Predicting Character Traits Through Reddit
Naval Research Labratory

Clarissa Scoggins

Thomas Jefferson High School for Science and Technology
Information Technology, 5584
Mentor: Myriam Abramson

Abstract

This paper discusses the use of subreddits in the prediction of personality as determined by the Myers Briggs Personality Identification system. For the personality types of ESTJ and INFP the users are gathered from the respective subreddits and then the posts of those users are collected. Extracting only the surbeddit names of the posts, the data is aggregated then clustered. The centroids of these clusters are used to classify the data. This method proves to be more successful for the INFP personality type than the ESTJ personality type suggesting that through this method, certain users may be easier to classify than others.

1. Introduction

In 2015 the Internet reached it's three-billion user milestone including penetration levels up to 80-100% in highly developed countries (Int, 2014). With such high levels of Internet usage comes increased social media activity, making it not surprising that what we post online can be used to define our characteristics (Golbeck et al., 2011). In fact, not only what we post online, but also where we post online mirrors users' personalities (Kosinski et al., 2012). The significance behind personality is monumental. As a known predictor for behavioral trials and attitudes, personality can be utilized by psychologists, criminal profilers, and even employers (Res). Companies like Netflix also use personality classification algorithms in order to provide users with predictions of movies they might enjoy, clustering users with similar preferences together (Vanderbilt, 2013).

Recently researchers have been analyzing popular social media platforms, such as Facebook and Twitter, to predict the personalities of users (Schwartz et al., 2013). The question is, can these predictions be extended to other platforms and can the inputs be simplified? Reddit, which first went live in 2003, now has over 170 million users per month, hailing from 213 different countries (Abo). In comparison Facebook has 1.3 billion monthly users and Twitter has 271 million monthly users (McCarthy, 2014). This paper details the attempt to use publicly available Reddit data in order to classify users' personalities.

The research is grounded in the assumption that users within a certain subreddit share similar characteristics. Subreddits are an inherent classification within the structure of Reddit itself, making the the retrieval of key identifiers of a user's personality easy. By simplifying classification down to two of the sixteen personality types as categorized by the Myers Brigg's personality assessment, the problem became whether a certain user was more similar to one personality or the other. Examples of each personality type were initialized from the data as the centroids produced from the k-Means algorithm; then, using k-Nearest Neighbor algorithm, a test was conducted to determine whether or not the examples of each personality type were true representations of that personality. The correct classification of the original data proves that this simplified model for personality prediction is accurate.

2. Background and Related Work

An initial assumption fundamental to studying the relationship between social media and personality is that social media is indeed reflective of users' real personalities and not an idealized version (Back et al., 2010). Despite this, there is still a lack of ground truth

inherent in social media evaluation. This lack of ground truth makes it difficult to validate assumptions and conclusions, however many scientific methods can be tweaked in order to evaluate social media and draw valid conclusions from the results (Zafarani and Liu, 2015).

Direct personality classification based on publicly available social media data has been attempted successfully my prior researchers. A study done by Gobleck et al. focused on personality prediction based off of Twitter. They administered a Big-Five Personality Inventory to their subjects at the start of their study. While they tested a variety of variables, they concluded that only social processes, negative emotions, cognitive mechanisms, perceptual processes, biological processes and punctuation were statistically significantly language features in personality prediction. The non-Linguistic Inquiry and Word Count features they analyzed all proved to be insignificant(Golbeck et al., 2011). The significance of language processing and word choice of users has previously been proven to correlate with users' personalities, however these algorithms require significant analysis of users' posts (Schwartz et al., 2013). A major difference this paper brings is the proposal of subreddits as the main feature used in personality prediction, choosing to bypass language analysis. Prior research has proven that website choice can be used to predict personalities; and as each subreddit is a separate page with its own distinct topic, there are underlying similarities (Kosinski et al., 2012).

