Gender Recognition Algorithm for Social Media: Twitter case

1. INTRODUCTION

Social Media platforms handle users' profile at different levels of details, collecting information (e.g. date of birth, gender) which users explicitly provide.

Most of these profile are not complete or can even be misleading. They also don't provide information related to people sentiment in relation to topic discussions. We'll aim at determining more fine grained information about users behavior and reaction to discussion topics, for the time being we concentrate on profiles and in this document we describe the gender recognition algorithm.

On Twitter the information about user gender is not specified. Nonetheless, it is interesting having such an information for analytics purposes (e.g. for marketing research or having a clue if a target event was more interesting for male or for female users could be very useful). Providing support to business analytics is the reason why of our work: the development of a gender recognition algorithm (GRA) whose purpose is to classify the gender of twitter users.

2. USER PROFILE ANALYSIS

Twitter profiles have been chosen as the baseline of our data collection. When a user registers himself/herself to the Twitter platform, has to fill a profile form, consisting of about 30 fields containing biographical and other information, such personal interests and hobbies.

However, many of those fields are optional, and therefore a substantial set of Twitter users leave blank many (or all) of those optional fields. Moreover, Twitter's profile form does not include a specific "gender" field, which complicates gender identification for Twitter users.

Additionally, in Twitter there are profiles not related to a single user: for example companies profiles, fan pages, press profiles and so on. In these case is useless to assign a specific "gender" to these profiles, since people that write on behalf of such profiles could be of both sex (there isn't just one person writing).

In many cases, identifying specific features to be collected is the most challenging part of using machine learning.

For our analysis we identified in the Twitter user profile 3 groups of fields that can help in devising an algorithm to classify a generic profile into one of the following categories:

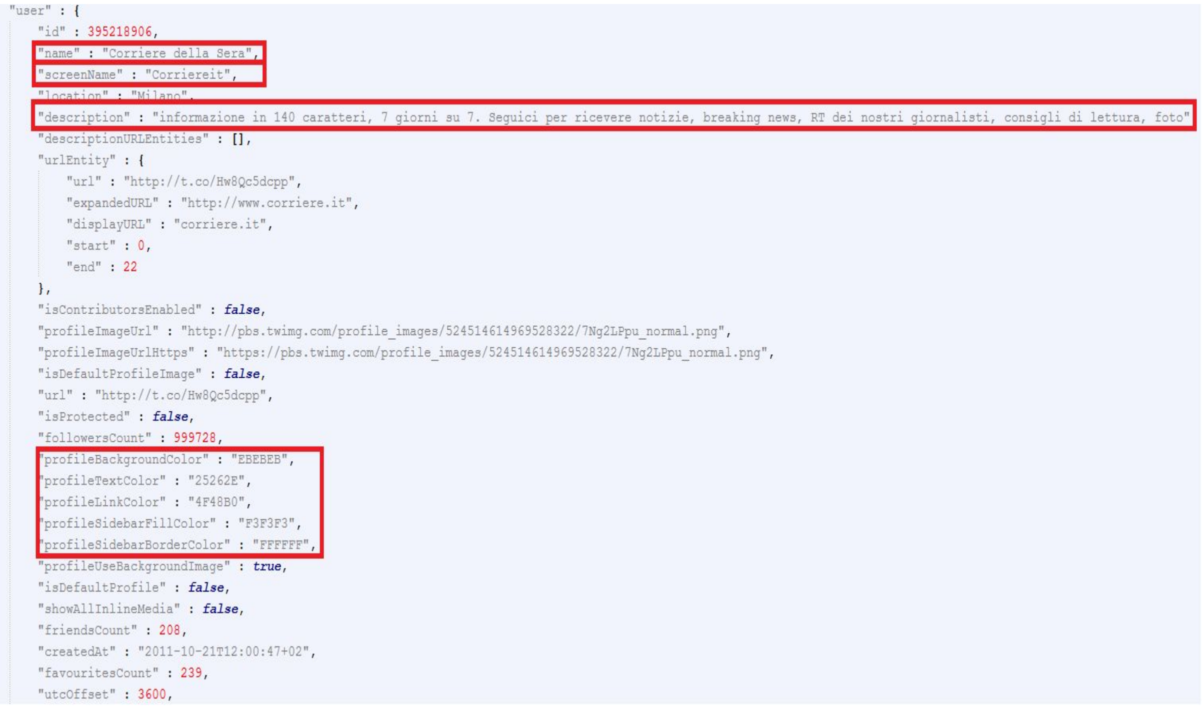
- Male
- Female
- page
- unknown

The generic term page group all the profiles which are not related to a single user (fanpages,official pages, advertising pages and so on).

