# MBTI Personality prediction using social media posts

...

## Final Project Presentation

~Ritik Khanna(2018084), Mrinal (2018398), Prutyay Gautam(2018403)

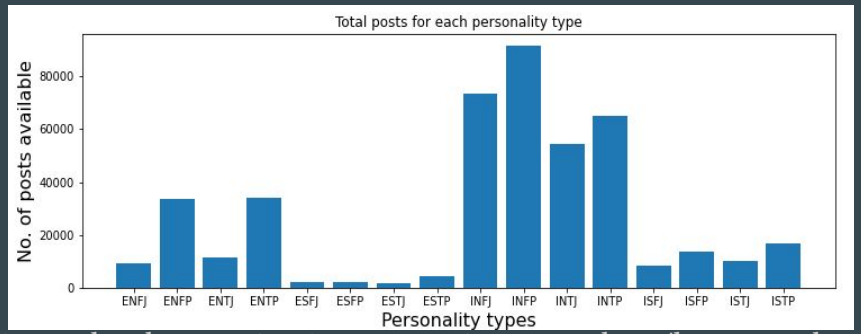
# Problem Statement

**Providing a reliable prediction tool to classify the personality of people using their past 50 social media posts**

- The amount of information being uploaded on social media is increasing by the second.
- This accumulated data can be used to predict the behaviour and personality of the user.
- Can be used by companies to understand their users or by people to exercise better mental/social health
- Four binary personality categories:
  - Introversion/Extroversion
  - Sensing/Intuition
  - Thinking/Feeling
  - Judging/Perceiving

