ENGINEERING
W Booth School of Engineering
Practice and Technology

McMaster
University

SEP 788 – Neural Networks and Development Tools

Personality Prediction

Submitted by
Arshdeep Singh- ID 400328272
Kunal Garg- ID 400387739

Date
February 20th, 2022

Problem Statement: With the increasing accessibility of the internet across the world, the use of social media is at an all-time high and is growing exponentially. According to Datareportal's article on 'Global Social Media Stats' (2022), "there are more than 4.62 billion social media users around the world in January 2022, equating to 58.4 percent of the total global population."

People respond differently to everyday situations based on their cognitive abilities and accordingly express their opinions on social media. In light of this, the huge amount of data available in the form of tweets, Instagram posts, Facebook posts, etc can be used to predict the personalities of individuals. Such information can be extremely useful, especially for the purposes of talent management. Recruiters and hiring managers can use these personality prediction models to gain insightful information about potential employees and hire the right talent for their companies. Moreover, the data obtained from personality prediction can also be used for the counselling of individuals, online marketing, personal recommendation systems, and bank credit scoring systems (Christian, 2021).

Through this project, we aim to identify the personality type of a person from a given social media post on the internet. We will use Myers–Briggs Type Indicator (MBTI) which is based on C. G. Jung's 'Theory of Psychological Type' and predicts personality types based on a person's preferences and decisions (Myers, 1998). It takes into account four dichotomies: introversion or extraversion, sensing or intuition, thinking or feeling, judging or perceiving, and hence, identifies 16 different personality types.

About the Dataset: The dataset is taken from Kaggle (MBTI Myers-Briggs Personality type Dataset). It contains 8600 rows of data and each row contains a person's:

- Type (This person's 4 letters MBTI code/type)
- A section of each of the last 50 things they have posted (Each entry separated by "|||" (3 pipe characters))

Comparison between Classification and Regression: Classification and Regression are both supervised learning approaches. The difference is that classification algorithm is used to predict discrete values such as spam or not spam, male or female, etc. A classic example of such a problem is the identification of whether an email is a spam or not. This is a binary classification problem since there are two target classes. When there are more than two classes, it is known as multi-class classification. Examples of classification algorithms include Support Vector Machine, Naive Bayes, K Nearest Neighbours, Random Forest, etc.

On the other hand, regression algorithms are the ones where the prediction is a continuous value such as the age, height, price of an article, etc. An example of a regression problem is predicting the temperature of a city. Based on the previous data, the model must be able to predict the temperature of future days which can be any numeric value. Examples of regression algorithms include simple linear regression, multiple linear regression, support vector regression, decision tree regression, etc.

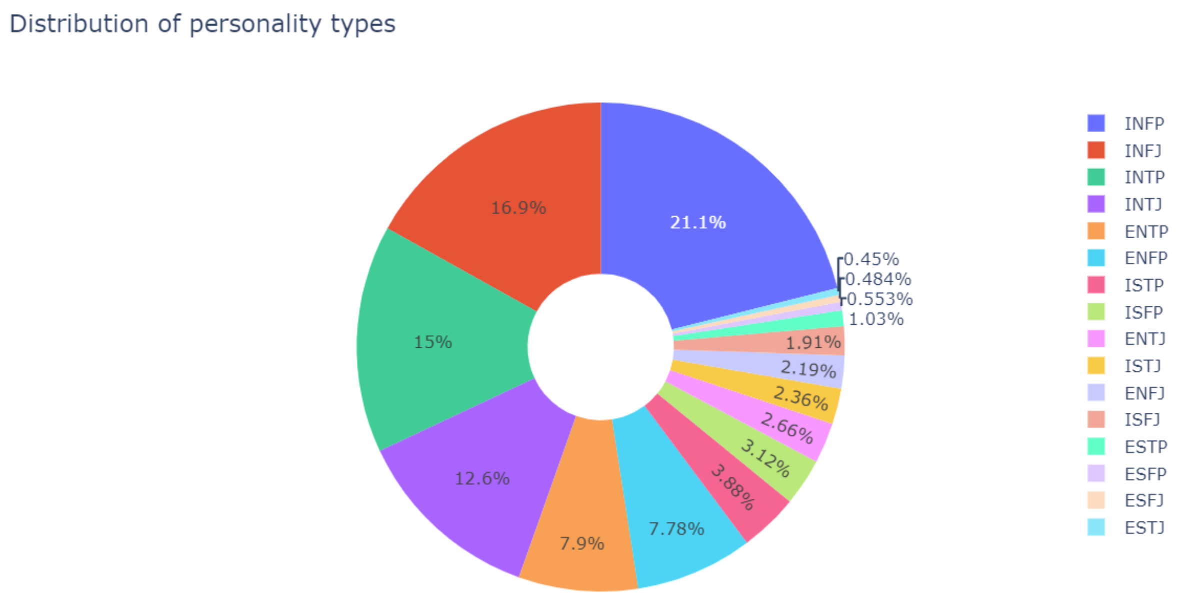
The project dataset has 16 target labels. The prediction of our machine learning model is supposed to be one of the 16 classes which signify a discrete output. Therefore, we can say that personality prediction using the MBTI dataset is an example of a classification problem.

