Problem Statement

Identification of the gender of a Twitter user using the user's profile information

Problem Statement: With the growth of social media in recent years, there has been an increasing interest in the automatic characterisation of users based on the informal content they generate. Gender recognition is essential and critical for many applications in the commercial domains. Imagine that Twitter needs to push advertisements based on gender. As there are many fake accounts or accounts belonging to organisations, Twitter cannot rely on what the users themselves mention in their respective profile descriptions. Hence, Twitter would need to determine the gender of the profile based on user behaviour on the platform.

To enable this, you would need to train an algorithm to determine if a Twitter account belongs to a man or a woman or an organization. You need to build two models based on two different classification algorithms and compare the results. You may choose any of the algorithms that you have been introduced to, throughout the course. Moreover, feel free to proceed with any data preparation technique which suits the given dataset (whether it is covered in the course or not).

Data Description:
The dataset contains 20,000 rows, each with a particular username, a random tweet, account profile and image, location, link, sidebar colour and other miscellaneous data.

The dataset contains the following fields:

Note: You can think of a golden account as a verified account and relevant columns have "gold" either as a prefix or suffix. Twitter verifies the accounts of famous organisations and people so that Twitter users can be sure if the account is fake or not.

Variable Name	**Description**	**MissingValues**
_unit_id	unique id for a user	no
_golden	whether the user was included in the gold standard for the model; TRUE or FALSE	no
_unit_state	state of the observation; one of finalised (for contributor-judged) or golden (for gold standard observations)	no
_trusted_judgments	number of trusted judgments (int); always 3 for non-golden, and what may be a unique id for gold standard observations	no
_last_judgment_at	date and time of last contributor judgment; blank for gold standard observations	yes
gender	one of male, female, or brand (for non-human profiles)	yes
gender:confidence	a float representing confidence in the provided gender	yes
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	no
profile_yn:confidence	confidence in the existence/non-existence of the profile	no
created	date and time when the profile was created	no
description	the user's profile description	yes
fav_number	number of tweets the user has favourited	no
gender_gold	if the profile is golden, what is the gender?	yes
link_color	the link colour on the profile, as a hex value	no
name	the user's name	no
profile_yn_gold	whether the profile y/n value is golden	yes
profileimage	a link to the profile image	no
retweet_count	number of times the user has retweeted (or possibly, been retweeted)	no
sidebar_color	colour of the profile sidebar, as a hex value	no
text	text of a random one of the user's tweets	no
tweet_coord	if the user has location turned on, the coordinates as a string with the format "[latitude, longitude]"	yes
tweet_count	number of tweets that the user has posted	no
tweet_created	when the random tweet (in the text column) was created	no
tweet_id	the tweet id of the random tweet	no
tweet_location	location of the tweet; does not seem to be particularly normalized	yes
user_timezone	the timezone of the user	yes