The unknown category is related to the profiles that cannot be recognized from the algorithm because of missing parameters (missing description and name not recognizable) or not known from the algorithm.

The three group of fields we choose are the following:

1. user name and user screen name
2. user profile description
3. user profile colors[1]



```
"user" : {
    "id" : 395218906,
    "name" : "Corriere della Sera",
    "screenName" : "Corriereit",
    "location" : "Milano",
    "description" : "informazione in 140 caratteri, 7 giorni su 7. Seguici per ricevere notizie, breaking news, RT dei nostri giornalisti, consigli di lettura, foto"
    "descriptionURLEntities" : [],
    "urlEntity" : {
        "url" : "http://t.co/Hw8Qc5dcpp",
        "expandedURL" : "http://www.corriere.it",
        "displayURL" : "corriere.it",
        "start" : 0,
        "end" : 22
    },
    "isContributorsEnabled" : false,
    "profileImageUrl" : "http://pbs.twimg.com/profile_images/524514614969528322/7Ng2LPpu_normal.png",
    "profileImageUrlHttps" : "https://pbs.twimg.com/profile_images/524514614969528322/7Ng2LPpu_normal.png",
    "isDefaultProfileImage" : false,
    "url" : "http://t.co/Hw8Qc5dcpp",
    "isProtected" : false,
    "followersCount" : 999728,
    "profileBackgroundColor" : "EBEBEB",
    "profileTextColor" : "25262E",
    "profileLinkColor" : "4F48B0",
    "profileSidebarFillColor" : "F3F3F3",
    "profileSidebarBorderColor" : "FFFFFF",
    "profileUseBackgroundImage" : true,
    "isDefaultProfile" : false,
    "showAllInlineMedia" : false,
    "friendsCount" : 208,
    "createdAt" : "2011-10-21T12:00:47+02",
    "favouritesCount" : 239,
    "utcOffset" : 3600,
```

Fig. 1: main fields of a Twitter user profile

In Figure 1 the structure of a Twitter user profile is reported, with the the main field groups highlighted.

[1] A user profile is made of multiple “colours” attributes: backgroundColor, textColor, linkColor, profileSidebarFillColor and profileSidebarBorderColor.

3. ALGORITHM DEFINITION

The algorithm consists of four sub-algorithms that evaluate the following fields of a user profile:

1. name
2. screen name
3. user profile description
4. user profile colours

The GRA algorithm can proceed then in 3 different ways to interpret the results of the sub-algorithms:

- Waterfall: if the result of the first evaluation is no matching or a low score proceed to the next evaluation until one of the sub algorithms has a match with a satisfying score, otherwise return undefined.
- Combining results with a weighted matrix: the scores of the different algorithm are combined taking in account the weight of each feature (i.e if the name feature has a perfect matching has more weight than the feature based on profile color or description)
- The results can be used to feed another classifier (for example SVM, decision tree or a simple ANN)

Optionally can be also defined a score that describes the accuracy of the classification.

3.1 DETECTING USER GENDER FROM NAME AND SCREEN_NAME

In order to get the gender from the profile name, we split the string and check if there is a matching with a dictionary containing the relation name/gender.

Obviously the name is cleaned from useless characters. If a multiple matching is found, the algorithm tries to detect if there is a second name or the surname itself can be used also as a name. The same procedure will be done for the screen name field.

From the collected profiles, we found that roughly 66% of users maintain their name in a way that is recognizable for a machine.

Furthermore the recognition is based on a dictionary of names so even if the name of the person is written in a clean manner with no noise (e.h. addition of special characters in the screen name, or name misspellings..) if it is not listed into the reference dictionary it won't be recognized.

Therefore, information in name and screen name fields is not sufficient alone to achieve a good accuracy.

3.2 DETECTING USER GENDER FROM PROFILE DESCRIPTION

Unlike traditional detection mechanisms which are based on samples of hundreds of words in length, the analysis of Twitter profile description (bio) is hindered by a 160 character limit.

