**Data Science and Big Data Analytics Project**

Twitter user gender classification

Information systems department

First semester 2018\2019

Instructor: D. Sahar Ismail

Students name: Manar Abdelkarim

Table of Contents:

[Abstract: 3](#_Toc531522340)

[Introduction: 3](#_Toc531522341)

[What is big data? 3](#_Toc531522342)

[Why Twitter data? 3](#_Toc531522343)

[Related Work 4](#_Toc531522344)

[Methodology 5](#_Toc531522345)

[1. Discovery phase 5](#_Toc531522346)

[2. Data preparation 8](#_Toc531522347)

[3. Model planning phase 9](#_Toc531522348)

[4. Model building phase 10](#_Toc531522349)

[5. Communication results 11](#_Toc531522350)

[Conclusion 12](#_Toc531522351)

[References: 13](#_Toc531522352)

**Table of Figures:**

[***Figure 1:Frequency distribution of gender 7***](#_Toc531522369)

[***Figure 2: Time zone frequency 7***](#_Toc531522370)

[***Figure 3: Null values bar chart 8***](#_Toc531522371)

[***Figure 4: Accuracy boxplot 10***](#_Toc531522372)

# Abstract:

The rapid growth of social networks has produced an unprecedented amount of user-generated data, which provides an excellent opportunity for text mining. Authorship analysis, an important part of text mining, attempts to learn about the author of the text through subtle variations in the writing styles that occur between gender.

Such information has a variety of applications including advertising and law enforcement. One of the most accessible sources of user-generated data is Twitter, which makes the majority of its user data freely available through its data access API. In this research we will train an algorithm to determine if a Twitter account belonged to a man or a woman (male/female) based on many criteria that we will demonstrate and detail. This prediction will be conducted to apply to the data for classifying by using RandomForestClassifier

and Linear Support Vector Classification models (SVC). [3]

# Introduction:

## What is big data?

Big data is exactly what it sounds like — a lot of data. Alone, a single point of data can’t give you much insight. But terabytes of data, combined together with complex mathematical models and boisterous computing power, can create insights human beings aren’t capable of producing. The value that big data Analytics provides to a business is intangible and surpassing human capabilities each and every day. [1]

## Why Twitter data?

Twitter is a gold mine of data. Unlike other social platforms, almost every user’s tweets are completely public and pullable. This is a huge plus if you’re trying to get a large amount of data to run analytics on. Twitter data is also pretty specific. Twitter’s API allows you to do complex queries like pulling every tweet about a certain topic within the last twenty minutes or pull a certain user’s non-retweeted tweets. [1]

# Related Work

Some​ ​related​ ​work​ ​that​ ​used​ ​machine​ ​learning​ ​to​ ​classify​ ​a​ ​Twitter​ ​user’s​ ​gender includes​ ​“Gender​ ​Recognition​ ​Algorithm​ ​for​ ​Social​ ​Media:​ ​Twitter​ ​case”​ ​by​ ​Fi-Ware Consoft​ ​and​ ​“Detecting​ ​the​ ​Gender​ ​of​ ​a​ ​Tweet​ ​Sender”​ ​by​ ​the​ ​University​ ​of​ ​Regina’s Department​ ​of​ ​Computer​ ​Science.​ ​The​ ​former​ ​article​ ​used​ ​decision​ ​trees​ ​and​ ​what​ ​they called​ ​a​ ​“waterfall​ ​evaluation”​ ​to​ ​classify​ ​such​ ​attributes​ ​as​ ​screen​ ​name,​ ​description, and​ ​profile​ ​color.​ ​The​ ​latter​ ​used​ ​a​ ​support​ ​vector​ ​machine​ ​along​ ​with​ ​ranker-filter algorithms​ ​they​ ​obtained​ ​from​ ​the​ ​WEKA​ ​Toolkit.​ ​The​ ​latter​ ​also​ ​used​ ​feature​ ​selection algorithms,​ ​including​ ​chi​ ​square,​ ​information​ ​gain​ ​ratio,​ ​relief,​ ​and​ ​symmetrical uncertainty.​ ​They​ ​were​ ​both​ ​able​ ​to​ ​achieve​ ​a​ ​gender-classifying​ ​accuracy​ ​of​ ​about 87%.​ ​Our​ ​project​ ​is​ ​different​ ​in​ ​that​ ​we​ ​parse​ ​an​ ​individual’s​ ​metadata tweet, ​in​ ​order​ ​to​ ​classify​ ​the author’s​ ​gender.