Literature Review and Domain Knowledge: In recent years, there has been much research in the domain of personality prediction. Due to the availability of a large amount of data on social media, most researchers have used it to make predictions about a person's traits. Not only textual data but, some have also attempted to predict personalities based on the images or emoticons that users post in their social media feeds. To categorize the personalities, the most commonly used taxonomies are the MBTI instrument and the Big Five Personality Trait Model also known as the OCEAN model which includes openness, conscientiousness, extraversion, neuroticism, and agreeableness (Songqiao, 2020). In addition, different classification approaches have been used for the prediction model such as the Support Vector Machine, Naive Bayes, Logistic Regression, or deep learning approaches such as LSTM with additional NLP features such as sentiment analysis.

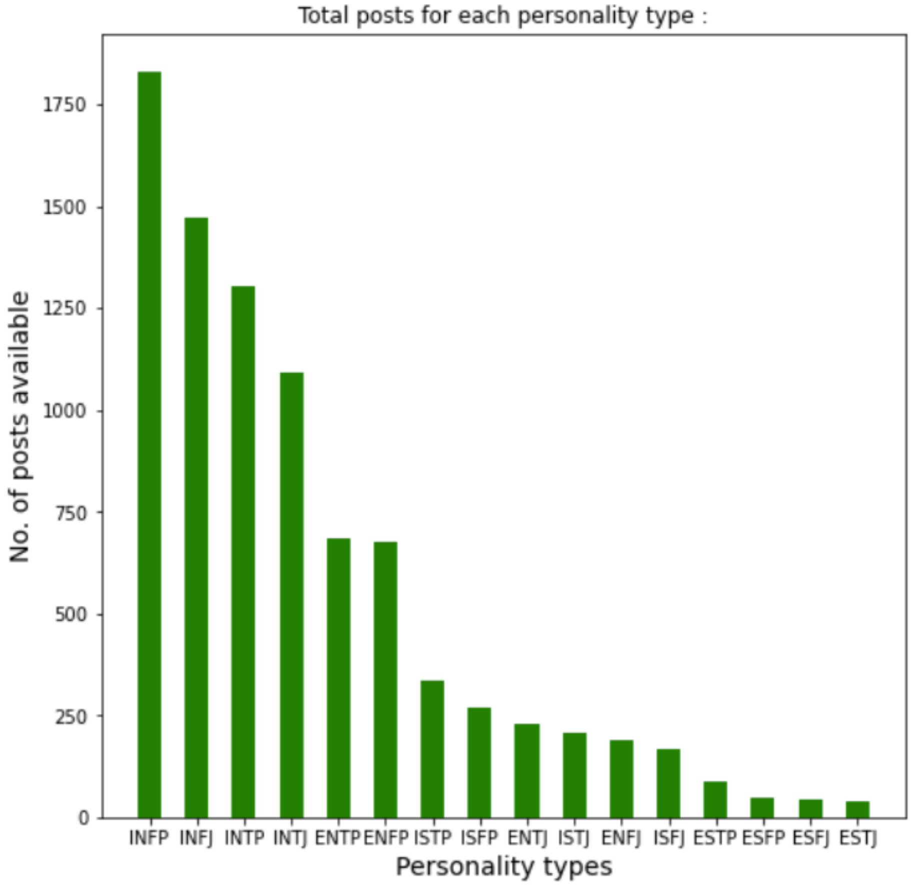
1. Christian et al. proposed a ‘multi-model deep learning architecture’ using Bidirectional Encoder from Transformer (BERT), A Robustly Optimized BERT Pretraining Approach (ROBERTA), and a Generalized Autoregressive Pretraining for Language Understanding (XLNet). With this model, they have used NLP (sentiment analysis, NRC emotion lexicon database) to analyse the features.
2. In another research (Cui, 2017), classification techniques Naive Bayes, SVM, and deep learning method (encoder-decoder method using LSTM and feed-forward network) have been used together with NLP. Out of all the approaches, the deep learning method is observed to be the best with the highest accuracy.
3. The way people behave or express themselves on social media reflects their personalities and thought processes. Alam et al. studied this using data from Facebook statuses of 250 users. In their analysis, Sequential Minimal Optimization (SMO) for SVM, Bayesian Logistic Regression (BLR), and Multinomial Naive Bayes (MNB) techniques were compared. The evaluation metrics were precision, recall, and F1 score, and MNB was observed to perform better than SMO and BLR.