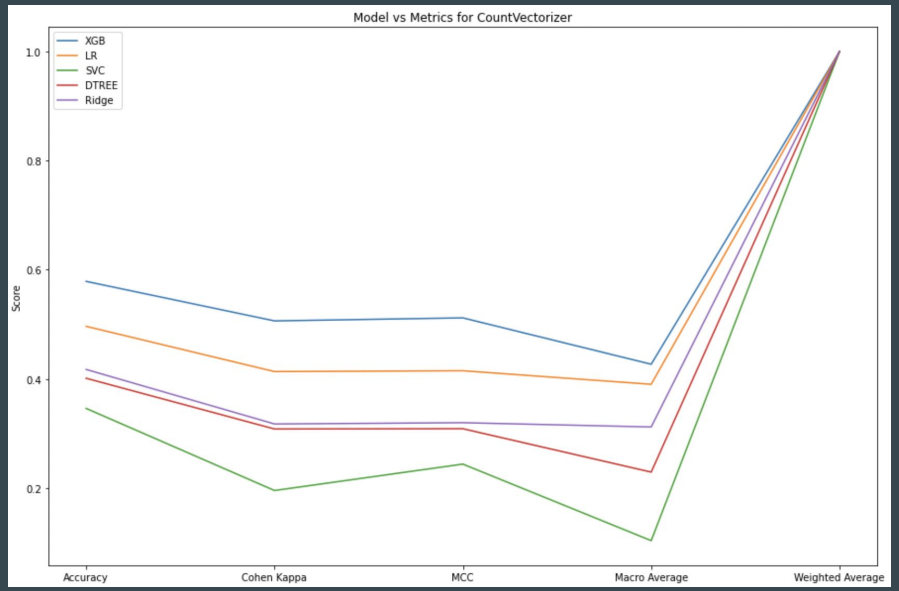
# Approaches

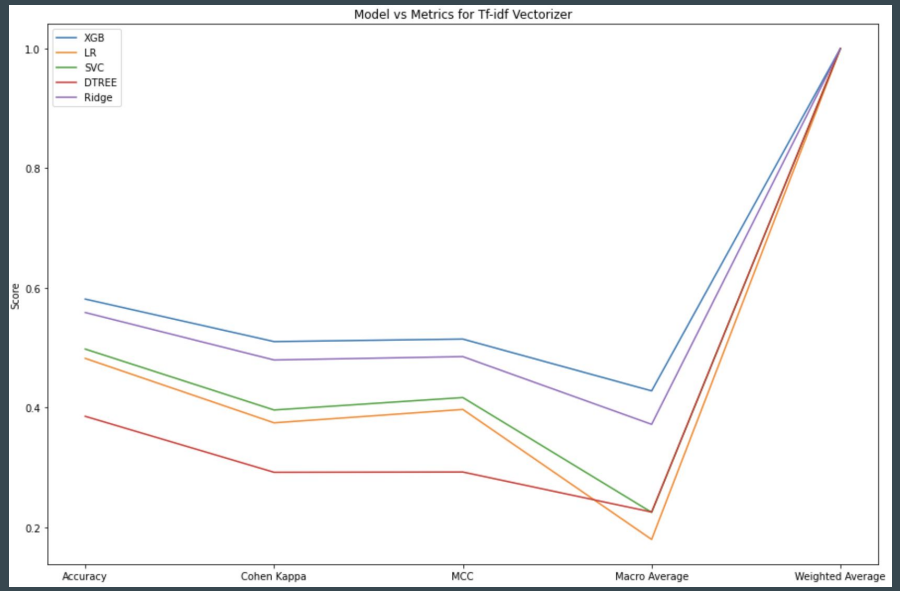
1. 16-Class Classification:




## Steps:

1. Pre-process(Removing links, removing personality codes, lemmatization, removing very short/long words,etc)
2. Finding best performing pair of feature set(CountVectorizer, Tf-idf Vectorizer) and ML model (RF, XGB, LR, SVM,  MLP, DT, KNN, MNB, Ridge, Perceptron)



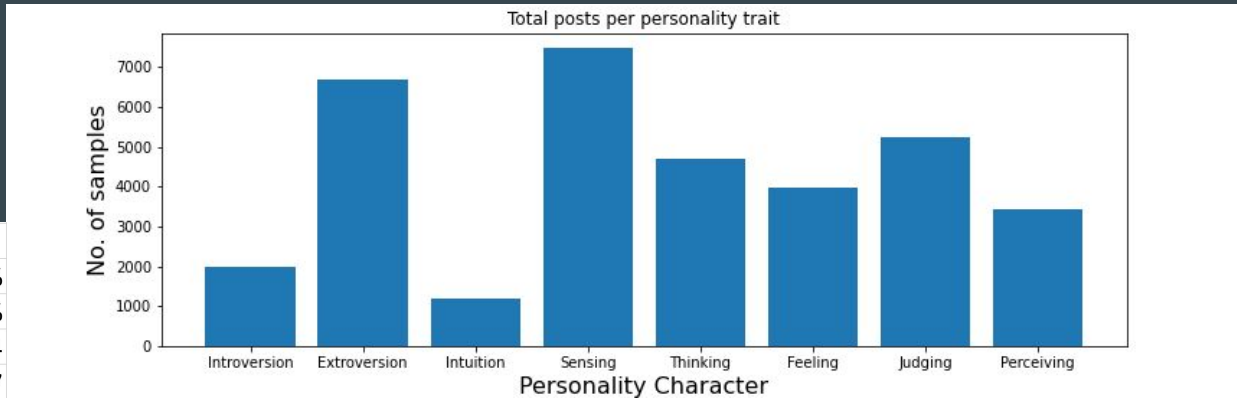



# Approaches

2. Ensemble Approach: Applied ML Model for each personality trait separately

## Steps:

1. Removing trait wise data imbalance (Downsampling)
2. Pre-Processing (Removing Links, case-folding, lemmatization, removing stop words etc. )
3. Finding best performing pair of Feature set (CountVectorizer, Tf-Idf Vectorizer, Word2Vec, BERT) and ML Model (RF, XGB, MLP, LR, SVM, ExtraTree, DT, KNN) **# 32 Possible pairs modelled for each personality trait (5-CV)**
4. Hyper-Parameter tuning of chosen models
5. Error Analysis




Results:

| Trait | Feature Set | ML Model | Accuraccy |
|---|---|---|---|
| IE | Word2Vec | MLP | 70.26% |
| NS | Tf-Idf | LR | 69.71% |
| TF | Word2Vec | MLP | 81.24 |
| JP | Tf-Idf | SVM | 63.87 |

