University of Management and Technology

Machine Learning

Assignment 4

Text Classification in Scikit-Learn

Submitted To

Sir Haseeb Younis

Submitted by

Muhammad Abdullah Shahid

F2019027004

School of Professional Advancement, UMT

C-II, Johar Town, Lahore

Introduction

This assignment is about text classification. We train ML Algorithms on a sample text data and then model will predict on real time that which text is written by a male and which text is written by female. We have PYN dataset this dataset have 215 text files about female comments and tweets and 210 text files about male comments and tweets. In this assignment we will train Ml algorithm with three different approaches or methods.

1. Content based methods
2. Stylometry based methods
3. Text Clustering

1. Work on task 1 Content based methods

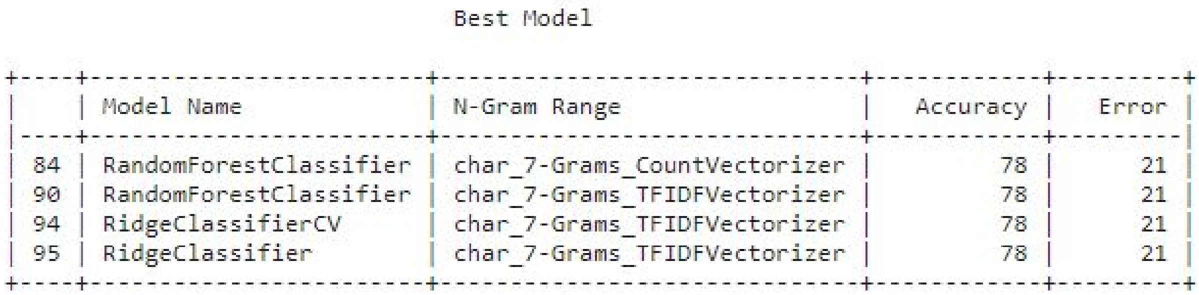
In content based methods 1st we steps load all male text files into python list and convert this list into pandas DataFrame and add the male label with all these instances. Than we load all female files in DataFrame and add female label with that. Than combine these two DataFrame and visualize this dataset through graphs and different techniques. Do some preprocessing on that dataset like remove punctuation and stop words after preprocessing we do label encoding on gender column.

After that we start the feature extraction step and we do this phase with **Word uni-grams, bi-grams, tri-grams** and **character 3 to 10 grams** with **max_features=8000** and we do feature extraction with **CountVectorizer and TfidfVectorizer** methods. After this phase we have total 22 feature vectors.

Than we split data into train and test data than train 6 ML Algorithms on that 22 feature vectors these Algorithms are:

1. RandomForestClassifier
2. BernoulliNB
3. ExtraTreeClassifier
4. GaussianNB
5. RidgeClassifierCV
6. RidgeClassifier

And we store all Algorithms prediction (22 * 6 = 132) into a DataFrame and check that which Algorithms gives the best performance on which feature vector and we that find these Algorithms gives the best performance on these feature vector.



```
                          Best Model

+----+-------------------------+----------------------------+------------+---------+
|    | Model Name              | N-Gram Range               |   Accuracy |   Error |
|----+-------------------------+----------------------------+------------+---------|
| 84 | RandomForestClassifier  | char_7-Grams_CountVectorizer |         78 |      21 |
| 90 | RandomForestClassifier  | char_7-Grams_TFIDFVectorizer |         78 |      21 |
| 94 | RidgeClassifierCV       | char_7-Grams_TFIDFVectorizer |         78 |      21 |
| 95 | RidgeClassifier         | char_7-Grams_TFIDFVectorizer |         78 |      21 |
+----+-------------------------+----------------------------+------------+---------+
```

Than we select **RidgeClassifierCV** as the best model and train this best model on best feature vector that is generate by TfidfVectorizer and set **ngram_range=(3,7), max_features=8000,** after that we execute this model and take some unseen data and do the same preprocessing on that unseen data and convet this input into same feature vector and then pass this input to model and model gives the prediction.

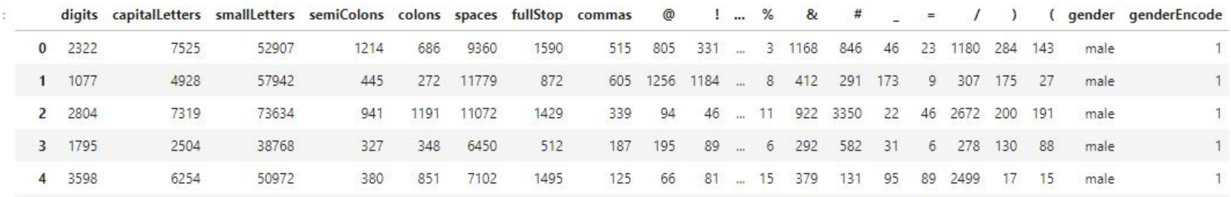
Conclusion:

We find that **TfidfVectorizer** with **ngram_range=(3,7) and max_features=8000**, the best feature vectorizer for this dataset and RidgeClassifierCV and RandomForestClassifier are best model for that feature vector.