Visualizing the data:

1. **Distribution of data based on different personality types:**



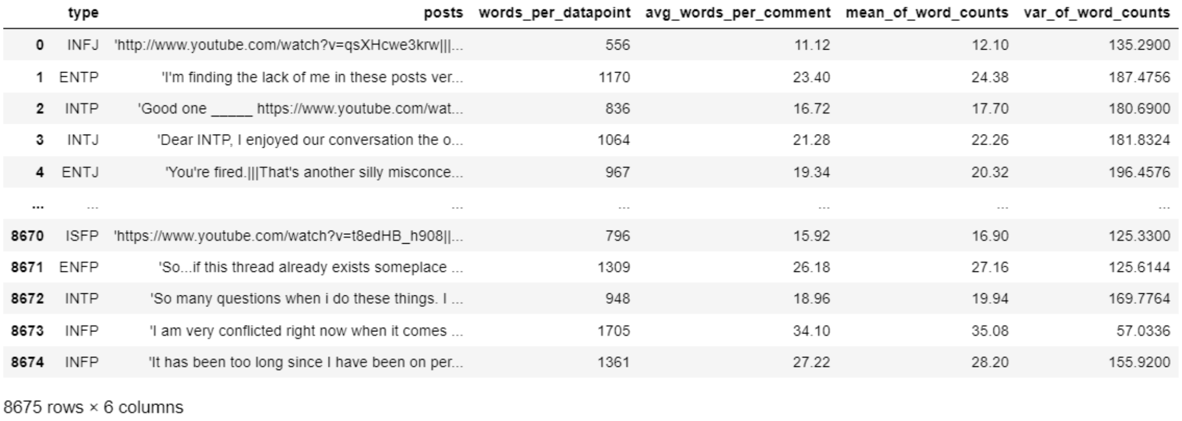
2. **Bar Distribution of the number of posts available for each personality type:**



<u>Inferences:</u>

1. Above we can see that there is great unbalance in Introvert/Extrovert and Intuition/Sensing pairs.
2. Whereas Feeling/Thinking and Perception/Judgment pairs are quite balanced.

