# ML-MAJOR-JUNE-ML063B10

The dataset contains the following fields:

- **_unit_id**: a unique id for user
- **_golden**: whether the user was included in the gold standard for the model; TRUE or FALSE
- **_unit_state**: state of the observation; one of finalized (for contributor-judged) or golden (for gold standard observations)
- **_trusted_judgments**: number of trusted judgments (int); always 3 for non-golden, and what may be a unique id for gold standard observations
- **_last_judgment_at**: date and time of last contributor judgment; blank for gold standard observations
- **gender**: one of male, female, or brand (for non-human profiles)
- **gender:confidence**: a float representing confidence in the provided gender
- **profile_yn**: "no" here seems to mean that the profile was meant to be part of the
- **profile_yn:confidence**: confidence in the existence/non-existence of the profile
- **created**: date and time when the profile was created
- **description**: the user's profile description
- **fav_number**: number of tweets the user has favorited
- **gender_gold**: if the profile is golden, what is the gender?
- **link_color**: the link color on the profile, as a hex value
- **name**: the user's name
- **profile_yn_gold**: whether the profile y/n value is golden
- **profileimage**: a link to the profile image
- **retweet_count**: number of times the user has retweeted (or possibly, been retweeted)
- **sidebar_color**: color of the profile sidebar, as a hex value
- **text**: text of a random one of the user's tweets

- **tweet_coord**: if the user has location turned on, the coordinates as a string with the format "[latitude, longitude]"
- **tweet_count**: number of tweets that the user has posted
- **tweet_created**: when the random tweet (in the text column) was created
- **tweet_id**: the tweet id of the random tweet
- **tweet_location**: location of the tweet; seems to not be particularly normalized
- **user_timezone**: the timezone of the user

We have worked on some questions and provided answers to that such as

Question 1:

What are the most common words used by males and females?

Answer 1:

Males

```
('that', 760),
 ('with', 534),
 ('have', 534),
 ('thi', 473),
 ('just', 465),
 ('your', 429),
 ('like', 391),
 ('they', 351),
 ('what', 303),
 ('when', 283),
 ('time', 258),
 ('from', 258),
 ('will', 257),
 ('love', 252),
 ('about', 251),
 ('make', 242),
 ('know', 202),
 ('look', 200),
 ('peopl', 197),
 ('there', 192),
 ('good', 189),
 ('think', 185),
 ('want', 179),
```

```
 ('need', 171),
 ('follow', 170)
```

## Females

```
('that', 802),
 ('with', 657),
 ('just', 612),
 ('have', 577),
 ('thi', 552),
 ('your', 530),
 ('like', 484),
 ('what', 397),
 ('love', 394),
 ('when', 364),
 ('make', 338),
 ('they', 318),
 ('about', 297),
 ('time', 290),
 ('want', 256),
 ('peopl', 249),
 ('from', 247),
 ('know', 233),
 ('follow', 218),
 ('look', 217),
 ('there', 205),
 ('will', 203),
 ('best', 192),
 ('thank', 189),
 ('here', 185)
```

## Question 2:

## Which Gender made more typos?

## Answer 2:

## Males made more typos.

## Males

```
No. of typos in males:  21413
Total words:  51081
Error rate:  0.41919696168829895
```

## Females

```
No. of typos in females:  21777
```

```
Total words:  53025
Error rate:  0.41069306930693067
```

We have cleaned the data and extracted the features as required to get the best possible results.

Results:

## Naïve Bayes

```
text, fav_number, link_color     61.17764471057884
```

## K-Nearest Neighbors

```
---------------------------------k = 3 ----------------------------------
--------
text, fav_number, link_color     53.243512974051896
---------------------------------k = 5 ----------------------------------
--------
text, fav_number, link_color     53.89221556886228
---------------------------------k = 10 ---------------------------------
---------
text, fav_number, link_color     52.99401197604791
```

Since k = 5 gives us the maximum accuracy

## Support Vector Machine

```
              precision    recall  f1-score   support

           0       0.61      0.76      0.68      1083
           1       0.60      0.44      0.51       921

    accuracy                           0.61      2004
   macro avg       0.61      0.60      0.59      2004
weighted avg       0.61      0.61      0.60      2004

accuracy : 0.6102794411177644
```

- Naive Bayes - 61.17764471057884%
- K-Nearest Neighbors - 53.89221556886228%

- Support vector machines - 61.02794411177644%

After ensembling :

Ensembling – 89.02195608782435%