American Journal of Applied Psychology
2021; 10(1): 21-26
http://www.sciencepublishinggroup.com/j/ajap
doi: 10.11648/j.ajap.20211001.14
ISSN: 2328-5664 (Print); ISSN: 2328-5672 (Online)

The Interdependency of the Diction and MBTI Personality Type of Online Users

Seoyoon Choi

Seoul International School, Seoul, Republic of Korea

Email address:
choieunie@gmail.com

To cite this article:
Seoyoon Choi. The Interdependency of the Diction and MBTI Personality Type of Online Users. *American Journal of Applied Psychology*. Vol. 10, No. 1, 2021, pp. 21-26. doi: 10.11648/j.ajap.20211001.14

Received: January 25, 2021; **Accepted**: February 25, 2021; **Published**: March 3, 2021

Abstract: This paper offers insight into the 16 Myers-Briggs Type Indicator (MBTI) personality types and how they may affect the diction used by online users on social media platforms such as Twitter and YouTube. The Myers-Briggs Type Indicator categorizes individuals who take the indicator test into one of 16 different personality types, and each of these types have distinct characteristics, from the simple Introverted versus Extraverted to Intuitive or Sensing, Feeling or Thinking, and Judging or Perceiving. These 4 sets of binary characteristics produce 16 different personalities that are often used to create general pictures or summaries about the individual who was assigned a certain personality type. The characteristics can, on occasion, even predict the potential actions of the individual based on their assigned personality type. This is what allows for the objective of this paper to be achieved - to use data analysis and machine learning to identify the number of times certain words were used by those of different personalities on online platforms, find patterns, and observe if the mechanic prediction of MBTI type based on words used in online posts is possible. The three machine-learning algorithms used to predict the personality types were the Naive Bayes, Gradient, and Random Forest algorithms, with a randomly-selected 80% of the data being used to train the algorithms and the remaining 20% being used to test the machine-learning for accuracy and specificity. This paper will analyze 433,750 total individual posts made online, along with the programming-processed data and the final results of the predictions, identifying which algorithm was most effective in predicting MBTI type and what future steps could be taken to increase accuracy and capacity.

Keywords: Myers-Briggs Type Indicator, Personality Types, Data Analysis, Machine Learning

1. Introduction

Personality tests have gained popularity in the twenty-first century as they became more easily accessible through the Internet. Online assessments such as the Myers Briggs Type Indicator and the Big Five Personality Test sort individuals into different personality categories, complete with explanations for each of the personality types and, in the case of the Myers Briggs Type Indicator, percentages for the test-taker's different regions of personality [1].

Created by the Myers-Briggs mother-daughter duo near the end of World War II, the Myers Briggs Type Indicator, or MBTI for short, is arguably the most well-known online personality test [2]. It sorts individuals into 16 different personalities, assessing different aspects in four sets of binary characteristics (introverted/extraverted, intuitive/sensing, feeling/thinking, and judging/perceiving) [1]. The computer programming and machine learning in this paper will analyze these four binary characteristics in the Myers-Briggs Type Indicator. Though there is also one additional binary characteristic (Turbulent vs. Assertive), this will be disregarded as it has more to do with how the individual takes the type indicator test than their actual resultant personality.

Though many psychologists have shunned this particular test for being ineffective at assessing personality, as it relies on a limited set of binary characteristics when humans actually fall on an entire spectrum [3], the MBTI test is currently being used in environments outside of casual passerby interest. Widely used by the general public, the MBTI test is also being used in corporations, workplaces, and even government agencies such as the state department and the CIA to separate employees into different categories and allocate them to different work programs and

assignments based on the resultant personality type [6].

The final purpose of this paper is to utilize computer programming, data analysis, and machine-learning algorithms to analyze individuals' online posts to identify the user's MBTI personality type. Though personality tests are never exactly precise, they do create general pictures and identify trends among certain groups of people, as revealed by a reported sense of identity among people of the same MBTI type [4, 5]. This would allow the computer to identify certain patterns in the diction of the online users during its machine-learning algorithm training, allowing the computer to identify the personality type of the online user with a much higher accuracy rate than random guessing.

