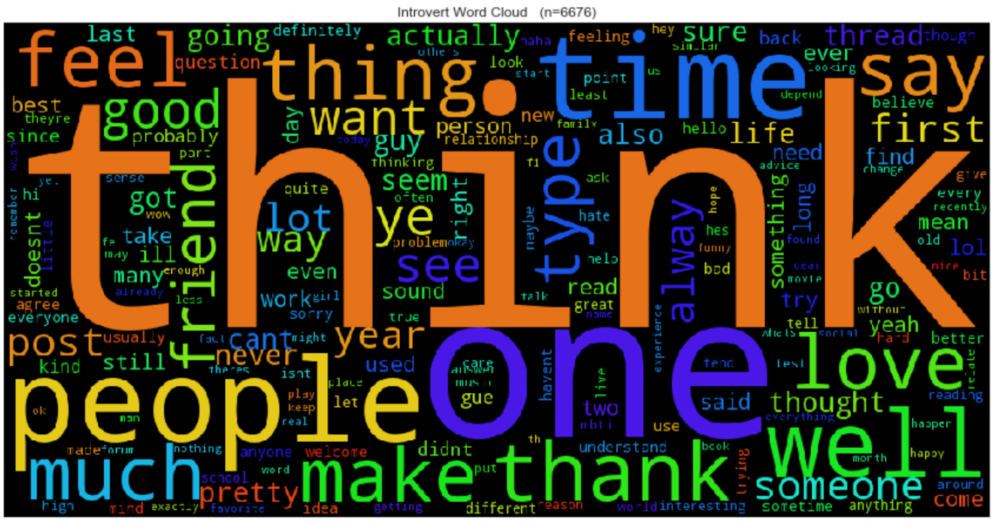
[1] http://www.indiana.edu/~jobtalk/HRMWebsite/hrm/articles/develop/mbti.pdf
[2] https://www.kaggle.com/datasnaek/mbti-type

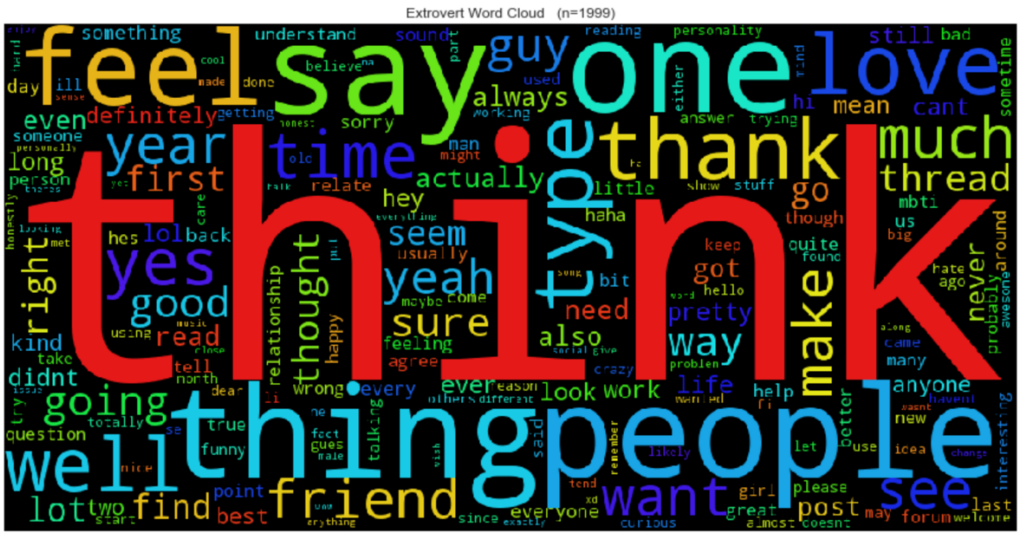
To better classify users as a Introverts or Extroverts, a new “attitudes” column was generated using the first character of the user’s MBTI; if user’s MBTI began with an “E” they were classified as Extroverts, and conversely, if the user’s MBTI began with and I they were classified as Introverts.

In order to better prepare the data for analysis and predictive modeling, each post was stripped of all punctuation, URLs, and numeric values. Lastly, each post was lower-cased and stripped of all stop words and any MBTI values embedded in each post.