While the correlation between personality and social media is often studied using "The Big Five" personality test; however, for this research the Myers Briggs Personality Identification was used. The reasoning for this choice revolves around the fact that the sixteen distinct personalities as determined by the Myers Briggs Personality test each have separate subreddits that data may be pulled from, while the Big Five personality assessment produces scores for each personality characteristic (Quercia et al., 2011). The Myers Briggs Personality Identification revolves around four different aspects: favorite world, information, decisions, and structure. Favorite world is introversion (I) and extroversion (E), asking whether one enjoys their own inner world or the outer world better. Information is the difference between sensing (S) and intuition (N), and whether or not one focuses on the facts or likes to interpret and add meaning. Whether one thinks (T) with logic or feels (F) with their heart is the difference in terms of decisions. Structure is the preference between judging (J), having decisions made and sticking to schedules, or perceiving (P), doing things on a whim. These different traits make up a four letter code with opposite personalities having different letters for each aspect (MBT).

3. Methodology

As shown in figure 1 the main parts to this research are data collection followed by clustering and classification of data which feeds into a web interface. The data collection process started by retrieving users from personality subreddits and then the posts of those users. The only feature observed was the subreddit name and all other collected features were ignored. The data was then aggregated for comparison and the resulting feature vectors were input to a k-Means algorithm from which centroids were extracted. Those centroids were then used as training data in a k-Nearest Neighbor algorithm which took a username as input in order to retrieve the test data from the feature vectors. This

algorithm produces a prediction for that user's personality which was then displayed on a web page. These processes are detailed as follows.
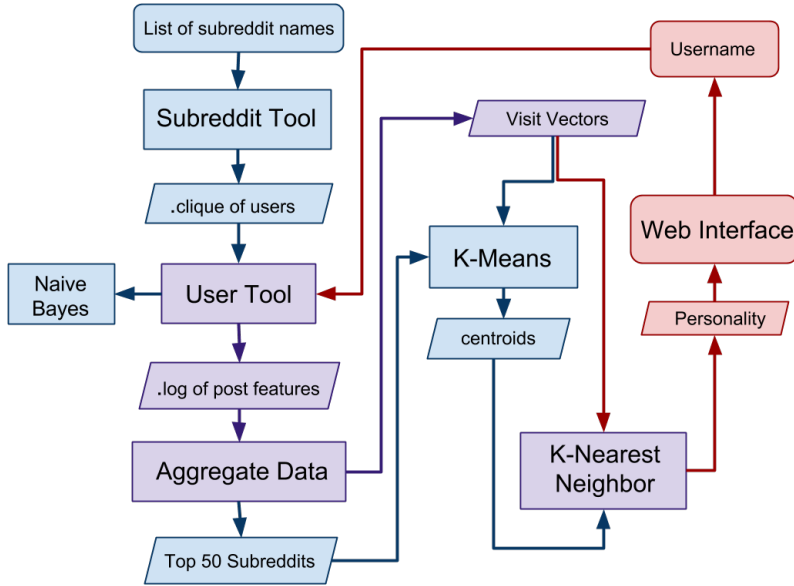

Figure 1: Boxes colored blue are part of the data collection and analysis. The red boxes are processes involved in classification and the web interface. The boxes colored purple are part of both. The first column depicts the initial data gathering.

3.1 Data Collection

In order to retrieve data a Python API tool, the "User Tool," created by a student previously under my mentor was altered and used as a base model for another API tool, the "Subreddit Tool." The Subreddit Tool created took a input file with the names of subreddits, retrieved the 100 most recent posts from each subreddit and then output a .log file containing the users of the 100 most recent posts. The Subreddit Tool was run on two opposing personality types, INFP and ESTJ. The altered User Tool took an input .log file with names of redditors and created .clique files for each redditor, containing their 1000 most recent posts and comments with a feature vector containing a timestamp, title, and subreddit name. While many feature can be pulled using PRAW, all that was needed for this particular study was the subreddit name, which would later be extracted from the data during aggregation. The User Tool was then run on the result files from the Subreddit Tool for each personality type. Across both tools PRAW, the Python Reddit API Wrapper, was used to pull all the information needed from Reddit (PRA). This wrapper simplified interaction with Reddit's API significantly.