Other difficulties include both accidental and purposeful misspellings, use of internet slang for profile description, the fact that this field is optional (so just some of the users will have it filled) and can be partially filled (i.e. descriptions too short, e.g. consisting of one or two words) - in addition to the fact that sometimes users write in their profile “useless” information (e.g. urls of favorite websites, mentions to other users, quotes to famous authors and so on).

Nonetheless, certain distinctive traits provide the possibility for accurate analysis: of particular utility to classify the gender are the emoticons, on purpose misspellings and slang (e.g. emoticons like “<3” are more likely to be used from females users while use of bad words are more likely to be used from males[1]).

It has been decided to filter the words contained into the user profile description removing all the mentions, all numbers and non-alphabetical letters (allowing just the character useful to represent emoticons) and stop words (either in Italian and English).

It has been decided to use each word contained in the cleaned description as a feature (even emoticons and misspelled words) computing for them the Term Frequency–Inverse Document Frequency (TF-IDF) (it is usually applied on documents but we believe it will have good performance also in our case, giving more weight to specific gender distinctive terms).

TF-IDF, is a simple way to generate feature vectors from text documents. It computes two statistics for each term in each document:

1. the term frequency (TF), which is the number of times the term occurs in that document
2. the inverse document frequency (IDF), which measures how (in)frequently a term occurs across the whole document corpus.

The product of these values, TF × IDF, shows how relevant a term is to a specific document.

In this environment, consisting of dictionary words (roughly 270.000 words of the Italian dictionary, a quarter of a million distinct English words, excluding inflections, and words from technical and regional vocabulary [3][4]), acronyms, emoticons and misspellings there are hundreds of thousand eligible words and tracking a distinct mapping from each word to an index in the vector would be expensive.

In order to map terms to vector indices, a technique known as the hashing trick can be used. It takes the hash code of each word modulo a desired vector size, S, and thus maps each word to a number between 0 and S–1.

This always yields an S-dimensional vector, and in practice is quite robust even if multiple words map to the same hash code. To make this technique robust enough will be tried to dimension S as 2^{18}, 2^{19} and 2^{20} switching between these three values in different running with which size the classification algorithm that will be used will be reliable enough for our purpose.

The words of each description will be mapped into LabeledPoints [5] that will be used to feed different classification algorithms. To start Naive Bayes, SVM and Decision Tree classification algorithms have been chosen. The accuracy of each of these algorithms will be tested applied to our classification problem and will be chosen the one that perform the best.

3.3 IDENTIFY USER GENDER FROM PROFILE COLORS

The approach in this case is to predicts gender using five color-based features extracted from Twitter user profile. We decided to add this sub-classifier for two main reasons:

1. Profile colors are a customization of the twitter account that can be highly related to the gender
2. This approach is language independent so it can work no matter in which language is the profile description and which username the user has chosen.
3. colors alone can provide reasonably accurate gender predictions as shown in [2].

The drawback to this approach however is that it is hard to identify pages (most of the time they use the default colors provided by twitter) The colors in the raw profile are in their hex triplet form (ie. C0DEED). A hex triplet is a six-digit, three-byte hexadecimal number used to represent colors.

The bytes represent the red, green and blue components of the color. One byte represents a number in the range 00 to FF (in hexadecimal notation), or 0 to 255 in decimal notation. The hex triplet is formed by concatenating three bytes in hexadecimal notation, in the following order:

1. Byte 1: red value (color type red)
2. Byte 2: green value (color type green)

3. Byte 3: blue value (color type blue)

This gives a total of 256^3 colors combinations. As pointed from [2] this is a large number of combinations ($16*10^6$) that has to be reduced since there is no point to have all the possible shades of the colors as features so we can use quantization to reduce the number of color combinations. It has been decided to try 2 different features set: the first one with all the five profile colors and the second with three over five (background,text,link) as may be the other 2 could be less meaningful (users won't modify them directly but with the choice of a theme). Also these features will be transformed into LabeledPoints [5] and then used to train different classification algorithms to evaluate which one perform the best.

3.4 POSSIBLE IMPROVEMENTS

The algorithm can be improved in a number of different ways, but two fields in particular, that have not been included in the current algorithm but which analysis can lead to interesting results are:

- **url field:** this is an optional field that can contain a link to another social network or about page which can have explicit gender field. A possible improvement could be make the algorithm open these links to obtain useful information about user gender.
- **profile image field:** this field contains the url to the user avatar image. Some of twitter users use their pictures as avatar image. The gender of the user can be guessed by using an ANN (Artificial Neural Network) that is trained to recognize user gender from picture.