This work also done byNishant Salvi I found on GitHub.com**.** He combined the features Text and Profile Description. Using the NLTK library he was able to separate the stop words from them. Then removed words that were not part of the English language. Also, he dropped the rows, whose tweets were not in English.

The visualization of the most frequently used relevant English words (barring stop words) can be done using word clouds: [4]

Male word cloud:

Female word cloud:

We think this attempt of extracting the most frequent words of male and female with this result will not be very insightful due to their similarity as we also get the same result, we will demonstrate that later and show how we changed our features.

# Methodology

## Discovery phase

We installed our dataset from kaggle.com website.

The dataset contains 20,000 rows, each with a user name, a random tweet, account profile and image, location, and even link and sidebar colour. With shape function we can observe this output (20050, 26) representing the dimensionality of the data frame. The type of the problem is classification, Our​ ​project​ ​aims​ ​to​ ​use​ ​Twitter​ ​metadata​ ​to​ ​accurately​ ​determine​ ​and​ ​classify​ ​a user’s​ ​gender.

We will need to have some understanding of programming languages such as R or Python which we choose it to code with, as we are more familiar with.

Python is an increasingly popular tool for data analysis and we find it to be more flexible to work with. [2]

This prediction is valuable for marketing, personalization, and legal investigation most important is that it can help companies when targeting the gender with products that suits them specifically and gain better profits in return.

**Hypothesis**

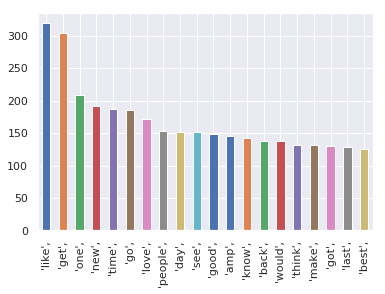
H1: Words in tweets and profiles don’t predict user gender well.

H2: Tweets’ meta data can contribute well to have good prediction of users’ gender.

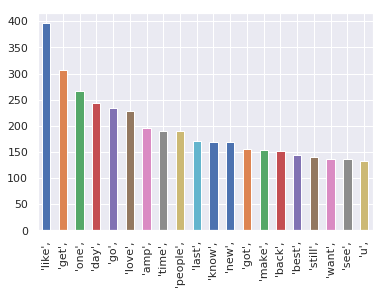
We took a small portion of data, almost 30 rows. From these small amounts, we discovered that male and female could be our label to predict for this problem.

We also provide some interesting visualization such as the most predictive words for both male and female. We can observe that both of them has similar frequency of words.

Male words:



Female words:



This is the target that we will train our model with, and it’s sort of balanced.

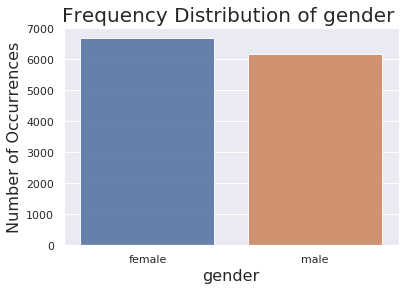


Figure 1:Frequency distribution of gender

This shows the time zone frequency among tweeters

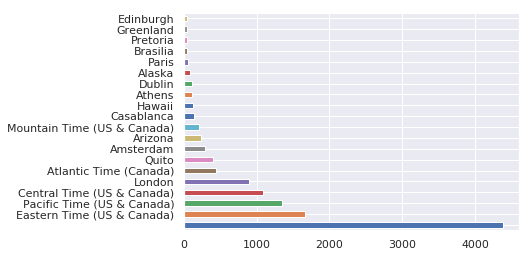


Figure 2: Time zone frequency

## Data preparation

Twitter user gender classification dataset:

**Data conditioning:**

Our target is the gender. The dataset has to be cleaned in a way that we will mention in the following steps:

* Filtered gender column from (brand) values because it is out of our project target which concentrate on male\female only.
* Filtered gender column from (unknown) because these are values that have been handled some missing values but either ways they won’t be providing much of an insight.
* Handling null values in six columns (visualization is provided next section) by filling

them with values.

* Balance the male and female amount via gender.value\_counts() function.

Male was (6194), Female was (6700).

* Cleaning text and description from an ambiguous letters and symbols.
* Filter text with Stop words to remove the most common words in a English.

**Survey & visualize**

* Create a bar chart showing how many missing values are in each column
* This shows which column has the most NaN values and how many cells in that column are empty.

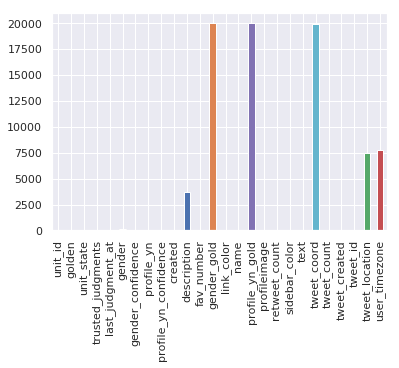
****

Figure 3: Null values bar chart

**Preprocessing**

* **Convert categorical variables before modeling**

Most of the algorithms (or ML libraries) produce better result with numerical variable. In python, library “sklearn” requires features in numerical arrays.

So because some of them do not take categorical variables as input. Thus, we convert them into numerical variables.

* **Using LabelEncoder() for target**

**Label Encoder:**It is used to transform non-numerical labels to numerical labels (or nominal categorical variables). Numerical labels are always between 0 and n\_classes-1.

* **Using get\_dummies for features**

Dummy’, as the name suggests is a duplicate variable which represents one level of a categorical variable. Presence of a level is represented by 1 and absence is represented

by 0.

## Model planning phase

From among the dataset we select some of them to be the features that we will build

our mode based on.

***golden, unit\_state, trusted\_judgments, last\_judgment\_at, profile\_yn,***

***profile\_yn\_confidence, fav\_number, link\_color, name, profile\_yn\_gold, retweet\_count, sidebar\_color', tweet\_coord, 'tweet\_count', tweet\_created, tweet\_id, tweet\_location,***

***user\_timezone***

All of the variables we decided to select except of these ('text','unit\_id','description',

'profileimage','gender\_confidence','created',’Tweets, 'Description

We exclude text due to their similarity for both male and female and thought it won’t be

a good predictor for our model as we established before. Description has emojis that framed a

challenge for us to trace them and convert them because they weren’t appearing as emojis!

As our goal is to classify the user’s gender, we will use RandomForestClassifier and Linear Support Vector Classification.

1. Model building phase

Training / Testing

We spilt the dataset into two datasets (training and testing) we toke 80% from original dataset for training and 20% for testing.

Because the more percent of train data the more accurate the model get.