The User Tool was altered to include a simple positive-negative sentiment analysis. This was done using the Python NLTK, Natural Language Toolkit, and their movie review library. The library was used to train a Naive Bayes classifier that was called when the API tool processed each post or comment. The classifier would return whether the post or comment was positive or negative and the probability for each (Bird et al., 2009). Although this classifier could have benefited from a neutral class, the goal was to do simple language analysis and it was simpler not to include one. Although this sentiment analysis was not used in later data analysis it was valuable for potential future use.

4

In order to do comparisons between two users, the data was aggregated and a new feature vector was created for each user. After simplifying the features to simply the user and the subreddits they have visited, the top 50 most visited subreddits of each personality type were defined. Using that vector, the number of each user's visits to each subreddit in the top 50 vector was put into another feature vector. It is important to note that the feature vectors were created for all users in respect to both personality types. This was in order for proper comparison of test data and training data. Since the top 50 were different for each personality type, one feature vector would not work in comparison to both personality types. Following all these processes, a set of vectors for all users visits in comparison to both top 50 subreddit vectors were obtained and put into a .log file.

#	ESTJ	INFP	#	ESTJ	INFP
1	AskReddit (2)	Infp (5)	26	Nootropics	mbti (2)
2	mbti (26)	AskReddit (1)	27	Fitness	videos (23)
3	Portland	AskMen (15)	28	casualiama	cringepics
4	intj	funny (19)	29	councilofkarma	pokemon
5	Infp (1)	AdviceAnimals (9)	30	technology	infj (20)
6	keto	leagueoflegends	31	IAmA	entp (14)
7	ENFP	exmormon	32	Denver	redditgetsdrawn
8	EdenDirect	inFAMOUSRP	33	Christianity	Anxiety
9	AdviceAnimals (5)	pics (16)	34	canada	cringe
10	leangains	ffxiv	35	ShittyPoetry	aww (48)
11	orangered	survivor	36	ADHD	DotA2
12	worldnews (43)	gaming (42)	37	offmychest	todayilearned (22)
13	INTP (20)	WTF	38	titanfall	DinosaurDrawings
14	entp (31)	Rabbits	39	linux	katawashoujo
15	AskMen (3)	Music	40	me_irl	INFPmusic
16	pics (9)	Psychonaut	41	ESTJ	Showthoughts
17	TrollXChromosomes	trees	42	gaming (12)	news (47)
18	periwinkle	depression	43	conlangs	worldnews (12)
19	funny (4)	DebateReligon	44	MakeupAddiction	Gaming4Gamers
20	infj (30)	INTP (13)	45	fffffffuuuuuuuuuuuuu	Warframe
21	philosophy	atheism	46	politics	writing
22	todayilearned (37)	CasualConversation	47	news (42)	formula1
23	videos (27)	bicycling	48	aww (35)	DnD
24	entj	gifs	49	GreatAurantiaco	anime
25	actuallesbians	gamegrumps	50	TrueReddit	cars

Table 1: These are the top 50 subreddits for each personality types. The blue boxes are the subreddit of that column's personality while red boxes are the subreddits of other personality types. The number in parenthesis beside certain subreddits is the position of that subreddit in the top 50 vector of the opposing personality type.

The top 50 subreddits contained some interesting trends, such as the fact the top 50 subreddits for ESTJ users had significantly more subreddits of other personalities within it in comparison to the top 50 subreddits of INFP users. This can be seen in table 1 which depicts the top 50 vectors for each personality. The number four most visited subreddit is intj and the number five is infp. Their own subreddit, ESTJ, is in the number 41 spot, significantly below several other personality types. This might indicative that the assumption that individuals visiting a personalty's subreddit is of that personality. However, the top subreddit for INFP users is INFP and the few other personality subreddits within the top 50 are personality types very similar to INFP. While there are 16 subreddits that appear in both personality's top 50 vectors, not a single subreddit has

the same ranking. This appears to signify that there is a discernible difference between the trends of each type of user. It is also likely that several of the 16 overlapping subreddits would also appear in the top 50 vectors for other personality types, as they are simply popular subreddits, neutral to personality.

3.2 Clustering

In order to find the patterns that could be used to define the data, the feature vectors for users of each personality were run through a clustering algorithm. The goal was to define visit trends unique to each personality that could create a model for that personality. The k-Means algorithm as defined in algorithm 1 was chosen for its simplicity and efficiency.