III. Data Story

Initial analysis showed that ~75% of the users were classified as Introverts, with the remaining 25% were classified as Extroverts. Additionally, there were many more comments made by introverts than by extroverts. These users comments were quantified and organized into the word clouds displayed below.

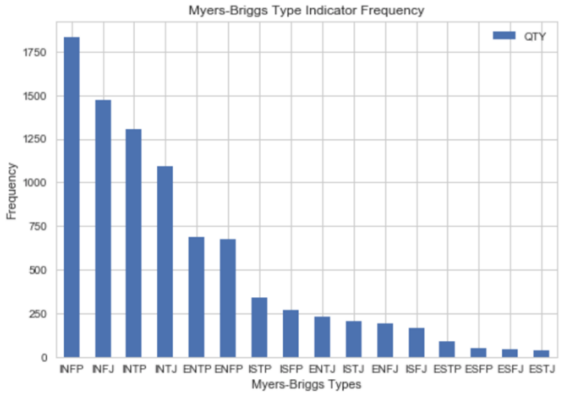




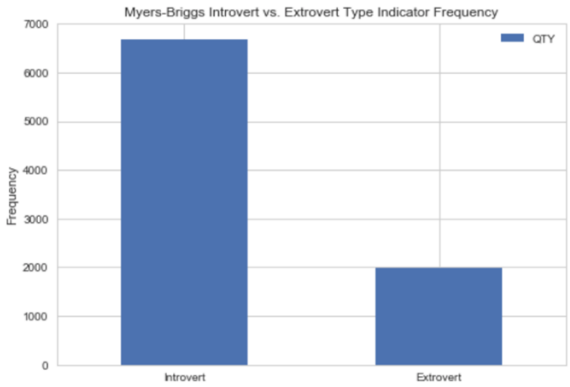
Visual inspection of these word clouds illustrates that the most commonly occurring word for both personality attitudes was the word “think.” It is interesting to note that other commonly occurring words such as “people,” “feel,” “thing,” “one,” “love”, and “say” are more prevalent in usage by Extroverts than by Introverts. This would suggest that the comments used by Extroverts are more social in nature given the frequency of this word usage. Conversely, Introverts appear to have more quantitative language patterns given their high frequency use of words such as “time”, “much,” “make,” “one,” and “first.” Though every effort was made to remove all stop words, according to the word clouds, a few were able to escaped the pre-processing stage. Although their presence is minimal; example of stop words currently present include “got”, “also”, and “ye.”

IV. Inferential Statistics

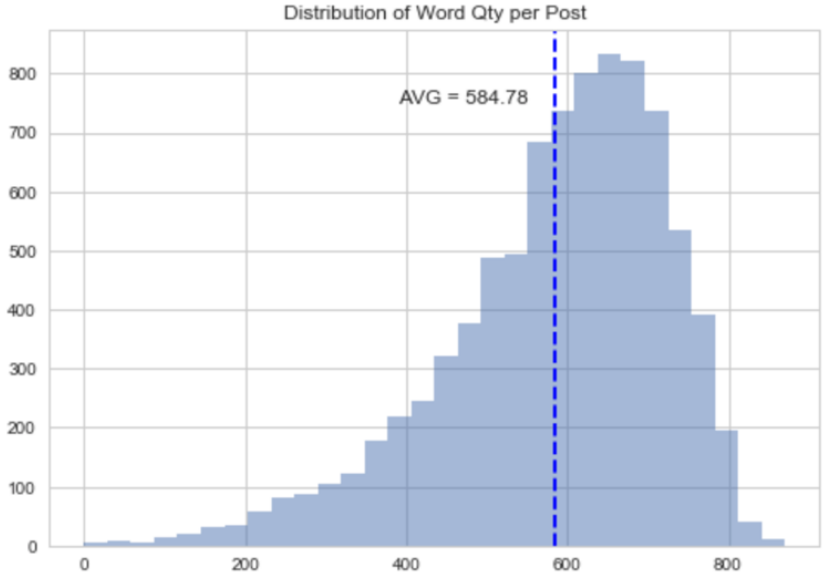
The dataset features a higher frequency of users identified as INFP (Introverted-Intuition-Feeling-Perceiving, n=1832) along with other users listed as introverts. Conversely, users listed as extraverts were fewer overall (6676 Introverts to 1999 Extroverts). The graphs bellow illustrates both the frequency per MBTI along with Introvert vs Extrovert classifications.