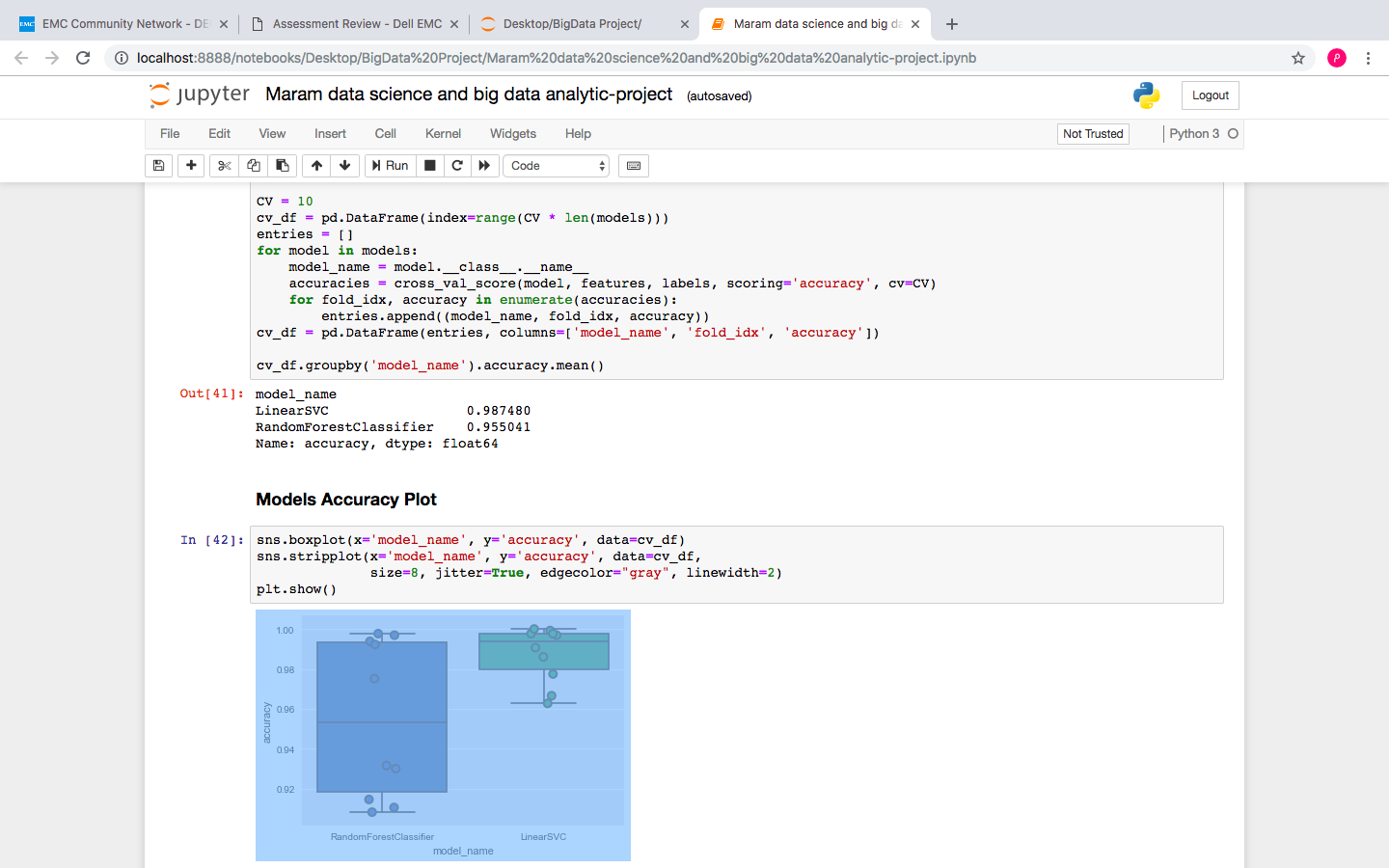
RandomForestClassifier and Linear Support Vector Classification are the models we run our prediction with.

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. [5]

Linear Support Vector Classification.

Similar to SVC with parameter kernel=’linear’, but implemented in terms of liblinear rather than libsvm, so it has more flexibility in the choice of penalties and loss functions and should scale better to large numbers of samples. [6]

**Accuracy:**



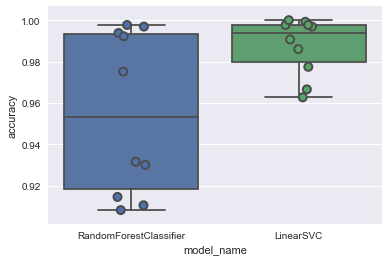


Figure 4: Accuracy boxplot

Linear SVC predict the gender higher than the random forest classification thus we will choose Linear SVC as our final model for the gender classification.