In order to complete an initial prototype of the function, the value of ten was arbitrarily chosen as k for both sets of data although repetition in the original gathering of users had altered the true data set's size. After this first version was completed, a program was developed to calculate the average diameter of the clusters and the average distance between clusters, outputting these statistics for varying k values. Using LibreOffice Calc the average cluster diameter for each k averaged over 10 trials was graphed. Through a power regression trend line, the correct k value for each data set was estimated visually to be around 6 for both personality types, and then the k values were determined again through the L method.

As shown in figures 3 and 4, the graph of average cluster diameter in comparison to the k value has a clear bend, referred to as the "knee" of the graph. Since the k value lies at the bend of the knee, the L method looks to isolate the bend using the distinct patterns portrayed by each half of the graph before and after the bend. By iterating through different combinations of points on each half of the knee, a pair of lines that fits the graph with the least amount of total error was determined (Salvador and Chan, 2004). The total error was a scaled sum of the Root Mean Square Errors (RMSE) for each half of the graph. The x-value at the intersection of the pair of lines, called the c value, is the set as k value. (Salvador and Chan, 2004). The graph and pair of lines determined through this method are depicted for each personality in figures 3 and 4. Any RMSE or c values less than 0 were ignored in the collection of data. The best c value for ESTJ was 7 when lines were drawn from the points at k values 2 and 4 on the left and 16 and 19 on the right. The best c value for INFP was 6 as a result of lines drawn using k values 3 and 5 on the left and 14 and 19 on the right. The k-Means algorithm was run with the result k values to collect a final set of centroids.

3.3 Classification

The centroids were then tested by classifying the data, using the centroids as the training set. The algorithm used was the k-Nearest Neighbor algorithm for its simplicity and efficiency, much like the k-Means algorithm. The steps of the k-Nearest Neighbor algorithm are as detailed in algorithm 2. For k-Nearest Neighbor, the k value is usually determined empirically. A common method of selection is to set k as the square root of the number of elements in the data set (Duda et al., 2012). The value of 5 was later determined to be the best k value.

Input: Visit vectors
Initialize centroids randomly
while centroids have not converged **do**
    Associate each data point with the centroid closest to it
    Reassign each centroid as the geometric mean of the data points associated with it
end while
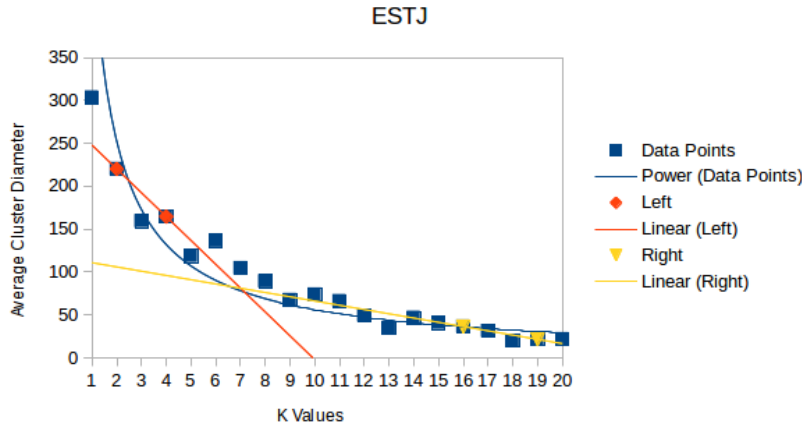**Output:** K centroids