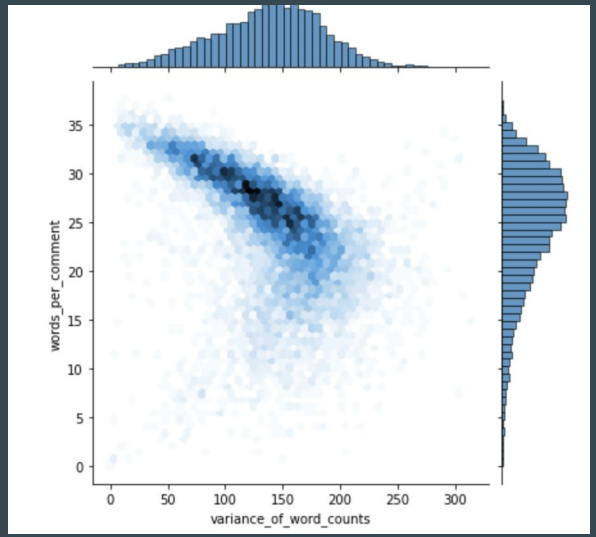
# Dataset and Evaluation Metrics

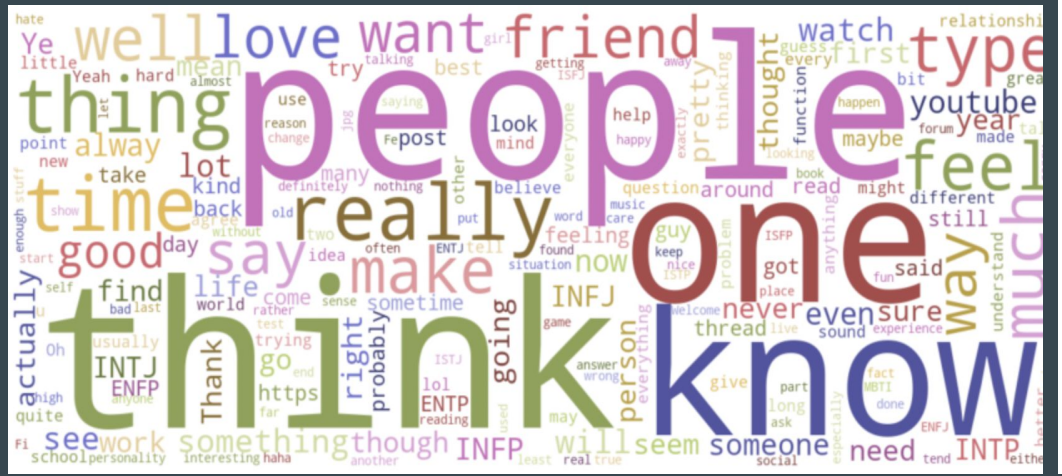
## Dataset:

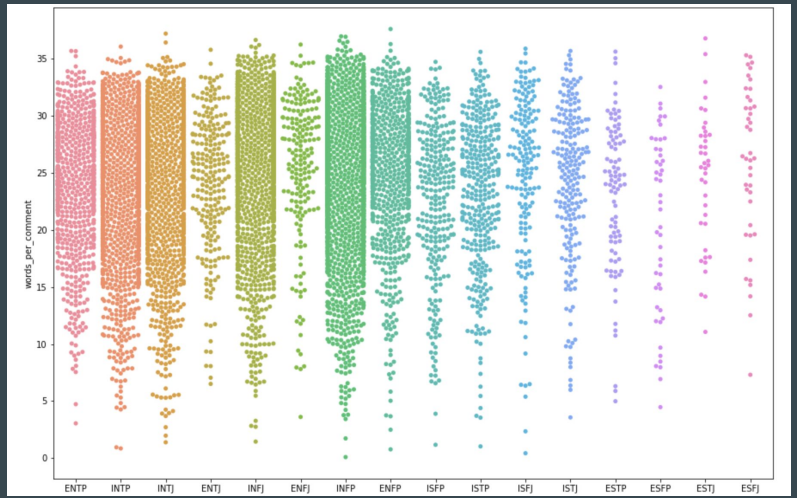
- Social media posts from 8675 Users (Samples)
- Past 50 social media posts per sample
- Each sample labeled a 4 letter code (Introversion(I)/Extroversion(E), Intuitive(N)/Sensing(S), Thinking(T)/Feeling(F), Judging(J)/Perceiving(P))