2. Preprocessing & Transformation

2.1. Preprocessing

The preprocessing began with adjusting all words to be lowercase. This was because later on, when the words were to be counted to observe usage frequency, the words with uppercase letters would be counted separately from their lowercase counterparts, and the preprocessing was to prevent that from happening. For example, the word "Math" would be counted separately from "math", and the preprocessing would change "Math" to "math" so that the words were counted together.

The second step was to remove the stopwords in the data. Stopwords are English words that do not add significant meaning to a sentence, thus making it safe to disregard and remove them without undermining the meaning of the sentence [7]. Examples of stopwords are "the", "a", "and", and "he". These were removed from the dataset for more efficient, meaningful analysis later on.

The third step was to delete the URLs of the links to the webpages on which the phrases were posted. Because most URLs were redundant and didn't hold much meaning, all beginning with "https://" followed by a series of numbers and letters, they could be safely removed from the data for more efficient counting and analysis.

The next step was to choose between the lemmatization and stemming methods. Lemmatization is a linguistic process with which the different forms of a word are grouped together to be analyzed as a single item that is identified by the word's dictionary form. Stemming is the process of reducing words to their root form to be identified as the same word [8]. Both processes aim to lessen the number of inflectional forms of a word and, sometimes, the other derivative forms of a word, to get the base form. However, stemming aims to just remove the end of the word in the hopes of getting the base form right with the removed words, while lemmatization would attempt to return different forms of the word depending on the state of the word that was used initially (verb or noun). For example, if the initial word was "saw", then the stemming process could reduce the word to "s" while lemmatization would return either "see" or "saw" [8].

Lemmatization was chosen for this paper for this very reason; it would alter the different states of words to find their base form, which would reduce the number of random words resulting from this process (for example, if the word was "organization", stemming could produce "organ" as the base term while lemmatization would attempt to find the base form and would recognize that "organ" is not the base form of "organization".

2.2. Transformation

In statistics, variance essentially measures how far apart certain data points are from their average values [12]. Variance here was applied to measure whether or not each online user, regardless of MBTI type, was consistent in the length of their posts in word count. Variance here not only allows the different lengths of posts by each online user to be measured but also provides another pattern to be analyzed and used later on in the prediction process, aside from the frequency of each word used by the different users.

The average number of words used per post by each personality type was also calculated, which aids in the measuring of variance for each online user as well as gives yet another pattern to be analyzed, potentially increasing the accuracy of the final predictions for which personality type created which post made by the machine learning program.

Because the computer cannot comprehend the letters of the 16 MBTI personality types (ex: ESFP, INTJ) the way people do, the letters also had to be changed into either 0 or 1 through the process of mapping. Because the 16 personality types are determined by a combination of 4 binary characteristics that produce 16 total possible results, the mapping was done so that 0 would represent one of the binary characteristics, and 1 would represent the other characteristic. The 4 sets of binary characteristics are introverted (I, represented by the number 0) or extraverted (E, represented by 1), intuitive (N, 0) or sensing (S, 1), thinking (T, 0) or feeling (F, 1), and judging (J, 0) or perceiving (P, 1). So, for example, the ESFJ personality type would be the numbers 1110, being Extraverted, Sensing, Feeling, and Judging. INTJ would be 0000, being Introverted, iNtuitive, Thinking, and Judging. This allows the computer to analyze the data with vectors, or numbers, instead of an aggregation of letters no different from other words.

3. Exploratory Data Analysis (EDA)

EDA, or Exploratory Data Analysis, is a method of data analysis that utilizes several generally graphical approaches in order to highlight variables, identify outliers or anomalies, test theories or assumptions, develop models, and foster an overall more thorough understanding of the specific data set being analyzed [9]. EDA is not necessarily a set of techniques defined solely with that term - it is simply an approach to data analysis that allows the data set to present itself, instead of having a preconceived notion

forced onto it (for example, the analyzer could push a certain type of graph onto a data set due to that graph form being the one they desire, when the data could actually be more efficiently analyzed when presented in another type of graph, and EDA attempts to prevent this from occurring as it shows whether or not certain statistical techniques are appropriate for that set of data) [10]. This allows for more than just the surface-level graphical model and hypothesis-testing task to be observed, and can go beyond Initial Data Analysis, or IDA.

