CSCI 720: Big data analytics project report
Gender prediction based on Twitter profile

Business Understanding:

"Commercials that appear in social networking sites, as with advertisements in all other forms of media, can be a source of gender stereotyping."[1]. How likely are women to be interested in automobiles? Would men be interested to view a commercial of a jewellery shop or beauty products? Answer to all these questions could be achieved by analyzing a user's social media profile. The business understanding of this project constitutes the idea of audience targeting. This concept is primarily used by advertisers to create custom advertisements for users based on their gender.

Our goal in this project is to identify the gender of the Twitter user given some data about the user. This would make advertisements more profitable as they could be targeted to the proper users. The analysis could be performed based on the user's tweet, description associated with profile, color backgrounds used on profile etc. The project is an example of classical two-class classification problem.

Data Understanding:

The most important aspect of any data mining project is the data. If the data does not have sufficient attributes and instances then the analysis results are usually very poor and such a poor model is of no use to the business process. For this project data has been obtained from the following website: https://www.kaggle.com/crowdflower/twitter-user-gender-classification. Kaggle is a platform where researchers and companies post their data making it public for the analysis. The user community use this data and devise models to be run on this data. There are frequent competitions where the users with best model or solution are awarded for their implementations.

The dataset on the website consists of 26 attributes and around 20000 records. The attributes are a mix of numeric, categorical and boolean data. Although one might be under impression that this is a great amount of data but actually it is not a sufficient amount for a text mining related task. And apart from that not the entire 20,000 records are usable for building model. The data understanding is the most crucial step in an analysis task.

There might be some attributes which do not contribute to analysis as there might not be a pattern in it. Some types of such data are monotonic variables, constants and sparse variables. Other than that there are some features which make it difficult to perform analysis. The data contains profile image attribute which provides the link to the user's profile image. Although this could be used for analysis but the links cannot be used directly. We first need to extract the image from the given link and then perform feature detection on it. The original size of the picture is very small and thus might not provide much of information.

There are many features in dataset which have been included considering a different analysis task. The original data consisted of results from an experiment which involved having

people look at the profile and asking them to predict gender based on details. So the data has attributes which have information regarding this task. There are certain gold standard columns which correspond to the feedback collected from some standard certified users. We decided that these attributes were more or less associated with the other task and hence decided to set them aside.

Finally we decided to work with only the following attributes:
Name: User name of the associated profile
Description: User provided self description on the profile
Text: Latest tweet from the user
Link_color: Color on the profile page
Sidebar_color: Color associated with the sidebar
Tweet_count: Number of tweets the user has posted

Data Preparation:

The next major step is data preparation and cleaning. Not all the data is useful for analysis. So we may need to do some cleaning and remove certain records from the data. The first and the most essential task in this phase was to remove records with gender as brand. Now there is one thing about twitter which needs to be taken into account. Twitter supports profiles' for brands. As we are more interested in the gender of profile, having brand related profiles would make no sense. So we decided to remove these entries. After these entries were removed, we were left with only 12,000 records in the data. There are some records where the gender is mentioned as unknown and some of them have a NA value in it. All of these records are of no use as having the target variable is important for classification task.

The color related attributes in the dataset require some cleaning. The color values are in hexadecimal number system. These represent the RGB values of that corresponding color. These need to be 6 characters long to be converted to RGB equivalent. Therefore some padding is needed to make it in proper format. After this the padded hexadecimal values can be converted to RGB tuple using matplotlib library. After conversion we get a tuple value representing the color value of the three color channels namely Red, Green and Blue. Depending on the maximum channel value the record is assigned a Red, Green or a Blue value. Same process is followed for the other color based attribute.

Example:
0084B4 is a hexadecimal value with more Blue component as the last two hex digits are higher in value. Hence this value becomes Blue.

The tweet count field is a numeric attribute, hence it needs normalization. We performed min-max normalization on this attribute and used the normalized values in the model. The text attributes Name, Description and Text need the most amount of cleaning. Each string for these attributes for all records needs some preprocessing. Following is a list of actions we performed on each of the strings:

1) Converting to lower-case

2) Remove extra whitespaces
3) Remove punctuations
4) Remove hashtags
5) Remove emoticons
6) Remove html characters
7) Remove numbers
8) Remove stopwords
9) Perform stemming

Out of the above tasks, stopword removal and stemming are important. Stopwords are the words present in any language which occur very frequently but are contextually not very important. Consider the words ‘the’, ‘is’, ‘and’, ‘who’, etc. These occur very frequently in day to day conversations but they do not point to any gender directly or indirectly. They are more responsible for connecting other words or are needed for grammar purposes. Hence these words can be removed from the data.

