**Gender classification of twitter users with NLP**

# Muhammad Shahrooz Khan

## Master in Computer Science Superior University Lahore

# Malik Muhammad Adnan Sami

## Master in Computer Science Superior University Lahore

# Ammaz Baig

## Master in Computer Science Superior University Lahore

# Laiba Imran

## Master in Computer Science Superior University Lahore

# Wasif Gill

## Master in Computer Science Superior University Lahore

# **1.Introduction:**

With the rise of social media in this ear, there has been a surge in interest in automatically identifying users based on their informal content. In this context, the research of other aspects intrinsic to users, such as political inclinations, personality, and gender, as well as the categorization of users in categories such as age, ethnicity, origin, and race has gained a lot of interest notably based on Twitter data. The current work focuses on the job of gender categorization in tweets written in Portuguese by extracting gender expression linguistic cues utilizing 25 attributes, which are often employed on text attribution tasks.

# **2.Objective:**

The major objective is to predict user gender based on Twitter profile information.

# **3.Data Source:**

This data is extracted from an online source (Kaggle). This dataset contains 20050 tuples and 26 attributes. In 26 there are 25 predictors and 1 is target variable that is gender in this case.

# **4.Data Description:**

The data contains the following fields:

* **unitid**: a unique id for the user
* **\_golden**: whether the user was included in the gold standard for the model; TRUE or FALSE
* **unitstate:** state of the observation; one of finalized (for contributor-judged) or golden (for gold standard observations)
* **trustedjudgments**: number of trusted judgments (int); always 3 for non-golden, and what may be a unique id for gold standard observations
* **lastjudgment\_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 dataset but was not available when contributors went to judge it
* **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
* **profileyngold:** 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 time zone of the user

# **5.Methdology:**

The methodology has been explaining step by step as follows

## **5.1 Step 1- Install and import all important libraries**

In this step we install and import all libraries that are being used in this case. Numpy for linear algebra, pandas for data processing input output in csv files and OS for basic input and output operations. As shows in Figure 1.1

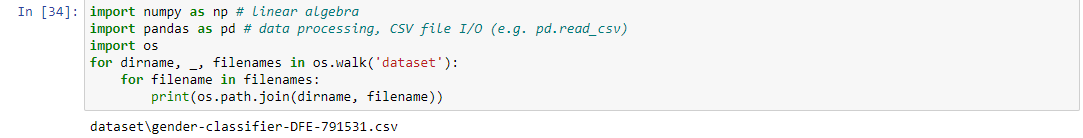


Figure 5.1Importing Libraries that are used in project

## **Step 2- Reading CSV File**

In this section we read CSV file of our data set that is mention above in data source section. As shows in Figure 5.2

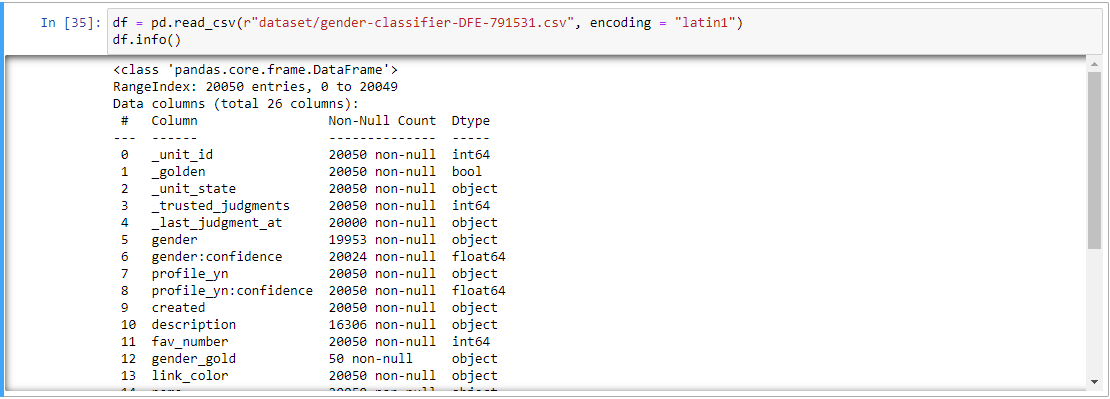
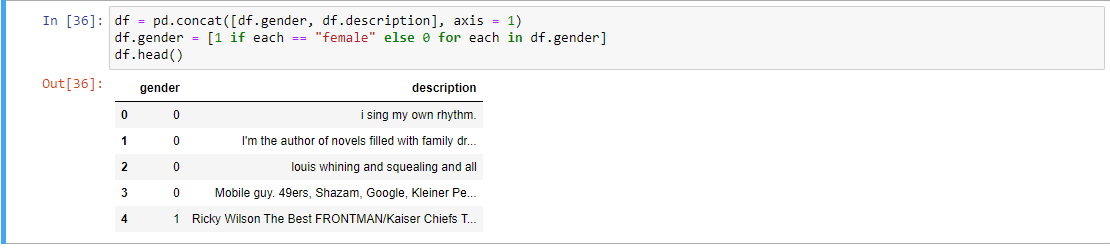