3.1. General Information About Data

The set of unprocessed data consisted of two columns: the MBTI types of the individuals who posted certain phrases online, and the links to the webpages on which the phrases had been posted. With those two columns, there were 8675 rows of data. The phrases were sections of the last 50 things the individuals posted online, and for each individual the 50 phrases were together in the same row of data. This amassed to 433,750 total individual posts.

Approximately 85.75% of the data came from YouTube, with the participants speaking or writing, identifying their MBTI type, and agreeing to have their words extracted to be used as data. A total of 8675 different individuals' words were extracted, each of them with 50 online posts (with one YouTube video counting as one post).

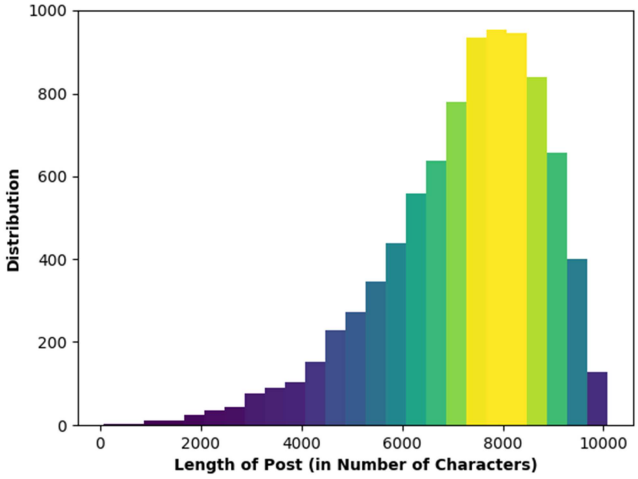

Figure 1. *The Frequency of Posts With Certain Lengths.*

As shown in Figure 1, the majority of the 50 collective posts of each individual had character lengths of around 7500 to 8200. The average English word is approximately 4.7 characters long, and each individual would have had around 1,700 words as the sum of the number of words in their 50 posts.