4. RESULTS

4. 1 DATA COLLECTION AND FEATURE EXTRACTION

By means of the Social Data Aggregator package twitter connector around 10 thousand distinct profiles have been collected.
We started to classify some of them to feed the algorithm (the training set). These profiles were manually labeled as male (m), female (f), or page (x). While labeling any profile where the gender was unclear or the description was missing or in a language other from English or Italian has been discarded (however the algorithm can work with every language by replacing the training set and the names storage with profiles containing the chosen language).

Until now around 3000 distinct users profiles have been classified.
Within the classified data there is a gender distribution slightly pending on male profiles (68% excluding page profiles). The reason of this can be related to the topics that were under monitoring during the data collection. May be some of them were more attractive to male users than females. To avoid data overfitting (the algorithm start to recognize mostly males than others) we will keep around the same number of profiles from the different types.

The name storage mostly keeps Italian names since our first implementation will focus on Italian profiles.

4.2 NAME/SCREEN NAME APPROACH

For the purpose of this sub algorithm we created a set of base file mapping keywords that can be used in name and screen name fields with the related gender.

Keywords can be names (like male name Francesco), personal titles (like mr,mrs,miss..) or words used to identify a page (like news, official,magazine..).

The algorithm tries to split the words contained in the name field and to find a match in the keywords dictionary. If there is at least one match the algorithms tries to figure out if the recognized names belong to the same gender otherwise guesses the gender following different heuristics (surnames that can be also names, female names used as second names for boys, the probability of put the first name before the second name and so on).

The same has be done with the screen name adding a further process to help for classification. During the collection of the training and test set we noticed that sometimes on screen name people don't use any character to distinct different words (for example screen names as phaigehope or katefinn). In this way is very difficult to extract the first name of the user so the algorithm tries to find a match with the names contained into the keywords with longest prefix match to avoid bad recognition with names contained one in another (daniel and daniela for example male and female respectively).

This sub algorithm has been ran on two samples:

1. the whole set of profiles collected (5000 profiles)
2. the whole set filtered only on user that had their name written explicitly (around 60% of collected data). In practice the test set 2 is a subset of the test set 1.

	total	male	female	pages
test set 1	56.78	67.30	70.13	6.26
test set 2	**86.35**	86.81	**88.99**	-

In the second row the percentage of page recognition is missing since test set 2 is related just to people that maintained their name not to pages. To justify this results three aspects has to be considered:

1) from our experience we have seen that roughly 60% of people write their name in the name or screen name section of their twitter profile
2) If we consider also pages the percentage of profiles recognizable from keywords decreases
3) Even if the name is clearly written the percentage of recognition is not 100% due to the fact that:
    a) some names could not be into the keywords file
    b) some names are unisex (can be used both for male and for female)

4.3 COLORS APPROACH

With respect to profile color approach we had some issues at beginning since 4k user profiles were not enough to find a clear pattern and even trying with different classifier and schema results were not so encouraging (around 48%). We needed an elevated number of profiles that could not be classified by hand, so the found solution was to take all the profiles we gathered from the collector (around 400k) and run on them the name/screenName approach that has a quite good accuracy when the name is written explicitly. In this way we obtained around 163k profiles.

Since the profiles set contained unbalanced number of male and females (the number of males overwhelmed the number of females by 40k profiles) we have decided to create a subset containing a balanced number of male/female profiles. Unfortunately a small percentage of page has been found that makes hard the classification for this type.

In total four sets of profiles has been created:

Set A	Contains a balanced number of male and female profiles
Set B	It is a subset of A without default colors (twitter provide a default set of colours if it is not modified by the user)
Set C	the whole unbalanced set
Set D	It is a subset of C without default colors

We divided each of the different sets in 70% of the data for the training and 30% for the tests, using the training data to feed four different classifiers:

- NB: Naive Bayes
- DT: Decision tree
- RF: Random forest
- LR: Logistic Regression

The tests on the accuracy of each classifier has been done by considering from 1 to 5 colors and mapping each color to 9 or 12 bits.

For almost all the sets the best performances have been obtained with the combination 4 colours 9 bits and 4 colours 12 bits.