## Evaluation Metrics:

- All Metrics used: Accuracy, Precision, Recall, F1 score, Cohen's Kappa, MCC, AUC-ROC score, Specificity, Sensitivity



# Main Results

The statistics on the test set for the 4 different curated models are as shown below:-

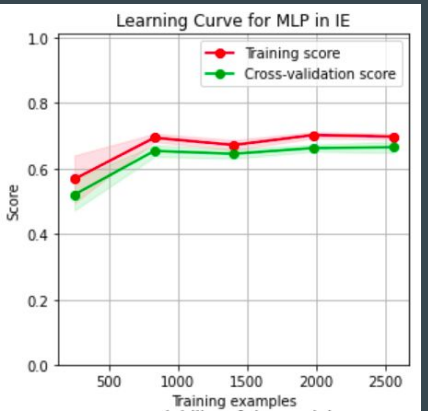
| Personality Trait | Model | Features | Accuracy | SOA Accuracy | Precision | Recall | F1 Score | Specificity | Cohen Kappa Score | MCC Score | AUC ROC Score |
|---|---|---|---|---|---|---|---|---|---|---|---|
| IE | MLP | Word2Vec | 67.37 | 67.6 | 67.5 | 72.5 | 67 | 60.28 | 34.54 | 35.17 | 67.5 |
| NS | LR | TF-IDF | 67.01 | 62 | 64 | 63.5 | 64 | 63.41 | 27.51 | 27.51 | 63.75 |
| TF | MLP | Word2Vec | 80.97 | 77.8 | 79.5 | 80 | 80 | 78.94 | 59.44 | 59.45 | 79.72 |
| JP | SVM | TF-IDF | 62.29 | 67.6 | 62.5 | 62.5 | 62.5 | 59.03 | 24.59 | 24.63 | 62.29 |

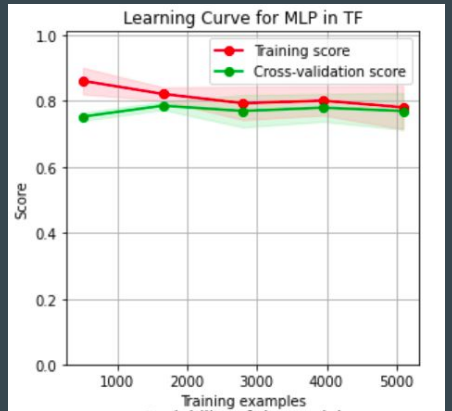
# Analysis

## Evidence of Proper Training

1. Choosing Best model for ensemble technique
   - 32 Possible models tried for each personality trait. Total 32x4=128 models
     - 5-CV was performed for each model and average accuracy obtained over all folds were observed
   - Model (Feature + ML Model) with best score was chosen for each trait.

| Trait | Feature Set | ML Model |
|---|---|---|
| IE | Word2Vec | MLP |
| NS | Tf-Idf | LR |
| TF | Word2Vec | MLP |
| JP | Tf-Idf | SVM |



# Analysis

## Error Analysis

1. Hyper Parameter Tuning
    - Hyperparameter tuning was performed for the best model and feature set for each personality trait.
    - MLP_IE: {'activation': 'relu', 'alpha': 0.001, 'hidden_layer_sizes': (20, 30, 50), 'learning_rate': 'constant', 'solver': 'adam'}
    - LR_NS: {'C': 10, 'penalty': 'l2'}
    - MLP_TF: {'activation': 'tanh', 'alpha': 0.01, 'hidden_layer_sizes': (50, 100, 50), 'learning_rate': 'constant', 'solver': 'adam'}
    - SVM_JP: {'C': 1, 'gamma': 1, 'kernel': 'rbf'}

# Conclusion

We have made a predictive model which is able to predict the 4 MBTI Personality traits with high accuracies:-

| | | |
|---|---|---|
| Introversion/Extroversion | : Accuracy = 67% | Auc-Roc = 0.675 |
| Intuitive/Sensing | : Accuracy = 64% | Auc-Roc = 0.637 |
| Thinking/Feeling | : Accuracy = 81% | Auc-Roc = 0.797 |
| Judging/Perceiving | : Accuracy = 62% | Auc-Roc = 0.622 |