Stemming replaces a word by its root word. Consider the word ‘play’: the words ‘player’, ‘played’, ‘playing’ are derived from ‘play’. They all mostly have a similar meaning. Hence we can do such stemming on other words to reduce the data size. For stopword removal and stemming we made use of nltk library in Python. For the stopword removal to work, it is necessary to have the nltk stopwords downloaded to the computer locally. After all these steps we get clean data which can be used for analysis. We stored this data in clean.csv file. The Python file cleanData.py performs the cleaning tasks as mentioned above and creates the new csv file.

Modelling:

The new csv file can now be used to perform analysis and create a model. Sklearn library in python provides various ways to use classification algorithms. But the text data cannot be used for analysis directly. We need some way to convert this data into a numeric format without any loss of information. This is where the concept of Bag of Words comes into picture. This method basically creates a dictionary of all the words in the entire data and creates a matrix representation. So basically the data is converted into a sparse matrix containing 1 or 0 depending on the condition that word is present or not. This way we would end up with columns equal to the number of words in the entire data and 1/0 as values in these columns for each row.

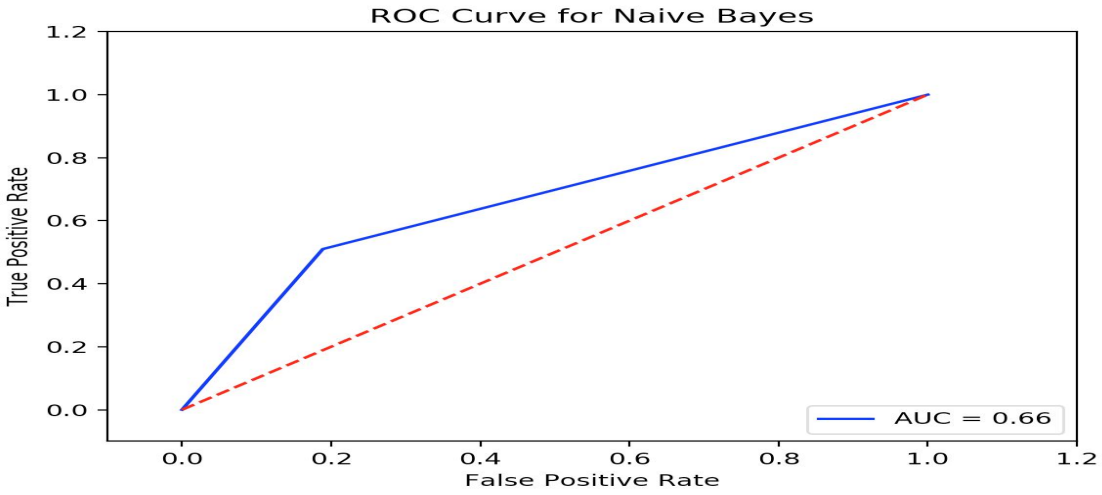
This concept is supported in scikit learn by the use of Tf-idf transform. This method goes one step further and creates the bag of words and assigns a weight based on the number of occurrences. It basically gives more priority to words which appear less frequently in the data.

Once we have this structure the classification task becomes very easy. We can simply pass this bag of words structure to the classification algorithms after performing some split to create test and training data. The algorithms now take very less time to train the model as compared to using the raw data and also provide promising results which we discuss in the next section. We used two algorithms for classification namely Random Forest and Naive Bayes. Both provided pretty identical results.