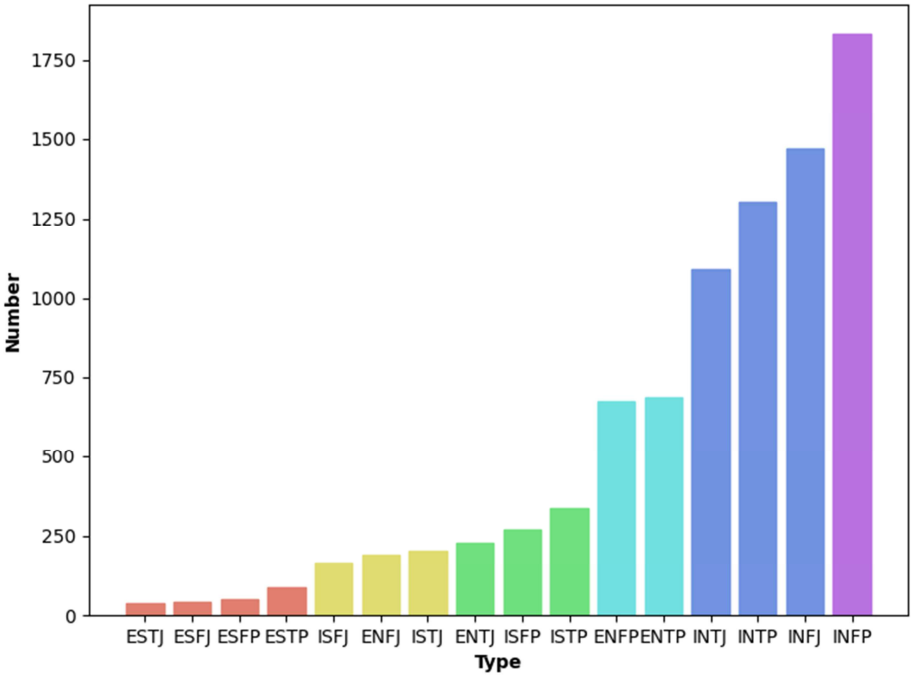

Figure 2. *The Frequency of the 16 MBTI Types in the Sample Data.*

As shown in Figure 2, the INFP MBTI type had the highest frequency of individuals who posted, with 1832, and ESTJ had the lowest frequency with 39.

3.2. Word Clouds

As shown in Figure 3, the three most frequently used words in total were *think, people,* and *know,* each used 68280, 46649, and 43129 times respectively. Because the highest amount of data was collected from online writers who identified as INFJs and INFPs, the word-clouds of the words most frequently used by those personalities appeared to be very similar to the total, all-encompassing word-cloud. The word-clouds below are of the personality types from which less data was collected, with 13 of the most commonly used words (think, people, know, one, say, feel, thing, go, like, get, would, really, make) removed from the word-cloud

for the sake of comparison and contrast. After the 13 most commonly-used words were removed from the dataset, the next most common words were “time” (32386), “want” (26090), “type” (25401), “see” (24787), and “love” (23406).

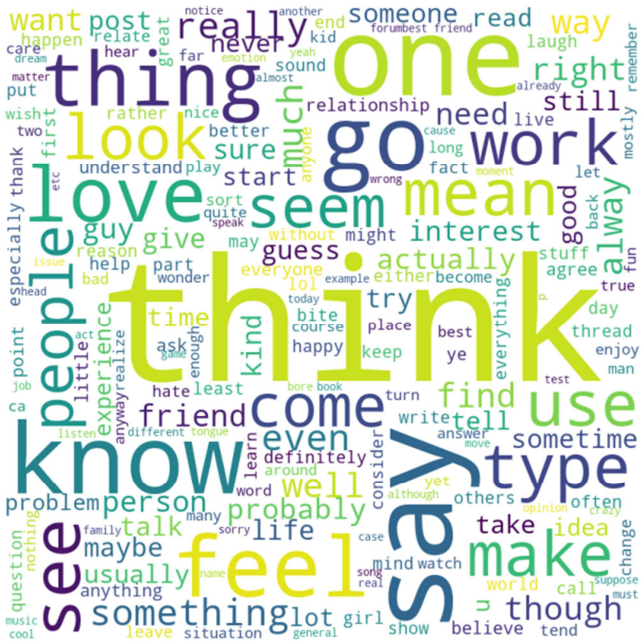

Figure 3. *A word-cloud depicting the frequency of words used in the total data using the size of the words in the word-cloud.*

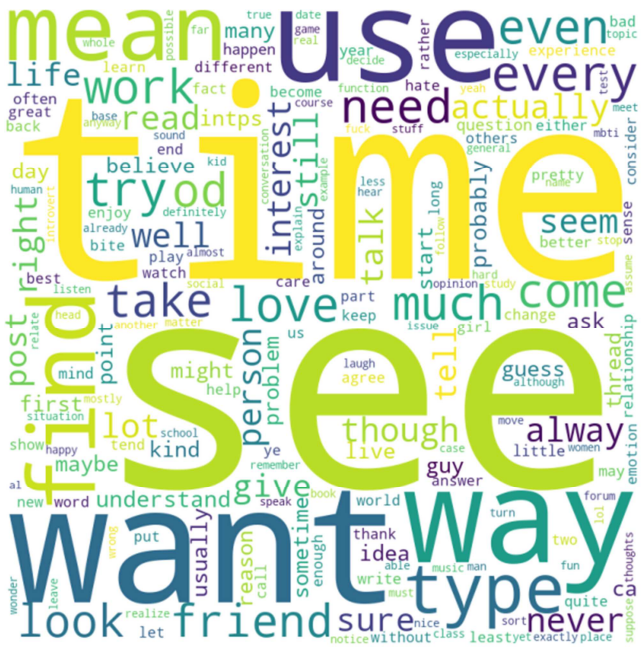

Figure 4. *A word-cloud depicting the frequency of words used in the data by INTP individuals using the size of the words in the word-cloud.*