Figure 5.2 Reading CSV file

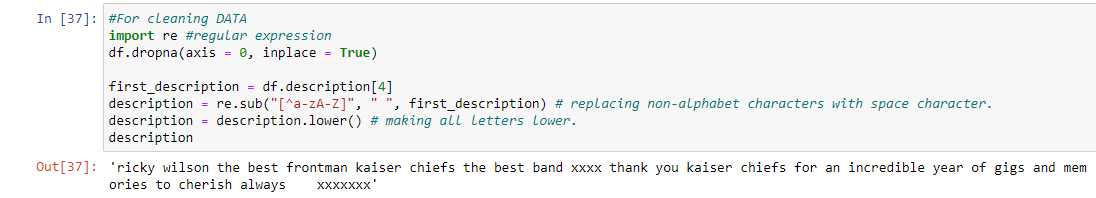
## **Step 3- Encoding of categorical variables ‘male’ and ‘female’.**

Encode male and female in 1 and 0 respectively as shows in Figure 5.3 and concatenate gender and description.



## **Step 4- Cleaning of the description column.**

keep only the words containing alphanumeric characters and remove punctuations. And replace every non alphabetic character with space, and convert all Upercase into lowercase. As shows in Figure 5.4.



## **Step 5-Tokenization and remove stop words from description column.**

Tokenization is way to break text into smaller pieces, which is known as tokens. In this case we use word\_tokenization() function, which splits into words. we could have use description.split() but it doesn't split words like this: "shouldn't = should not". These function are belongs to NLTK , and we download stopwords and punkt. Stopwords is used for removing stop words and Punkt is used for word\_tokenization(). As shows in Figure 5.5.

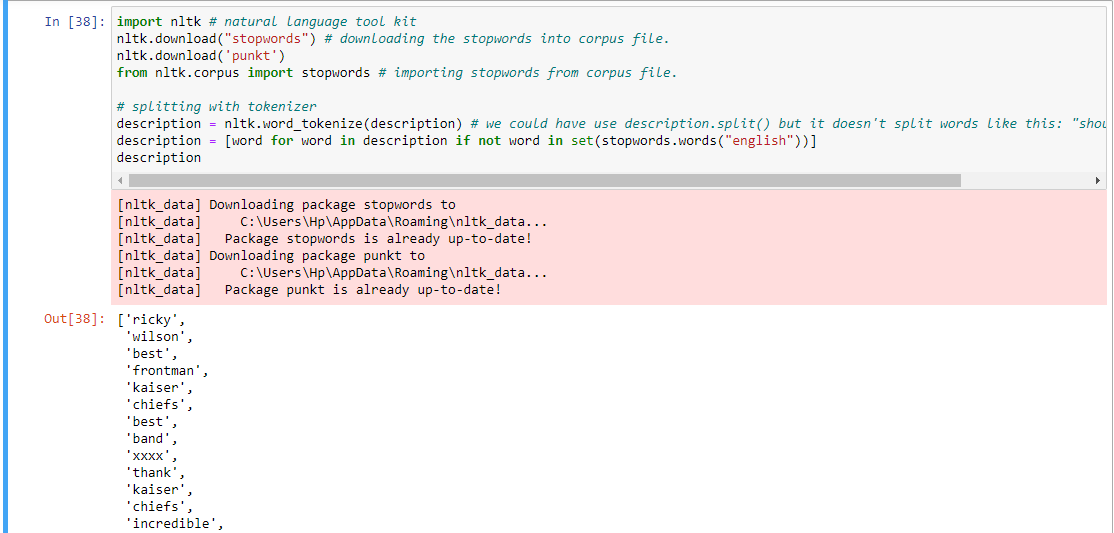


Figure 5.5 Tokenization and removing stop words

## **Step 6-** **Lemmatization of the text of the description.**

Lemmatization is the process of grouping together the different inflected forms of a word so they can be analyzed as a single item. Lemmatization is similar to stemming but it brings context to the words. So it links words with similar meanings to one word. (exp: Loved => love, memories => memory) As shows in Figure 5.6.