Result:
Following are the result metrics associated with the classification algorithm:

1) **Naive Bayes**

Metrics/ Data Split	Accuracy	Precision	Recall	F-score
80:20	69.72	71.58	51.00	59.56
70:30	67.45	74.00	50.08	59.73
60:40	66.18	71.85	48.13	57.64

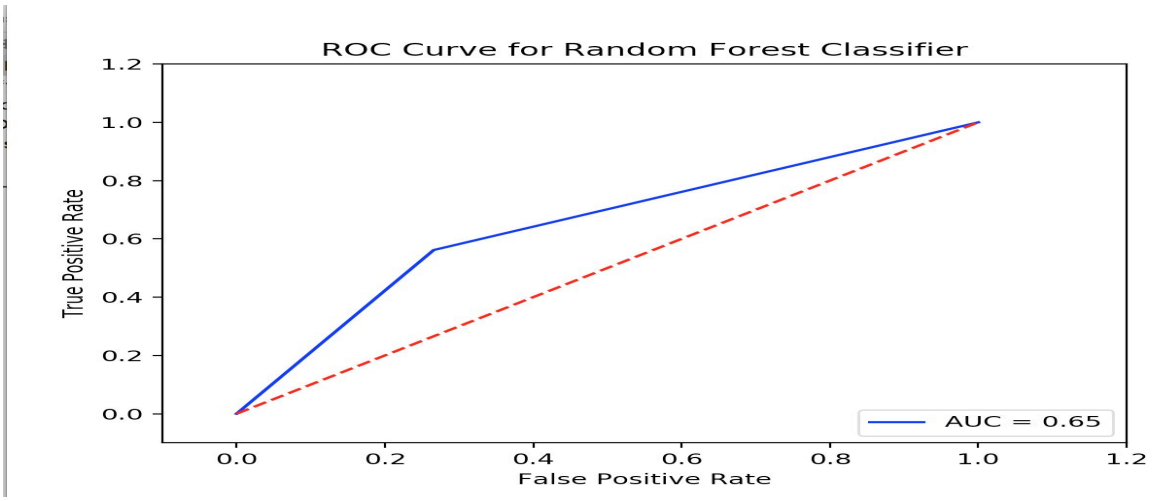


As it can be seen from the table above, the 80:20 split provides a good value on all the metrics. The recall value is not that great but accuracy and precision have decent numbers. The area under the curve for ROC is 0.66.

Now let us look at results from the Random Forest algorithm. These have been obtained by using 20 trees in the forest and using ‘gini’ criterion. Although we can use different numbers and criterion for these two metrics but the results do not vary significantly. Following are the metrics associated with the result.

Metrics/ Data Split	Accuracy	Precision	Recall	F-score
80:20	65.91	66.73	56.09	60.95
70:30	63.55	65.28	53.63	58.89

60:40	62.31	62.37	54.47	58.15



It can be clearly seen that again 80:20 split gives good results. The area under the curve metric matches up with the value for Naive Bayes but there is a significant increase in recall metric if random forest is used. The other metrics for random forest algorithm are pretty decent as well.

We found the colors used by Men and Women and following were the results of the task. Men used colors in the blue channel of RGB whereas women used colors in the red channel. We also found some words used by men and women:

Women used the following words frequently: people, love, last, new, best etc
Men used the following words frequently: don't, one, time, good, think etc

There are also some words used by both genders which do not actually help in identifying the gender, this is where the color would be helpful.

Conclusion:

The main task now remains to identify which algorithm would be actually used for modelling the problem. Depending on above results we can say that Random Forest is a better choice because it offers more consistent results. It has a good recall value and more or less similar values for other metrics as compared to Naive Bayes.

There are a couple of things which could be done to improve the accuracy of the model. The first and foremost of them is having more data to work with. The 10-12 thousand records are not sufficient to train such a model. The other thing would be to supply a list of known female and male words to train the model. Also we use only three colors in the model, in reality having more colors would be an ideal scenario as the model could have more to learn.

The brand related tweets could be kept in the dataset to make this a three class problem instead of a two class one. Apart from a few shortcomings such a system would be ideal for the given task and an improvement using the above factors will definitely lead to better results. The most important result in the project is an 12% increase in accuracy over the unclean data. Also the data reduces to 1/4th of the original size and also provides good results. This is a major achievement from the project.

Team Members:

1) **Akshay Renavikar(asr5422@rit.edu)**
2) **Gaurav Gawade(gdg6776@rit.edu)**

References:

[1]http://www.huffingtonpost.com/suren-ramasubbu/does-gender-matter-on-soc_b_7591920.html

[2] https://www.crowdflower.com/using-machine-learning-to-predict-gender/