Different from Figure 3, in Figure 4, the three most frequently used words in total by online users who identified as INTPs were *time, want,* and *see,* each used 4634, 3616, and 3495 times by INTPs respectively. INTP individuals are the “Thinkers” or the “Logicians”, who are Introverted, Intuitive, Thinking, and Perceiving. They are characterized by the Myers-Briggs Type Indicator as “innovative inventors” with an “unquenchable thirst for knowledge” [11].

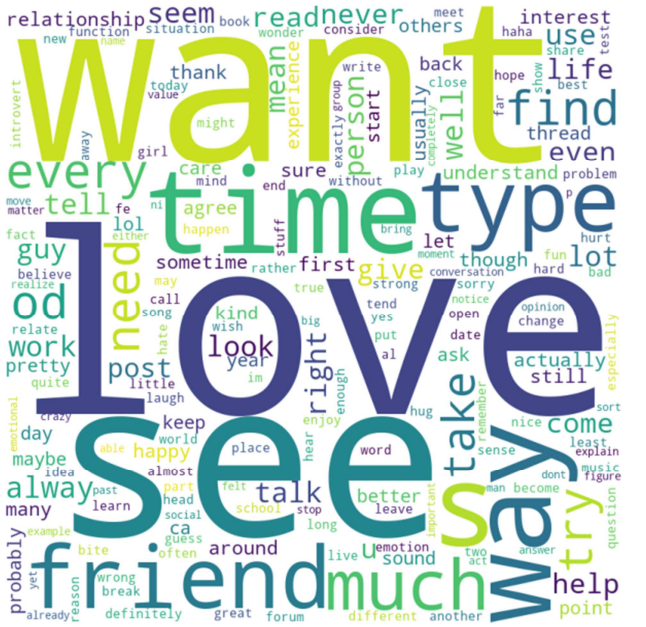

Figure 5. *A word-cloud depicting the frequency of words used in the data by ENFJ individuals using the size of the words in the word-cloud.*

On the other hand, in Figure 5, the three most frequently used words in total by users who identified as ENFJs were *love, want,* and *see*, each used 726, 630, and 587 times by ENFJs respectively. ENFJ individuals are the “Givers” or the “Protagonists”, and are Extraverted, Intuitive, Feeling, and Judging. They are characterized by the Myers-Briggs Type Indicator as “warm, outgoing, and sensitive”, as well as being “charismatic and inspiring leaders” who are perceived as the strongest “*people* people” of the 16 personalities [11].

4. Machine Learning

4.1. CountVectorizer

CountVectorizer is a tool in the programming language *Python* that changes a piece of text, in words, into a vector, or a number that describes the frequency of words used in a particular text [13]. The CountVectorizer can be extremely useful when there are more than one set of texts that need to be converted to be used for further data analysis, and as there were 8675 rows of data, the CountVectorizer can help convert each section of text into vectors.

4.2. TfidfVectorizer

TfidfVectorizer, also known as Term Frequency - Inverse Document Frequency Vectorizer, vectorizes the more “interesting” words of a group, categorizing them by frequency within a certain piece of text but not across all pieces of text [13]. A simple way to understand this concept is by comparing it to different categories of a newspaper: the word “quarterback” may appear frequently in the Sports section of the newspaper, but not so much in the Politics or Finance sections, and the TfidfVectorizer would be able to recognize that. The difference

between CountVectorizer and TfidfVectorizer is that CountVectorizer returns integers, or whole numbers, while TfidfVectorizer returns floats, or real numbers with decimals.