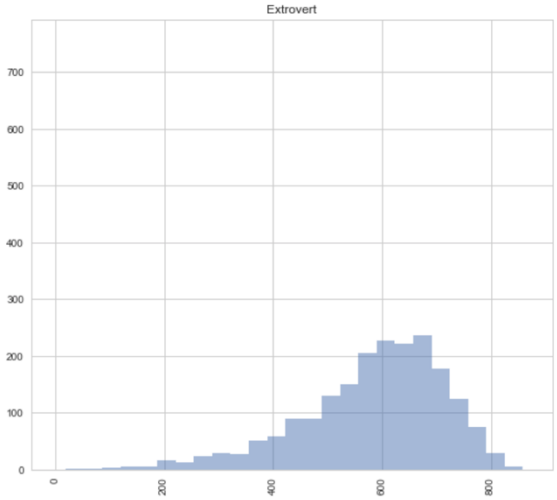
	INFP	INFJ	INTP	INTJ	ENTP	ENFP	ISTP	ISFP	ENTJ	ISTJ	ENFJ	ISFJ	ESTP	ESFP	ESFJ	ESTJ
QTY	1832	1470	1304	1091	685	675	337	271	231	205	190	166	89	48	42	39

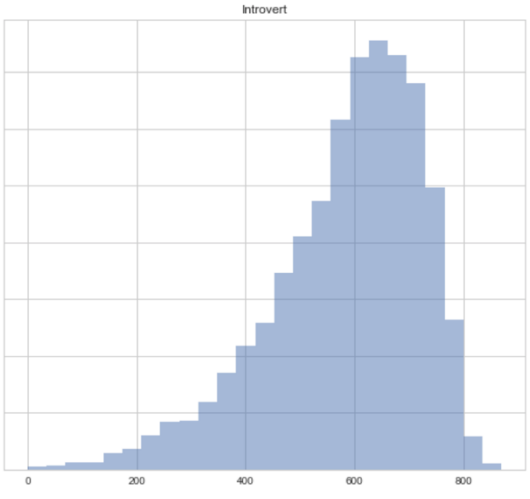


Additionally, each user's post was quantified and plotted below, the result is a negatively skewed distribution of total words per user. The average word count for all combined user posts is 584.78 with a median of 609 words per user posts for all MBTI types.



Additionally, the distributions are also negatively skewed for both introverts and extroverts, suggesting that both attitudes many not differ significantly in their words quantities from a proportional stand-point and the higher frequency of introverts accounts for the difference. It would be interesting to see if this is still the case with equal populations of extroverts and introverts.





ATTITUDE	MEAN # OF WORDS/TOTAL POSTS	MEDIAN # OF WORDS/TOTAL POSTS
INTROVERT	584.97	611
EXTROVERT	584.10	604

V. Machine Learning & Predictive Algorithm Results

To predict whether a user is an introvert or an extrovert the data was split into a training set consisting of 90% of the original dataset, and 10% into a testing set. Five predictive algorithms were trained on the training set, and tested for accuracy on the testing set. The five algorithms utilized were: Multinomial Naïve Bayes, Logistic Regression, Support Vector Machines, Decision Tree Classifier, and Random Forrest Classifier.

To prepare the data, the attitude column of the dataset, which denotes if the user is an extrovert or an introvert, was selected as the target variable. The posts column was selected as the features. The posts were transformed into numeric values using Scikit-Learn's CountVectorizer and TfidfVectorizer and allocated to both the training and testing sets to prepare for predictive analysis. The results are listed below:

Classifier 1: Multinomial Naïve Bayes

The simple probabilistic algorithm Multinomial Naïve Bayes was selected as the first algorithm due to its strong independence between features. The algorithm was executed for the Count Vectorizer and the Tfidf Vectorizer. In this scenario, the Covectorizer scored higher with an accuracy score of 77.76%.

```
---Multinomial Naive Bayes Model---

CountVectorizer Accuracy Score: 77.76

             precision    recall  f1-score   support

  Extrovert       0.55      0.03      0.06       194
  Introvert       0.78      0.99      0.87       674

avg / total       0.73      0.78      0.69       868

Tfidf_Vectorizer Accuracy Score: 77.65

             precision    recall  f1-score   support

  Extrovert       0.00      0.00      0.00       194
  Introvert       0.78      1.00      0.87       674

avg / total       0.60      0.78      0.68       868
```

