Project Proposal

Twitter User Gender Classification

Project Title:

Twitter User Gender Classification

Data Mining:

In data mining, we will focus on using Predictive Analysis in the field of Social Network Analysis.

Challenges Involved:

There are lots of challenges involved in the project. The first major challenge comes in the form of choosing the dataset. Having a dataset with a biased split on an attribute tends to skew the model and forces it to predict the majority class. So, it is important to use an unbiased dataset for classification. Next challenge involved is data cleaning. Looking for missing values or removing attributes which are not essential to our data model is preliminary tasks. Feature selection is another challenge which we will have to address as we progress along.

Then, choosing the split between training and testing data is also an important challenge that must be tackled since we do not want to underfit or overfit the model while training the data. Once the correct split is chosen, achieving a high accuracy rate is important too.

Description:

The dataset contains various attributes about a Twitter user such as a username, profile description, tweet, etc. The dataset contains about 20000 rows each with a unique id and username. The dataset contains missing values in the description attribute. The main aim of the project is to determine whether the profile belongs to a male, female or a brand. Our model tries to predict the gender of the user based on the parameters mentioned above. Our model will determine what words strongly separate the males from females.

Data:

The dataset we will be using is Twitter User Gender Classification and has been downloaded from Kaggle [1]. It has 20000 rows with a unique id for each row.

Evaluating the model:

We will be using the Hold-out method to split the data into training, testing and validation data sets. The results obtained from our model will be compared against the values given in the dataset for the target variable. We hope to achieve an accuracy of about 90% using our data model.

Selecting the project:

One of the main reasons for selecting this project is that we wanted to delve deeper into the field of Social Network Analysis. Social Network Analysis finds applications in areas such as Epidemic Detections, Crime detection, Marketing, etc. Choosing Twitter data as our focus was down to the fact that data is openly available. In terms of predictive analysis that we will be doing, it is because we want to be able to predict whether tweets and profile descriptions can help determine the gender or distinguish between a human and a brand. We hope to discover what words are strong predictors of the gender.

References:

[1] - https://www.kaggle.com/crowdflower/twitter-user-gender-classification

Team Members:

Akshay Karki avk1063@rit.edu

Sahana Murthy sam2738@rit.edu