RESULTS FOR SET A WITH 4 COLOURS:

Accuracy using 9 bits to represent colors:

	Total	**Males**	**Females**
NB	59,16	79,74	39
DT	57,79	76,24	39,79
RF	57,96	76,89	39,47
LR	59,73	80,35	39,54

Accuracy using 12 bits to represent colors:

	Total	**Males**	**Females**
NB	58,58	79,65	37,89
DT	56,72	63,84	50,37
RF	54,89	83,54	26,35

LR	59,38	80,39	38,77

RESULTS FOR SET B WITH 4 COLOURS:

Accuracy using 9 bits to represent colors:

	Total	**Males**	**Females**
NB	**64,85**	**70,32**	**60,36**
DT	61,65	55,57	68,64
RF	60,29	37,01	84,45
LR	**65,5**	**67,86**	**64,13**

Accuracy using 12 bits to represent colors:

	Total	**Males**	**Females**
NB	64,64	67,41	62,83
DT	59,76	33,93	86,45
RF	59,25	31,59	87,77
LR	64,86	67,04	63,65

RESULTS FOR SET C WITH 4 COLOURS:

Accuracy using 9 bits to represent colors:

	Total	**Males**	**Females**
NB	65,4	85,31	36,26
DT	63,87	82,06	37,31
RF	63,45	94,96	16,65
LR	65,94	91,15	28,75

Accuracy using 12 bits to represent colors:

	Total	**Males**	**Females**
NB	64,66	84,28	35,92
DT	62,24	98,75	7,81
RF	60,9	98,7	4,5
LR	65,45	88,03	32,25

RESULTS FOR SET D WITH 4 COLOURS:

Accuracy using 9 bits to represent colors:

	Total	**Males**	**Females**
NB	65,9	76,73	53,28
DT	62,52	64,86	60,56
RF	60,24	95,48	16,75
LR	66,36	82,62	46,9

Accuracy using 12 bits to represent colors:

	Total	**Males**	**Females**
NB	65,22	77,85	50,31
DT	59,56	97,62	12,51
RF	58,59	98,15	9,63
LR	65,54	76,55	52,7

Gathered data show that the best accuracy is reached with 9 bits to represents rgb colors,default colors filtered from the training/test set and using Naive Bayes or Logistic Regression for classification. It is also clear that the unbalanced set of profiles doesn't add a valuable increase of performances respect to the balanced set (set B against set D).

Looking at the results, the configuration chosen are:

- set B as training set
- 9 bits rgb colors scaling from 24 bits
- 4 colours over 5 (background,text,link,scrollbar)

4.4 DESCRIPTION APPROACH

In this case we classified by hand up to 5000 twitter user profiles discarding profiles that didn't have the description filled, with a too short description or with a description in a language that was neither English nor Italian.

Since the whole set contains a very high number of male profiles (2948) respect the number of females(1311) and pages(1005) we decided to create a balanced subset of the whole training set and try to evaluate the algorithms against this two different sets:

- Set A is the balanced subset containing 1100 male profiles, 1100 female profiles and 1005 pages.
- Set B is the whole unbalanced classified set

Each set of profiles has been divided randomly between 70% training data and 30% test, using the training data to feed four different classifiers:

- NB: Naive Bayes
- LR: Logistic Regression

The tests on the accuracy of each classifier has been done by considering four different values for the number of features:

1. 65536
2. 131072
3. 262144
4. 524288

The maximum number of features can be set a priori since we compute the Term Frequency by using the hashing trick.

RESULTS FOR SET A:

The best results for set A were reached with a number of features equals to **262144**. Anyway the difference with the others configuration was of few points on percentage.

	NB TF	**NB TF-IDF**	**LR TF**	**LR TF-IDF**
Total	70.03	68.03	66.04	66.04
Male	70.78	66.27	65.36	65.36
Female	59.62	56.73	56.09	56.09
Page	79.80	81.43	76.87	76.87

RESULTS FOR SET B:

The best results for set A were reached with a number of features equals to **262144**. Anyway the difference with the others configuration was of few points on percentage.