1. Communication results

Our goal is to predict Twitter users’ gender, we can see that the outcome of this model has achieved the goal we were looking for in a successful way.

**H1: Words in tweets and profiles don’t predict user gender well.**

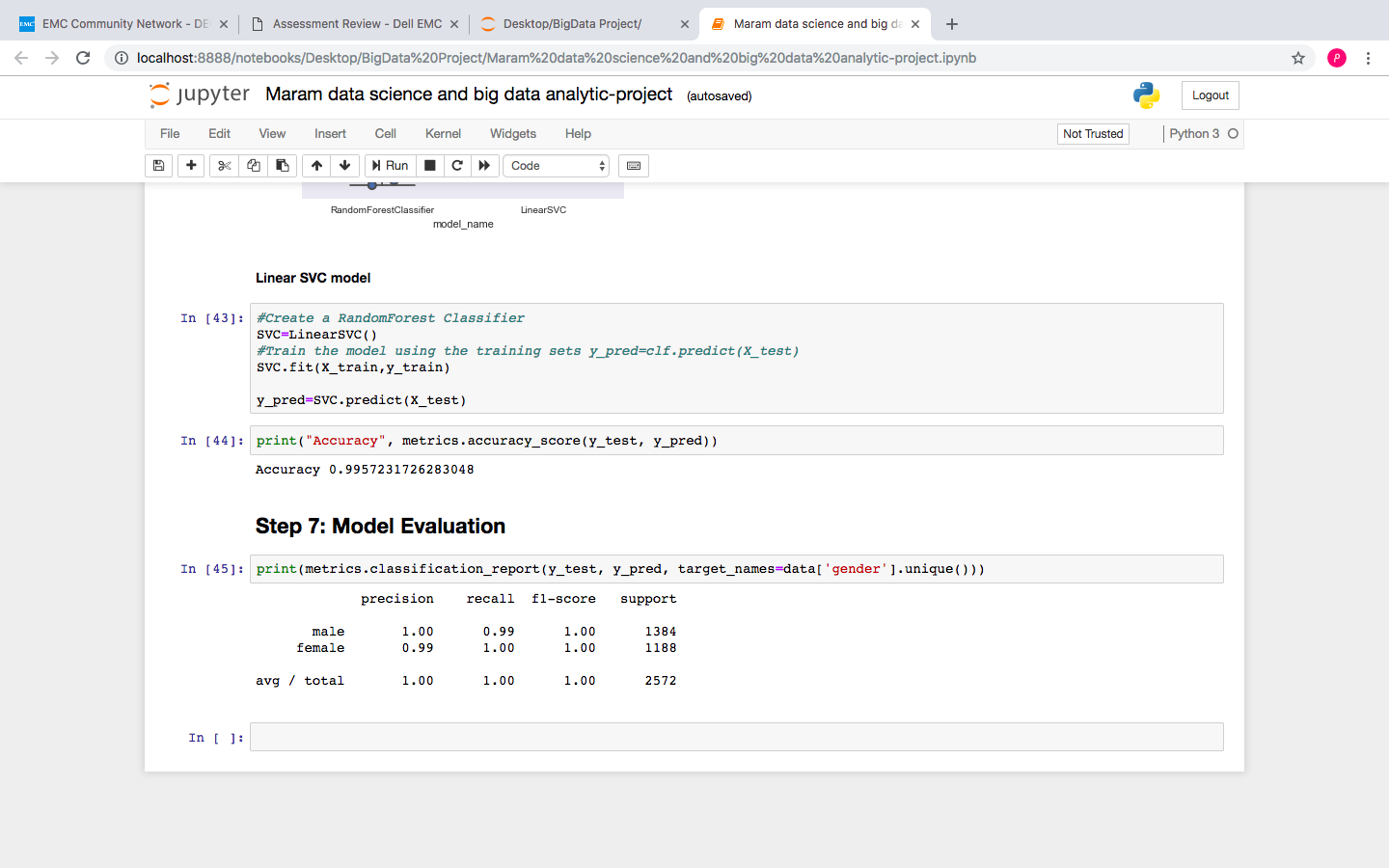
We tried to use text and description as feature and that attempt didn’t go well as we established before.