3. **Mean and Variance of words for each person for all his 50 comments in the Dataset:**

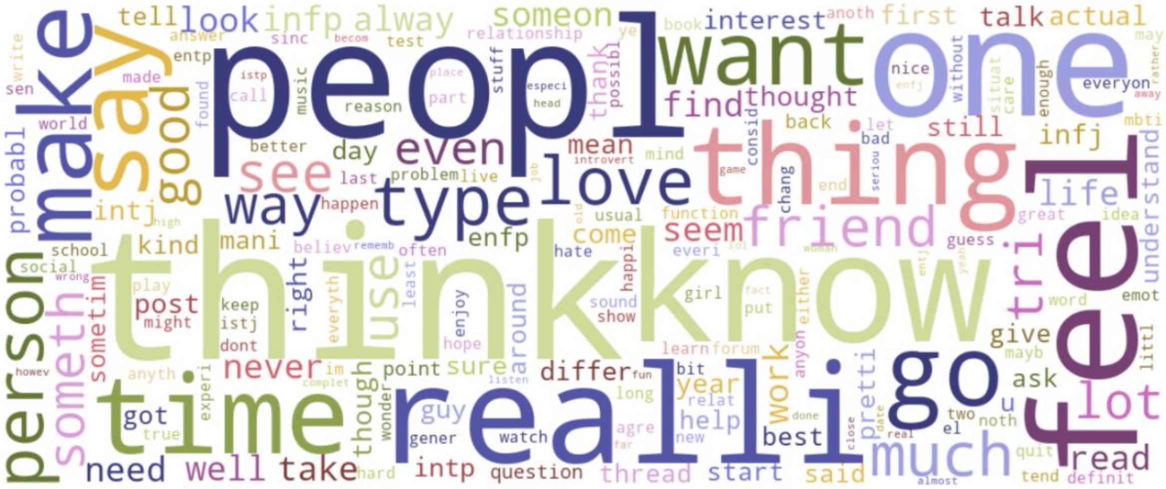


<table>
<tr><th></th><th>type</th><th>posts</th><th>words_per_datapoint</th><th>avg_words_per_comment</th><th>mean_of_word_counts</th><th>var_of_word_counts</th></tr>
<tr><td>0</td><td>INFJ</td><td>'http://www.youtube.com/watch?v=qsXHcwe3krw|||...</td><td>556</td><td>11.12</td><td>12.10</td><td>135.2900</td></tr>
<tr><td>1</td><td>ENTP</td><td>'I'm finding the lack of me in these posts ver...</td><td>1170</td><td>23.40</td><td>24.38</td><td>187.4756</td></tr>
<tr><td>2</td><td>INTP</td><td>'Good one _____ https://www.youtube.com/wat...</td><td>836</td><td>16.72</td><td>17.70</td><td>180.6900</td></tr>
<tr><td>3</td><td>INTJ</td><td>'Dear INTP, I enjoyed our conversation the o...</td><td>1064</td><td>21.28</td><td>22.26</td><td>181.8324</td></tr>
<tr><td>4</td><td>ENTJ</td><td>'You're fired.|||That's another silly misconce...</td><td>967</td><td>19.34</td><td>20.32</td><td>196.4576</td></tr>
<tr><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td></tr>
<tr><td>8670</td><td>ISFP</td><td>'https://www.youtube.com/watch?v=t8edHB_h908||...</td><td>796</td><td>15.92</td><td>16.90</td><td>125.3300</td></tr>
<tr><td>8671</td><td>ENFP</td><td>'So...if this thread already exists someplace ...</td><td>1309</td><td>26.18</td><td>27.16</td><td>125.6144</td></tr>
<tr><td>8672</td><td>INTP</td><td>'So many questions when i do these things. I ...</td><td>948</td><td>18.96</td><td>19.94</td><td>169.7764</td></tr>
<tr><td>8673</td><td>INFP</td><td>'I am very conflicted right now when it comes ...</td><td>1705</td><td>34.10</td><td>35.08</td><td>57.0336</td></tr>
<tr><td>8674</td><td>INFP</td><td>'It has been too long since I have been on per...</td><td>1361</td><td>27.22</td><td>28.20</td><td>155.9200</td></tr>
</table>