Algorithm 1 K-Means algorithm

Figure 2: This is the graph used to determine k for INFP. The intersection of the two lines as determined by the L method has an x-value of 7 with a RMSE of 13.96. The line on the left is determined by k values 2 and 4 and the line on the right is determined by k values 16 and 19.
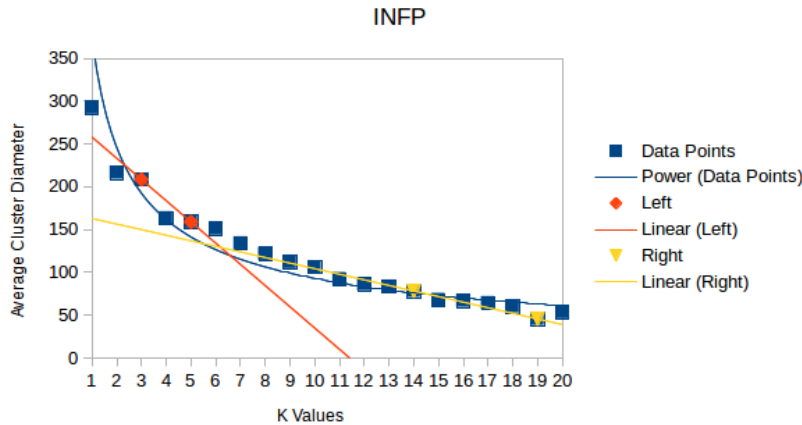

Figure 3: This is the graph used to determine k for INFP. The intersection of the two lines as determined by the L method has an x-value of 6 with a RMSE of 9.54. The line on the left is determined by k values 3 and 5 and the line on the right is determined by k values 14 and 19.

Input: Username
Retrieve the training and test data
Calculate the distance between the test data and each of the training data points and find the k nearest points
Have each neighbor vote for a prediction and return the predicted values
Calculate the accuracy of the prediction
Output: Class for the test user

Algorithm 2 K-Nearest Neighbor algorithm

	Right Predictions	False Predictions	Accuracy
ESTJ	3	36	7.69%
INFP	69	17	80.23%
Total	72	53	57.60%

Table 2: Overall final statistics run with k values of 7 for ESTJ and 6 for INFP.

ESTJ k Value	INFP k Value	ESTJ Right	INFP Right	Total Right	ESTJ Wrong	INFP Wrong	Total Wrong
5	11	0	82	82	39	4	43
5	13	0	82	82	39	4	43
6	14	0	82	82	39	4	43
5	11	0	81	81	39	5	44
6	13	0	81	81	39	5	44
Average 5.4	12.4	0	81.6	81.6	39	4.4	43.4

Table 3: The combination of k values for k-Means that produced the best overall accuracies. The best k value for k-Nearest Neighbor was 5 over every trial.

The algorithm was nested within a web page created to interface with the project in order for ease of understanding. The product, which is shown in figure 4, consisted of two main pages, the first of which was simply pure HTML with CSS formatting in order to input a Reddit username. This username was sent to the k-Nearest Neighbor algorithm which retrieved that user's vectors from the .log files containing all user's vectors. While the web interface was perfected using users whose data was used within the study, the could potentially retrieve data for users outside the study and create vectors based on that data in order to run the k-Nearest Neighbor algorithm. This version of the simply took an inconvenient amount of time to run for demonstration, so the web interface remained only functional with users within the study. The k-Nearest Neighbor program including CGI scripts to display the prediction along with statistics on the accuracy of the prediction and distances from each personality type.

In order to test the accuracy of the classification, a driver was created to classify each user within the study and print whether the output was right or wrong, as well as simple statistics about the success and failure of the overall research. These results can be seen in table 2. Another driver which ran different k values for k-Means ranging from 5 to 15 and k values from 3 to 7 for k-Nearest Neighbor was created to determine which combination of k values produced the highest overall accuracy. The results of this test can be seen in table 3.