Classifier 2: Logistic Regression

A Logistic Regression predictive algorithm was also selected for this problem given the binary/categorical nature of the dependent variable. The Logistic Regression algorithm was also applied to both the Count Vectorizer and the Tfidf Vectorizer, with the latter achieving a higher score than the Naïve Bayes classifier with 78.11 % accuracy.

```
---Logistic Regression Model---

CountVectorizer Accuracy Score: 77.19

             precision    recall  f1-score   support

  Extrovert       0.55      0.03      0.06       194
  Introvert       0.78      0.99      0.87       674

avg / total       0.73      0.78      0.69       868

Tfidf_Vectorizer Accuracy Score: 78.11

             precision    recall  f1-score   support

  Extrovert       0.62      0.05      0.10       194
  Introvert       0.78      0.99      0.88       674

avg / total       0.75      0.78      0.70       868
```

Classifier 3: Support Vector Machine

The non-probabilistic support vector machine algorithm was also selected for this problem also due to the dependent variable's categorical nature. In this scenario, both the CountVectorizer and the Tfidf Vectorizer achieved an equal accuracy score of 77.65%.

```
---SVM Model---

CountVectorizer Accuracy Score: 77.65

             precision    recall  f1-score   support

  Extrovert       0.55      0.03      0.06       194
  Introvert       0.78      0.99      0.87       674

avg / total       0.73      0.78      0.69       868

Tfidf_Vectorizer Accuracy Score: 77.65

             precision    recall  f1-score   support

  Extrovert       0.00      0.00      0.00       194
  Introvert       0.78      1.00      0.87       674

avg / total       0.60      0.78      0.68       868
```

Classifier 4: Decision Tree Classifier

A decision tree classifier with a max depth of 4 was also selected for this problem for its robust ability to classify target variables. Accuracy-wise the Counter Vectorizer scored higher with a score of 77.30 % over that of the Tfidf Vectorizer's 76.38%.

```
---Decision Tree Classifier---
CountVectorizer Accuracy on training set: 77.94
CountVectorizer Accuracy on testing set: 77.30

Tfidf Accuracy on training set: 78.19
Tfidf Accuracy on testing set: 76.38
```

Classifier 5: Random Forest Classifier

Lastly, to build on the decision tree model, a random Forest classifier was selected with 10 trees per forest. The random forest did no fare better than the decision tree model as its best score was the CountVectorizer score of 76.96%.

```
---Random Forest Classifier---
CountVectorizer Accuracy on training set: 98.77
CountVectorizer Accuracy on testing set: 76.96

Tfidf Accuracy on training set: 98.89
Tfidf Accuracy on testing set: 76.04
```

VI. Findings/Conclusion

To predict whether a user is an introvert or an extrovert, their internet posts on social media websites such as YouTube and Facebook, were cleansed, analyzed and converted to numeric vectors in preparation for predictive analysis. The cleansed dataset was partitioned into a training and testing sets and each algorithm was fitted to the training data and tested on the testing set to assess its predictive accuracy. The algorithm with the highest test set accuracy score of 78.11% was the Logistic Regression algorithm utilizing a Tfidf Vectorizer. The summary of the scores for the other algorithms can be found in the table on the next page.

Predictive Algorithm	*CountVectorizer score*	*Tfidf Vectorizer score*
Multinomial Naïve Bayes	77.76	77.65
Logistic Regression	77.19	**78.11***
Support Vector Machines	77.65	77.65
Decision Tree Classifier	77.30	76.38
Random Forest Classifier	76.96	76.04

*Best Accuracy Score

Though a score of 78.11%, achieved by the Logistic Regression algorithm is descent, it would be advisable to further refine the data and the algorithms until a minimum score of 90% or above can be achieved. To further refine this predictive algorithm, it is recommended that the posts be analyzed with further scrutiny and corrected for possible misspelled words, refined so that words with common roots be identified and eliminate additional stop words that may have lingered in the original analysis. It is also advised that more participants be assessed so that equal amounts of introverts and extroverts be included in a new study. Lastly, it is also recommended that additional data be provided such as demographic information about a participant, the time spent in each website where they composed a post, and the number of posts. The introduction of this additional data will make it possible to have a better picture of each participant and better identify the attributes that define them as either introverts or extroverts.