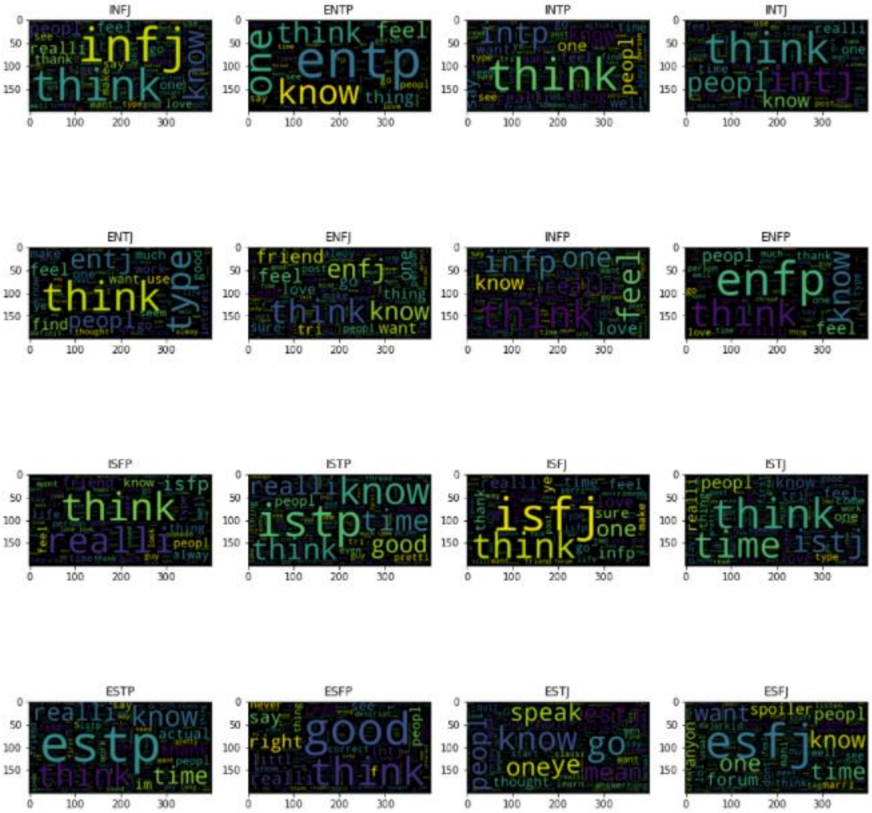
8675 rows × 6 columns

Inference:
Words per data point highly varies, but average word per comment by each person remains between 10-35 word/comment**.**

4. **Word Cloud (shows the most frequently occurring words Bigger) of processed data:**