4. Discussion

While the classification using the centroids resulting from the L-method k values had about an 57% accuracy, the separate accuracies were 80.23% for INFP and 7.69% for ESTJ. The tests conducted with varying k values showed that the best accuracy that could be extracted was 65.6% overall when the k value of ESTJ was 5 and the k value of
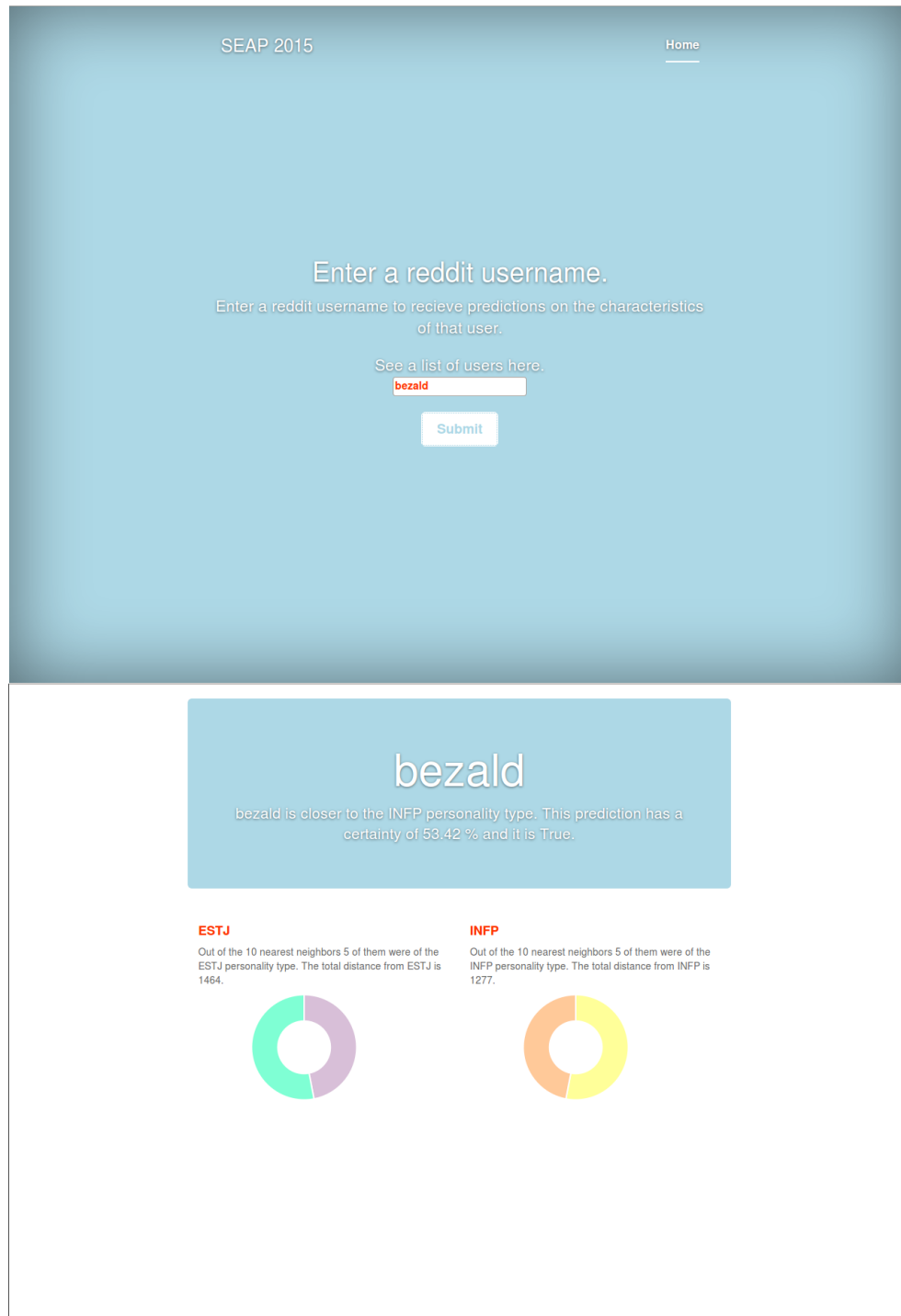
Figure 4: The web interface which can be accessed at www.clairescoggins.com/nrl.html. The first page allows for an input of a Reddit username and the second page classifies that user, displaying the results.

INFP was 11, 12 and 13. This difference in k values between ESTJ and INFP is likely a result of the variance in size of data as the INFP data set was larger than the ESTJ data set. The L method may have produced different results due to a small or inaccurate data set. However, as shown in tables 2 and 3 there is a clear discrepancy in the accuracy of predictions for each personality type regardless of the k value. The k-Nearest Neighbor algorithm using the k values determined by the L method predicted 73 personalities correctly out of 125, however this breaks down to 73 correct predictions of INFP out of 86 INFP users and 3 correct predictions out of 39 ESTJ users. When the overall accuracy is optimized, the total correct predictions increases to 82, all of which are INFP predictions. The classification algorithm then only produces 4 incorrect INFP predictions while all 39 ESTJ users are falsely classified. The overall trend appears to be that INFP was easier to predict, and if the accuracy in predicting ESTJ were to increase, the accuracy in INFP predictions would decrease.