	NB TF	**NB TF-IDF**	**LR TF**	**LR TF-IDF**

Total	63,11	66,12	67,10	67,10
Male	98,44	86,18	82,09	82,09
Female	5,06	30,38	40,51	40,51
Page	41,47	57,53	60,54	60,54

TOP 10 WORDS FOR MALES:

WORD	FREQUENCY
web	50
digital	49
developer	46
geek	41
fan	40
media	40
photographer	38
own	36
father	30

TOP 10 WORDS FOR FEMALES:

WORD	FREQUENCY
love	64

lover	48
life	41
media	39
music	38
digital	35
writer	34
i’m	34
food	32

TOP 10 WORDS FOR PAGES:

WORD	FREQUENCY
news	166
official	73
twitter	64
from	58
your	40
account	38
more	37
follow	33
notizie	30

4.5 PULLING THINGS ALTOGETHER: GRA APPROACH

After evaluating the performance of each sub system, the next step is to put all the partial results altogether in a single algorithm that from a twitter user profile, even if it does not contain all the required fields (for example missing description or unknown name) can be able anyway to classify it.

In other words GRA algorithm is a multiple classifier system. The problem to deal with how to combine the results of the 3 classifiers. We have decided to follow two different approaches:

- decision tree evaluation
- waterfall evaluation

The data set is composed of 5000 profiles containing all the required data (all the required fields are not empty) and with profile description in Italian or English (our description training set was composed following these criteria), so we are evaluating the algorithm in the best possible conditions.

DECISION TREE EVALUATION:

For this evaluation the initial data set has been split in two parts: 70% data training and 30% test since for this evaluation the algorithm itself has to be trained, using as training set the combination of the results of the sub algorithms with the expected results. Since the training and the test set are picked randomly from the whole data set we applied the algorithms several times to see how the accuracy percentage change with different sets of data. The results are the following:

	Accuracy (%)

Total	86.69
Male	92.48
Female	78.08
Pages	79.31

As shown from the results a pretty good accuracy is achieved, however the drawback of this solution is that we need another training set to teach to the GRA algorithm how to behave with the possible output of the sub algorithms and in any case all sub algorithms have to be ran in order to achieve the classification.

WATERFALL EVALUATION:

The waterfall evaluation is a very straightforward approach. It works as follow: it takes into consideration different classification sub algorithms sets as C1,C2,C3. It then chooses among these sub algorithms the one that achieve the best accuracy during classification (in our case the name/ScreenName approach achieve a good accuracy when able to recognize the name) and elect it as the first classifier. All the profiles not recognized from the first classifier are used as input to the second and so on.

The advantage of this technique is that if a match is found in the first algorithm there is no need to process the data with the others. However the drawback is that if an algorithm that is on the top of the chain gives a bad prediction (for example due to a name contained in a page type or a description in a language different from the one of the training set) there is no way to correct the result due to the absence of other predictions with which compare the result.

We have chosen the following configuration:

1. Name ScreenName approach as the first classifier since it has a good accuracy when there is a match into the names gender map
2. Description approach since it has a good accuracy to recognize pages
3. Profile colors approach. Despite it is language independent it is the less reliable since in gathered profiles we have noticed an high presence of default colors that makes difficult the classification

With these configurations we have obtained the following results:

	Accuracy
Total	86.76
Male	86.67
Female	90.07
Pages	81.74

As shown from the two tables above the two algorithms in the best conditions (profiles filled with all the required information and with the description in expected languages) have almost the same level of accuracy.

5 CONCLUSION:

In this document we studied gender classification on Twitter. We presented our approach to predict gender from different available fields composing twitter user profile. The approach to use different sub algorithms has different advantages:

- **Specialization**: each sub algorithm focus on a particular field
- **Elasticity:** Even if a field is missing or filled in a manner that is unrecognizable from a sub algorithm, there are other sub algorithms that can in any case classify the profile relying on other parameters.
- **Good Accuracy:** the algorithm provide a good classification accuracy in total and on the specific genders

In the future, we intend to study different characteristics of the dataset to classify gender (additional fields like url and profile image) and to incorporate them in our framework. Moreover we want also to try to figure out possible improvements into our multi classifiers system to make it more reliable in all the possible cases.

ACRONYMS:

GRA - Gender Recognition Algorithm

SDA - Social Data Aggregator

TF-IDF - Term Frequency–Inverse Document Frequency

REFERENCES:

[1] Gender Prediction on Twitter Using Stream Algorithms with N-Gram Character Features,July 26 2009

[2] Language Independent Gender Classification on Twitter,2013

[3] http://www.oxforddictionaries.com/words/how-many-words-are-there-in-the-english-language

[4] http://www.lingholic.com/how-many-words-do-i-need-to-know-the-955-rule-in-language-learning-part-2/

[5] http://spark.apache.org/docs/0.8.1/api/mllib/org/apache/spark/mllib/regression/LabeledPoint.html