5. **Word Cloud for each Personality Type for processed data:**



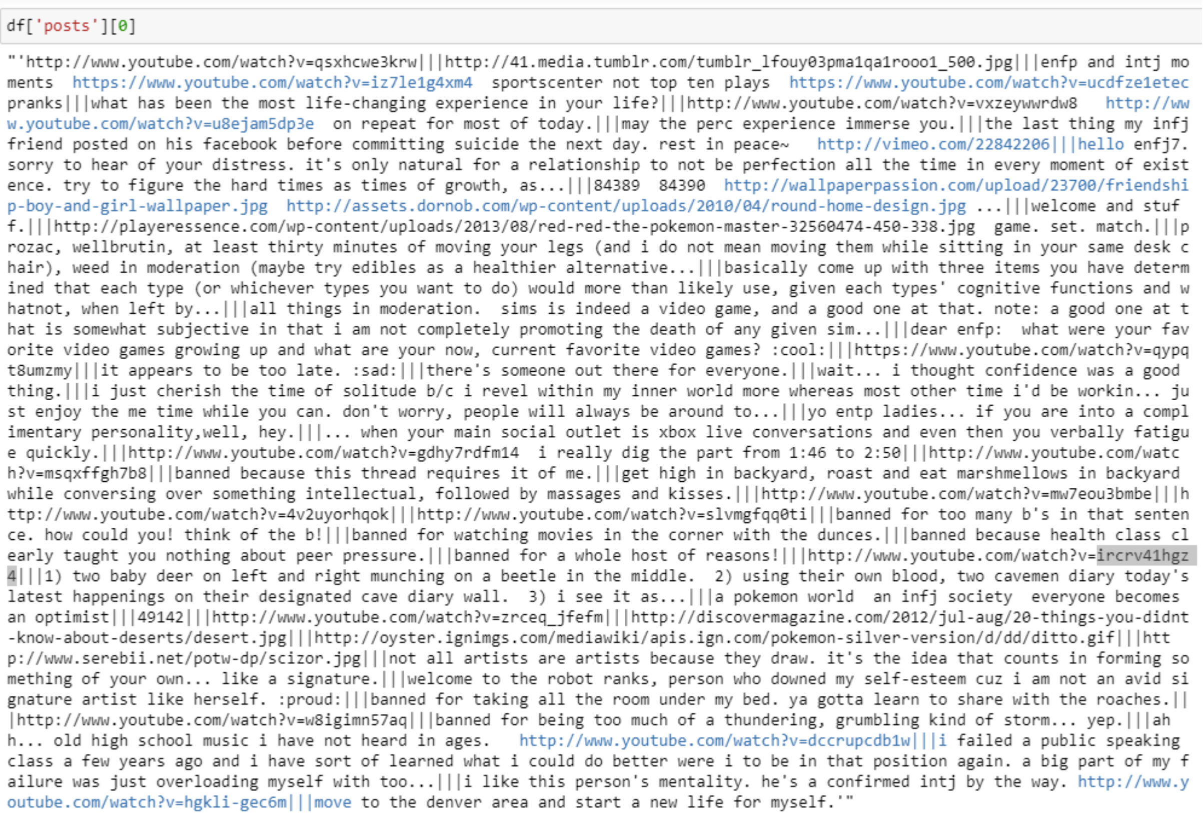
Inference: Each personality type has some dominant words in its data points that define the personality traits.

Data Pre-Processing:

1. Removal of Non-Words
2. Removal of Punctuations and Stop Words
3. Selective word-removal—might cause machine to cheat
4. Lemmatization—brings words to its root form

The snapshots shown below describes the transition of the first datapoint of the dataset from raw data to processed data after performing the pre-processing:

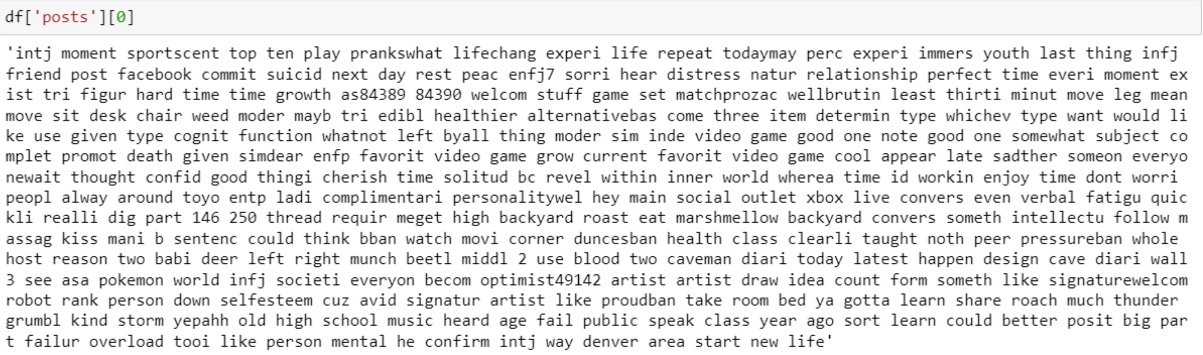
FROM:



```
df['posts'][0]
```

```
"'http://www.youtube.com/watch?v=qsxhcwe3krw|||http://41.media.tumblr.com/tumblr_lfouy03pma1qa1rooo1_500.jpg|||enfp and intj mo
ments  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://ww
w.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 exist
ence. try to figure the hard times as times of growth, as...|||84389  84390  http://wallpaperpassion.com/upload/23700/friendshi
p-boy-and-girl-wallpaper.jpg  http://assets.dornob.com/wp-content/uploads/2010/04/round-home-design.jpg ...|||welcome and stuf
f.|||http://playeressence.com/wp-content/uploads/2013/08/red-red-the-pokemon-master-32560474-450-338.jpg  game. set. match.|||p
rozac, wellbrutin, at least thirty minutes of moving your legs (and i do not mean moving them while sitting in your same desk c
hair), weed in moderation (maybe try edibles as a healthier alternative...|||basically come up with three items you have determ
ined that each type (or whichever types you want to do) would more than likely use, given each types' cognitive functions and w
hatnot, when left by...|||all things in moderation.  sims is indeed a video game, and a good one at that. note: a good one at t
hat is somewhat subjective in that i am not completely promoting the death of any given sim...|||dear enfp:  what were your fav
orite video games growing up and what are your now, current favorite video games? :cool:|||https://www.youtube.com/watch?v=qypq
t8umzmy|||it appears to be too late. :sad:|||there's someone out there for everyone.|||wait... i thought confidence was a good
thing.|||i just cherish the time of solitude b/c i revel within my inner world more whereas most other time i'd be workin... ju
st enjoy the me time while you can. don't worry, people will always be around to...|||yo entp ladies... if you are into a compl
imentary personality,well, hey.|||... when your main social outlet is xbox live conversations and even then you verbally fatigu
e quickly.|||http://www.youtube.com/watch?v=gdhy7rdfm14  i really dig the part from 1:46 to 2:50|||http://www.youtube.com/watc
h?v=msqxffgh7b8|||banned because this thread requires it of me.|||get high in backyard, roast and eat marshmellows in backyard
while conversing over something intellectual, followed by massages and kisses.|||http://www.youtube.com/watch?v=mw7eou3bmbe|||h
ttp://www.youtube.com/watch?v=4v2uyorhqok|||http://www.youtube.com/watch?v=slvmgfqq0ti|||banned for too many b's in that senten
ce. how could you! think of the b!|||banned for watching movies in the corner with the dunces.|||banned because health class cl
early taught you nothing about peer pressure.|||banned for a whole host of reasons!|||http://www.youtube.com/watch?v=ircrv41hgz
4|||1) two baby deer on left and right munching on a beetle in the middle.  2) using their own blood, two cavemen diary today's
latest happenings on their designated cave diary wall.  3) i see it as...|||a pokemon world  an infj society  everyone becomes
an optimist|||49142|||http://www.youtube.com/watch?v=zrceq_jfefm|||http://discovermagazine.com/2012/jul-aug/20-things-you-didnt
-know-about-deserts/desert.jpg|||http://oyster.ignimgs.com/mediawiki/apis.ign.com/pokemon-silver-version/d/dd/ditto.gif|||htt
p://www.serebii.net/potw-dp/scizor.jpg|||not all artists are artists because they draw. it's the idea that counts in forming so
mething of your own... like a signature.|||welcome to the robot ranks, person who downed my self-esteem cuz i am not an avid si
gnature artist like herself. :proud:|||banned for taking all the room under my bed. ya gotta learn to share with the roaches.||
|http://www.youtube.com/watch?v=w8igimn57aq|||banned for being too much of a thundering, grumbling kind of storm... yep.|||ah
h... old high school music i have not heard in ages.   http://www.youtube.com/watch?v=dccrupcdb1w|||i failed a public speaking
class a few years ago and i have sort of learned what i could do better were i to be in that position again. a big part of my f
ailure was just overloading myself with too...|||i like this person's mentality. he's a confirmed intj by the way. http://www.y
outube.com/watch?v=hgkli-gec6m|||move to the denver area and start a new life for myself.'"
```