There are several potential reasons for these results. The first potential reason is that this method of clustering and classification based on subreddit is not an accurate measure for predicting personality. Another potential reason is that the data collected is a faulty data set, since it is assumed that any user posting within a personality's subreddit has that personality. If a personality test was conducted prior to experimentation, as done in several other studies, perhaps more accurate and better results could be reached since in this study there was no confirmation on whether each user was actually of the personality type they were assumed. Less significantly, a larger data set and normalization of data could also improve predictions. This could also allow for a larger feature vector, expanding the scope beyond the top 50 subreddits. The initial data collection skewed the amount of data obtained, since the Subreddit Tool did not check for repetition in users when collecting them. This difference in size of intial data most likely led to the inconstancy seen in the accuracy of predictions of INFP users in comparison to the accuracy of predictions for ESTJ users.

5. Conclusion

This research has suggested that certain personality types are easier to classify than others using subreddit visit patterns. Expansion of the study to span over all 16 personality types would allow for further insight to the true success of subreddits in the classification of personality. This method could also be used in the prediction of other characteristics such as age, gender, location, or even political lenience for users, allowing researchers to potentially create a type of profile for a given user. This profile could be used to determine persons of interest or larger trends among types of users through Reddit or similarly structured social media.

Acknowledgments

References

About reddit. URL `https://www.reddit.com/about/`.

Mbti basics. URL
  `http://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/`.

Getting started. URL
  `https://praw.readthedocs.org/en/v3.1.0/pages/getting`$_s tarted.html$.

Research and validity. URL
  `http://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/reliability-and-validity.`

Internet society global internet report 2014, 2014. URL
  `https://www.internetsociety.org/sites/default/files/Global`$_I nternet_R eport_2 014_0.pdf$.

Mitja D. Back, Juliane M. Stopfer, Simine Vazire, Sam Gaddis, Stefan C. Schmukle, Boris
  Egloff, and Samuel D. Gosling. Facebook profiles reflect actual personality, not
  self-idealization. *Psychological Science*, 21(3), 2010.

Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python -
  Analyzing Text with the Natural Language*. O'Reilley Media, 1st edition, 2009. URL
  `http://www.nltk.org/book/ch06.html`.

Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley &
  Sons, 2012.

Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. Predicting
  personality from twitter. *IEEE International Conference on Privacy, Security, Risk, and
  Trust, and IEE International Conference on Social Computing*, pages 149–156, 2011.

Michal Kosinski, David Stillwell, Pushmeet Kohli, Yoram Bachrach, and Thore Graepel.
  Personality and website choice. *WebSci*, pages 1–4, June 2012.

Niall McCarthy. Facebook versus twitter in numbers, October 2014.

Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcoft. Our twitter profiles,
  our selves: Predicting personality with twitter. *IEEE International Conference on
  Privacy, Security, Risk, and Trust, and IEE International Conference on Social
  Computing*, pages 180–185, 2011.

Stan Salvador and Philip Chan. Determining the number of clusters/segments in
  hierarchical clustering/segmentation algorithms. In *Tools with Artificial Intelligence,
  2004. ICTAI 2004. 16th IEEE International Conference on*, pages 576–584, Nov 2004.
  doi: 10.1109/ICTAI.2004.50.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski,
  Sephanie M. Ramones, Megha Agarwal, Achal Shah, Michal Kosinski, David Stillwell,
  Martin E. P. Seligman, and Lyle H. Ungar. Personality, gender, and age in the language
  of social media: The open-vocabulary approach, September 2013.

Tom Vanderbilt. The science behind the netflix algorithms that decide what to watch next, August 2013.

Reza Zafarani and Huan Liu. Evaluation without ground truth in social media research. *Communications Of The ACM*, 58(6):54–60, June 2015.