4.3. Machine Learning Algorithms

4.3.1. Naive Bayes

Naive Bayes methods are a series of machine learning algorithms based on Bayes' theorem. Bayes' theorem, named after Reverend Thomas Bayes, states the probability of an event in relation to prior events and conditions that may be connected [14]. Naive Bayes applies this theorem but with the "naive" assumption that features of certain measurements and conditions are independent of each other.

This is a naive assumption because this is almost never true, meaning that almost all of the time, those features *are* dependent on each other [14]. This means that Naive Bayes' would only be useful for the classification of data if the data points were exactly identical.

Usually, this makes the Naive Bayes methods inefficient for data classification, but because of the preprocessing done in this case, different tenses and forms of words (which are the data points being used in this study) have been changed to their base forms, making identical pieces of data out of words with the same bases (ex: laughed, laugh).

4.3.2. Gradient Boosting

Gradient Boosting is a machine learning technique typically used for regression and classification of data, and creates prediction models through a series of weak prediction models (the most commonly used weak prediction models in Gradient Boosting are decision trees, which are tree-shaped models that display decisions and their possible outcomes or consequences) [15]. Gradient Boosting has proven to be successful in a multitude of different situations, as they are customizable to different models and data sets [15].

4.3.3. Random Forest

Random Forest is a machine learning method used for the classification and regression of data. It generates numerous decision trees and, as an output, produces the class that (for classification) is the mode (the value that appears the most often) of the classes among the individual decision trees [16]. Random Forests usually outperform, meaning they have higher accuracy than, single, basic decision trees, but they are generally less accurate than Gradient Boosted trees, which are mentioned above [16].

Table 1. *Algorithm's accuracy (i-e).*

Model	Accuracy	Specificity
Naive Bayes	0.768876080691642	0.781828703703703
Gradient	0.851152737752161	0.870903010033444
Random Forest	0.83371757925072	0.85209003215434

Table 2. *Algorithm's accuracy (n-s).*

Model	Accuracy	Specificity
Naive Bayes	0.862680115273775	0.85846331600231
Gradient	0.899279538904899	0.912989434431323
Random Forest	0.882132564841498	0.887222555488902

Table 3. *Algorithm's accuracy (t-f).*

Model	Accuracy	Specificity
Naive Bayes	0.66729106628242	0.602065131056393
Gradient	0.848270893371758	0.831619537275064
Random Forest	0.830115273775216	0.831117021276595

Table 4. *Algorithm's accuracy (j-p).*

Model	Accuracy	Specificity
Naive Bayes	0.656340057636887	0.828125
Gradient	0.795389048991354	0.80110497237569
Random Forest	0.779250720461095	0.832298136645962

The tables above describe the accuracy and specificity of each of the three machine-learning algorithms used to classify the data. The tables are split into fours - each one for the four binary characteristics that make up the 16 MBTI personality types. The algorithms identified these characteristics separately, and the accuracy describes the accuracy with which it predicted each characteristic based on the word data.

TfidfVectorizer, average words per post, and variance were all used to predict the characteristics using the machine-learning algorithms. Although the data was vectorized in order for the computer to read it, the goal was for the computer to correctly identify each of the characteristics without looking at the answers (the answers being the vectorized personality type names), so these other aspects were used to help the computer learn with the algorithms.

80% of the data was used to train the machine-learning algorithm, and 20% was used to test the algorithm, and the result of the testing of the 20% is depicted in the tables above.

The highest recorded accuracy was 0.8993 rounded, with the Gradient Boosting in the characteristic n-s (Intuitive vs. Sensing). The lowest recorded accuracy, on the other hand, was 0.6563 rounded, with the Naive Bayes in the characteristic j-p (Judging vs. Perceiving). The accuracy for the j-p characteristic was generally comparatively lower than the rest, with no accuracies in the 0.8~ range.

The machine-learning algorithm that generally recorded the lowest accuracies and specificities was the Naive Bayes method, while the one with the general highest was the Gradient Boosting.

The final outcome was that the computer, after the machine learning, was able to correctly guess approximately 80% of each of the four binary characteristics. This means that the computer was able to guess with 80% accuracy whether an online user was, for example, Introverted or Extroverted, but that this guess would be completely separate from its guess about whether that same user was Intuitive or Sensing, or Thinking or Feeling, or Judging or Perceiving.