**H2: Tweets’ meta data can contribute well to have good prediction of users’ gender.**

The model’s accuracy we conducted on some meta data of the tweets are proven that this hypothesis is valid.

**Model Evaluation**

* Precision: what percent of things we mark positive really are positive.
* Recall: what percent of positive instances did we correctly identify.
* The f1-score gives the harmonic mean of precision and recall.
* Support: is the number of samples of the true response that lie in that class.



**Software**​ ​**and**​ ​**Hardware**​ ​**Used**

We​ ​used​ ​various​ ​software​ ​packages​ ​in​ ​this​ ​study, ​ ​including​ ​Anaconda Notebook’s​ ​Jupyter​ ​platform,​ ​the​ ​pandas​ ​data​ ​analysis​ ​library, scikit-learn’s​ ​tools​ ​for​ ​machine​ ​learning,​ ​and​ ​matplotlib​ ​data​ ​plotting​ ​library​ ​for​ ​data visualization.​ ​We​ ​each​ ​used​ ​our​ ​own​ ​personal​ ​laptop​ ​computers​ ​to​ ​carry​ ​out​ ​the computation.

# Conclusion

In this paper, we have presented an implementation steps of developing a prediction model for Twitter users gender classification we conducted it with two kinds of model and selected Linear SVC with 98% accuracy to be our final model

​ ​It​ ​would​ ​be​ ​wise​ ​to​ ​consider​ ​emojis when choosing text and description as features. Personal​ ​experience​ ​and​ ​anecdotal​ ​observations​ ​show​ ​that​ ​there​ ​are​ ​some​ ​emojis​ ​that are​ ​used​ ​by​ ​one​ ​gender​ ​much​ ​more​ ​than​ ​the​ ​other​ ​(for​ ​instance,​ ​women​ ​use​ ​the​ ​heart emoji​ ​frequently​ ​and​ ​men​ ​are​ ​less​ ​inclined​ ​to​ ​use​ ​emojis​ ​overall).​ Emojis was a little bit of a challenge we may consider using it in the future with a better resources.

# References:

[1] "Twitter Data Mining: A Guide to Big Data Analytics Using Python", *Chatbots Life*, 2017. [Online]. Available: https://chatbotslife.com/twitter-data-mining-a-guide-to-big-data-analytics-using-python-4efc8ccfa219. [Accessed: 27- Nov- 2018].

[2] *Media.readthedocs.org*, 2017. [Online]. Available: https://media.readthedocs.org/pdf/dataanalysispython/latest/dataanalysispython.pdf. [Accessed: 27- Nov- 2018].

[3] 2012. [Online]. Available: https://www.researchgate.net/publication/276037940\_Gender\_Prediction\_on\_Twitter\_Using\_Stream\_Algorithms\_with\_N-Gram\_Character\_Features. [Accessed: 27- Nov- 2018].

[4]"Twitter User Gender Classification", *Kaggle.com*, 2016. [Online]. Available: https://www.kaggle.com/crowdflower/twitter-user-gender-classification/home. [Accessed: 27- Nov- 2018].

[5]"3.2.4.3.1. sklearn.ensemble.RandomForestClassifier — scikit-learn 0.20.1 documentation", *Scikit-learn.org*, 2018. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. [Accessed: 27- Nov- 2018].

[6]"sklearn.svm.LinearSVC — scikit-learn 0.20.1 documentation", *Scikit-learn.org*, 2018. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html. [Accessed: 27- Nov- 2018].