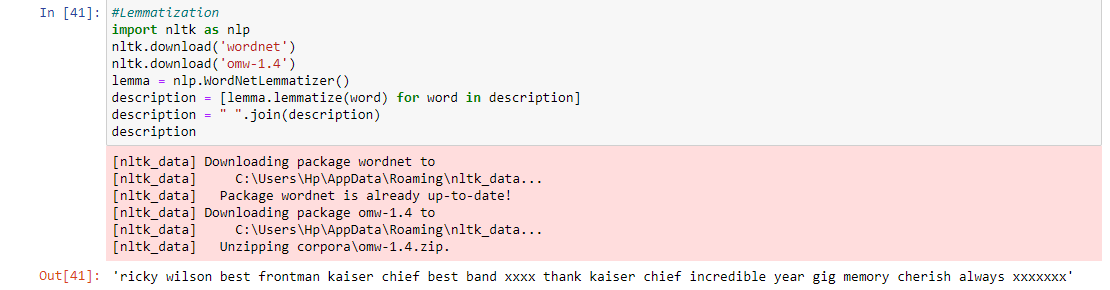


Figure 5.6 Lemmatization on description column

## **Step 7- Vectorization and Classification.**

Vectorization is a methodology in NLP to map words and phrases from vocabulary to a corresponding vector of real numbers which is used to find word predictions, word similarities/semantics. To make documents corpora more relatable for computers they must first be converted into some numerical structure. Few techniques are used to achieve this, is called ‘Bag of Words’. As shows in Figure 5.7 and there are many classifiers such as Lightgbm, and XGBoost but is this case we use Naive Bayes Classifier with the accuracy 0.5419987737584304. as shows in figure 5.8

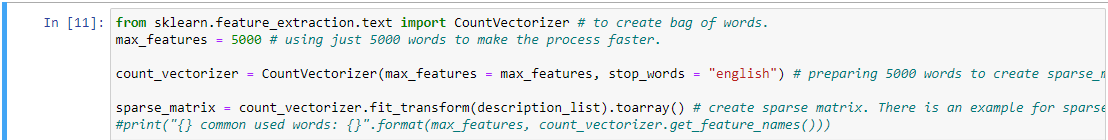


Figure 5.6 Vectorization process

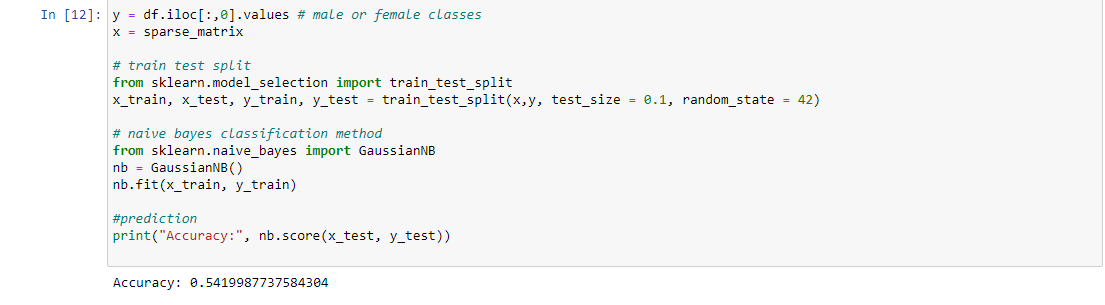


Figure 5.7 Classification Naive Bayes

# **6.Conculion:**

In this case we use naïve Bayes classification to predicate gender of twitter user, with the accuracy of 0.541998. as mention in Figure 5.7 in this case we perform read CSV file, Tokenization and removing stop words, and perform Lemmatization.

# **7.Referance:**

|  |  |
| --- | --- |
| [1] | K. CrowdFlower, "Twitter User Gender Classification," [Online]. Available: https://www.kaggle.com/datasets/crowdflower/twitter-user-gender-classification. |