TO:



```
df['posts'][0]
```

```
'intj moment sportscent top ten play prankswhat lifechang experi life repeat todaymay perc experi immers youth last thing infj
friend post facebook commit suicid next day rest peac enfj7 sorri hear distress natur relationship perfect time everi moment ex
ist tri figur hard time time growth as84389 84390 welcom stuff game set matchprozac wellbrutin least thirti minut move leg mean
move sit desk chair weed moder mayb tri edibl healthier alternativebas come three item determin type whichev type want would li
ke use given type cognit function whatnot left byall thing moder sim inde video game good one note good one somewhat subject co
mplet promot death given simdear enfp favorit video game grow current favorit video game cool appear late sadther someon everyo
newait thought confid good thingi cherish time solitud bc revel within inner world wherea time id workin enjoy time dont worri
peopl alway around toyo entp ladi complimentari personalitywel hey main social outlet xbox live convers even verbal fatigu quic
kli realli dig part 146 250 thread requir meget high backyard roast eat marshmellow backyard convers someth intellectu follow m
assag kiss mani b sentenc could think bban watch movi corner duncesban health class clearli taught noth peer pressureban whole
host reason two babi deer left right munch beetl middl 2 use blood two caveman diari today latest happen design cave diari wall
3 see asa pokemon world infj societi everyon becom optimist49142 artist artist draw idea count form someth like signaturewelcom
robot rank person down selfesteem cuz avid signatur artist like proudban take room bed ya gotta learn share roach much thunder
grumbl kind storm yepahh old high school music heard age fail public speak class year ago sort learn could better posit big par
t failur overload tooi like person mental he confirm intj way denver area start new life'
```

Choosing the ML Approach: Since this is a classification problem, we will choose classification algorithms. We will try out classic ML algorithms such as SVM, Naïve Bayes as well as deep learning method LSTM. In addition to this, we will also explore the features of NLP such as sentiment analysis to extract features and analyze if using NLP improved the accuracy of predictions or not. At last, the results of all approaches will be compared.

References:

1. https://datareportal.com/social-media-users#:~:text=Kepios%20analysis%20shows%20that%20there,of%20the%20total%20global%20population.

2. Christian, H., Suhartono, D., Chowanda, A. *et al.* Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging. *J Big Data* 8, 68 (2021). https://doi.org/10.1186/s40537-021-00459-1
3. Myers, I. B., McCaulley, M. H., Quenk, N. L., & Hammer, A. L. (1998). The MBTI® Manual: A Guide to the Development and Use of the Myers-Briggs Type Indicator. Palo Alto: Consulting Psychologists Press.
   Accessed from: https://www.tolarisd.org/cms/lib3/TX01000982/Centricity/Domain/27/Myers%20Briggs%20Personality%20Test%20Manual.pdf
4. Alam F, Stepanov EA, Riccardi G. Personality traits recognition on social network Facebook. AAAI Workshop—Technical Report, WS-13-01, 2013.
5. Songqiao Han, Hailiang Huang, Yuqing Tang,
   Knowledge of words: An interpretable approach for personality recognition from social media, Knowledge-Based Systems, Volume 194, 2020, 105550, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2020.105550.
   (https://www.sciencedirect.com/science/article/pii/S0950705120300459)
6. Cui, Brandon and Calvin Qi. “Survey Analysis of Machine Learning Methods for Natural Language Processing for MBTI Personality Type Prediction.” (2017).
   https://www.semanticscholar.org/paper/Survey-Analysis-of-Machine-Learning-Methods-for-for-Cui-Qi/08a3043e30ff342f9a92b438646e05d3eeeef6f4
7. https://saejournal.com/wp-content/uploads/2021/07/Personality-Prediction-Using-Machine-Learning.pdf
8. https://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/
9. https://saejournal.com/wp-content/uploads/2021/07/Personality-Prediction-Using-Machine-Learning.pdf