After each of these steps are completed, the final MBTI type is able to be pieced together with these four individual guesses. The computer is able to guess, with approximately 51.65% accuracy, the final four-letter MBTI type of the online user. Though at first glance 52% accuracy seems relatively low, randomly guessing one of the 16 personality types would result in 6.25% accuracy. The computer is able

to increase that percentage by over 8 times through the utilization of the machine-learning algorithms.

5. Conclusion

This paper used data analysis and machine learning to attempt to predict the MBTI type of online users based on the words they frequently used in their online posts. Programmed preprocessing, Exploratory Data Analysis, TfidfVectorizer, and three separate machine-learning algorithms were used to attempt this, with the different steps of the process accounting for different aspects of the raw data that could affect the final outcome of the percentage of MBTI types the machine guessed correctly based on the words used in the online posts.

The final result of this paper was that the computer was able to predict with an average of 80% accuracy the individual binary characteristics, and with approximately 52% accuracy the complete, four-letter MBTI type.

Because this paper was based on online posts, the vocabulary employed by the online users would have been relatively limited in comparison to, for example, that of written works such as diaries or journals in which individuals are free to express similar thought but with, perhaps, more extensive vocabulary and more personal ideas. Future works may employ data from more written works to improve the range and reach of the study.

Future works may also improve the accuracy of the machine-learning algorithms through the employment of different algorithms used together instead of separately. The improved models could then be applied to different, more refined data sets as mentioned above, in order to create a more comprehensive study of how MBTI personality type can affect the words used by individuals in written works, whether that be online or in personal pieces.

References

[1] T. Var, S. Adam, and S. Pridie “A Study of the Effect of the Myers-Briggs Type Indicator on Team Effectiveness”, American Society for Engineering Education 2003.

[2] “Myers Briggs Personality Types.” *Myers Briggs Personality Types - Introduction and Overview*, www.teamtechnology.co.uk/tt/t-articl/mb-simpl.htm.

[3] G. Boyle “Myers_Briggs Type Indicator (MBTI): Some psychometric limitations” Bond University.

[4] Cherry, Kendra. “Myers-Briggs Type Indicator: The 16 Personality Types.” *Verywell Mind*, 17 Sept. 2020.

[5] Riggio, Ronald E. “The Truth About Myers-Briggs Types.” *Psychology Today*, Sussex Publishers, 21 Feb. 2014.

[6] Zurcher, Anthony. “Debunking the Myers-Briggs Personality Test.” *BBC News*, BBC, 15 July 2014.

[7] W. Wilbur “The automatic identification of stop words”, Journal of Information Science, 1992.

[8] V. Bala “Stemming and Lemmatization: A comparison of Retrieval Performances”, Lecture notes on Software Engineering 2 (3): 262-267.

[9] B. Yan, X. Cheng, F. Yang, L. Yao “Research on EDA technology and its related issues”, International Conference On Computer Design and Applications, 2010.

[10] IBM Cloud Education. “What Is Exploratory Data Analysis?” *IBM*, www.ibm.com/cloud/learn/exploratory-data-analysis.

[11] “Personality Types.” *16 Personalities*, www.16personalities.com/personality-types.

[12] K. Hatam, M. Jaf, H. Az and H. Na “How should we report the variation of a study data in a biomedical literature?”, Iranian Journal of Public health, 2018.

[13] Brownlee, Jason. “How to Encode Text Data for Machine Learning with Scikit-Learn.” *Machine Learning Mastery*, 27 June 2020.

[14] “1.9. Naive Bayes.” *Scikit Learn, https://scikit-learn.org/stable/modules/naive_bayes.html*

[15] Natekin, Alexey, and Alois Knoll. “Gradient Boosting Machines.” *Methods Articles*, Frontiers in Neurobotics, 4 December 2013.

[16] “Random Forests Leo Breiman and Adele Cutler.” *Random Forests - Classification Description*, www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.