**For all cases,**

1. **data which has missing values for the target variable ‘gender’ are dropped.**
2. **Data with gender values ‘unknown’ are also dropped.**
3. **Only the data which has gender:confidence = 1 is considered!**

**Original data size: 20050 rows**

**After a) => 19953 rows**

**After b) => 18836 rows**

**Male Profile Count 6194**

**Female Profile Count 6700**

**Brand Profile Count 5942**

**After c) => 13804 rows**

**Now,**

**Male Profile Count 4653**

**Female Profile Count 5367**

**Brand Profile Count 3784**

**Training set size = 0.66, Test set size = Remaining**

**For 5 chosen attributes:**

Input: selected\_attributes = ['text\_cleaned', 'description\_cleaned', 'sidebar\_color', 'link\_color', 'name']

Output: gender

Input data shape (13804, 5)

Output data shape (13804,)

**Naive Bayes**

**0.66489135066**

[[ 932 186 179]

[ 94 1469 288]

[ 116 710 720]]

Stochastic Gradient Descent

0.635918193438

[[ 976 190 131]

[ 143 1361 347]

[ 199 699 648]]

MultiLayer Perceptron

0.655517682147

[[ 925 147 225]

[ 83 1211 557]

[ 108 497 941]]

Random Forest Classifier

0.542617871892

[[ 722 336 240]

[ 79 1428 321]

[ 112 921 535]]

Decision Tree

0.557733276523

[[ 722 336 240]

[ 79 1428 321]

[ 112 921 535]]

K-NN

0.460374946741

[[ 409 690 199]

[ 96 1382 350]

[ 120 1078 370]]

Support Vector Machines

0.389433319131

[[ 409 690 199]

[ 96 1382 350]

[ 120 1078 370]]

**For 4 chosen attributes:**

Input: selected\_attributes = ['text\_cleaned', 'description\_cleaned', 'sidebar\_color', 'link\_color']

Output: gender

Input data shape (13804, 4)

Output data shape (13804,)

**Naive Bayes**

**0.652322113336**

[[ 908 166 175]

[ 78 1447 303]

[ 126 784 707]]

Stochastic Gradient Descent

0.623135918193

[[ 979 144 126]

[ 202 1202 424]

[ 243 630 744]]

MultiLayer Perceptron

0.630166169578

[[ 897 162 190]

[ 94 1266 468]

[ 120 702 795]]

Random Forest Classifier

0.524494016608

[[ 745 304 200]

[ 93 1431 304]

[ 127 1018 472]]

Decision Tree

0.54260758415

[[ 745 304 200]

[ 93 1431 304]

[ 127 1018 472]]

K-NN

0.448444823179

[[ 436 590 223]

[ 147 1243 438]

[ 175 1016 426]]

Support Vector Machines

0.389433319131

[[ 436 590 223]

[ 147 1243 438]

[ 175 1016 426]]

**For a single attribute which combines the words from both the tweet text and description**

Input: Combined\_text (Strings from both the cleaned tweet text and description text are concatenated)

Output: gender

Input data shape (13804, 1)

Output data shape (13804,)

**Naive Bayes**

**0.638900724329**

[[ 893 175 183]

[ 103 1433 339]

[ 161 734 673]]

Stochastic Gradient Descent

0.609288453345

[[ 872 201 178]

[ 114 1349 412]

[ 191 738 639]]

MultiLayer Perceptron

0.604814657009

[[ 828 156 267]

[ 80 1150 645]

[ 129 578 861]]

Random Forest Classifier

0.512566023699

[[ 714 410 165]

[ 96 1469 237]

[ 124 1118 361]]

Decision Tree

0.515551768215

[[ 714 410 165]

[ 96 1469 237]

[ 124 1118 361]]

K-NN

0.428206220707

[[ 442 641 206]

[ 181 1126 495]

[ 198 963 442]]

Support Vector Machines

0.383894333191

[[ 442 641 206]

[ 181 1126 495]

[ 198 963 442]]

**Considering only the tweet text**

Input: ‘text\_cleaned’

Output: ‘gender’

Input data shape (13804, 1)

Output data shape (13804,)

Not imp: Broken into 39753 vectors using countVectorizer

**Naive Bayes**

**0.55112910098**

[[ 604 365 309]

[ 109 1414 318]

[ 162 844 569]]

Stochastic Gradient Descent

0.524499360886

[[ 659 260 359]

[ 168 1046 627]

[ 246 572 757]]

MultiLayer Perceptron

0.527055815935

[[ 600 337 341]

[ 97 1172 572]

[ 175 698 702]]

Random Forest Classifier

0.473797831398

[[ 549 423 351]

[ 91 1210 525]

[ 110 872 563]]

Decision Tree

0.479335321687

[[ 549 423 351]

[ 91 1210 525]

[ 110 872 563]]

K-NN

0.439071154666

[[ 421 663 239]

[ 118 1215 493]

[ 124 996 425]]

Support Vector Machines

0.389007243289

[[ 421 663 239]

[ 118 1215 493]

[ 124 996 425]]

**Considering only the profile description**

Input: ‘description\_cleaned’

Output: ‘gender’

Input data shape (13804, 1)

Output data shape (13804,)

Not imp: Broken into 34396 vectors using countVectorizer

**Naive Bayes**

**0.582658713251**

[[ 819 412 91]

[ 155 1324 374]

[ 196 731 592]]

Stochastic Gradient Descent

0.562207072859

[[ 993 182 147]

[ 372 1063 418]

[ 304 632 583]]

MultiLayer Perceptron

0.541542394546

[[ 674 470 178]

[ 89 1268 496]

[ 113 806 600]]

Random Forest Classifier

0.496168992235

[[ 869 251 132]

[ 346 1229 257]

[ 282 857 471]]

Decision Tree

0.519173412867

[[ 869 251 132]

[ 346 1229 257]

[ 282 857 471]]

K-NN

0.444397102684

[[ 550 585 117]

[ 307 1353 172]

[ 233 1194 183]]

Support Vector Machines

0.390285470814

[[ 550 585 117]

[ 307 1353 172]

[ 233 1194 183]]

Top 20 most frequent words used by Men

'like', 300

'get', 287

"i'm", 178

'new', 178

'one', 177

"don't", 169

'go', 159

"it's", 158

'time', 157

'love', 149

'see', 137

'know', 134

'people', 133

'make', 124

'good', 122

'think', 120

'got', 120

'still', 118

'back', 117

'best', 116

dtype: int64

Top 20 most frequent words used by Women

'like', 370

"i'm", 318

'get', 299

'one', 237

'go', 224

'love', 208

"don't", 197

"it's", 176

'day', 171

'time', 169

'people', 167

'know', 160

'new', 160

'got', 154

'want', 140

'make', 139

'still', 135

'back', 135

'best', 132

'makes', 127

dtype: int64

Top 20 most frequent words used by Brands

'weather', 2278

['get', 1159

'updates', 1142

'channel'] 1137

'new', 175

'get', 134

'us', 128

'like', 124

'love', 108

'via', 100

'one', 100

'best', 85

'see', 78

'makes', 75

'look', 74

'first', 74

'people', 73

'year', 72

'make', 70

'week', 67

dtype: int64