2. Work on task 2 Stylometry based methods

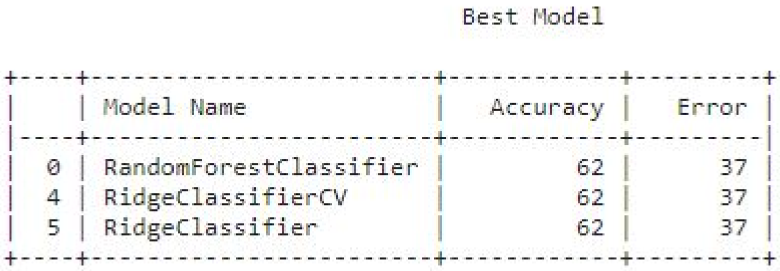
In stylometry based methods 1st is almost same that wo do in content based method after that we just remove the stope words from the data and do label encoding on gender.

In feature extraction step we write a function for extract the features in **20 different stylometery** which given in assignment 4 requirement. That function count the stylometery and add their sum in that particular row and column that shows in figure.



	digits	capitalLetters	smallLetters	semiColons	colons	spaces	fullStop	commas	@	!	...	%	&	#	_	=	/	)	(	gender	genderEncode
0	2322	7525	52907	1214	686	9360	1590	515	805	331	...	3	1168	846	46	23	1180	284	143	male	1
1	1077	4928	57942	445	272	11779	872	605	1256	1184	...	8	412	291	173	9	307	175	27	male	1
2	2804	7319	73634	941	1191	11072	1429	339	94	46	...	11	922	3350	22	46	2672	200	191	male	1
3	1795	2504	38768	327	348	6450	512	187	195	89	...	6	292	582	31	6	278	130	88	male	1
4	3598	6254	50972	380	851	7102	1495	125	66	81	...	15	379	131	95	89	2499	17	15	male	1

Than we split data into train, test data than train **same 6 ML Algorithms** in this dataset and compare the accuracy of all these Algorithms and select **RandomForestClassifire** as best model.



```
                    Best Model

+----+------------------------+------------+---------+
|    | Model Name             |   Accuracy |   Error |
|----+------------------------+------------+---------|
|  0 | RandomForestClassifier |         62 |      37 |
|  4 | RidgeClassifierCV      |         62 |      37 |
|  5 | RidgeClassifier        |         62 |      37 |
+----+------------------------+------------+---------+
```

And train RandomForestClassifire on all data. Than execute this model and tame some unseen data from user and convert that input into feature vector through that same function and pass this feature vector to model and model gives the output.

Conclusion:

We convert text data into **20 different features** on symbols base, than some Algorithms on that data and select one of best Algorithms and train this Algorithms on all data. After their training we give some unseen data and model gives the output.

3. Work on task 3 Text Clustering

In this task we load all text file into DataFrame and do not give the label because we want to implement unsupervised approach so we do clustering of our data because when we load all text file into Dataframe than we do not gives the label and that data is consider as unlabeled data.

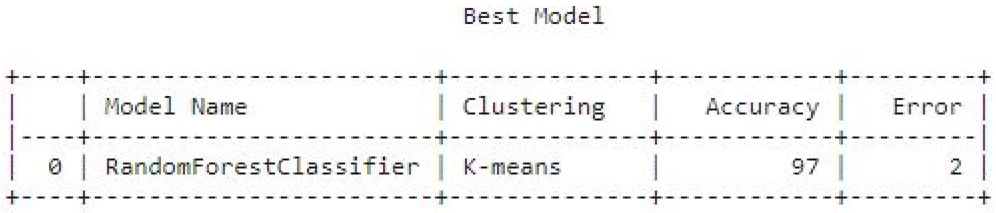
Than we clean this data and convert this text data in feature vector by using **TfidfVectorizer** and set their parameters value **ngram_range=(3,7) and max_features=8000** because we see in task 1 that this feature vector method and values are best for this dataset.

Use two clustering algorithm:

1. K-means clustering
2. MiniBatchKMeans

Try each algorithm one by one with parameter value **n_clusters = 2** and predict the positive and negative instance in our data and pass matrix which generate by **TfidfVectorizer** for prediction. After that we will train some model and find the prediction and measure the model accuracy.

After that we attach these cluster as label with our dataset one by one. And train same 6 ML Algorithms on these two clusters' labeled dataset and measure the accuracy and select the best model and best cluster.



```
                       Best Model

+----+------------------------+-------------+-----------+---------+
|    | Model Name             | Clustering  |  Accuracy |   Error |
|----+------------------------+-------------+-----------+---------|
|  0 | RandomForestClassifier | K-means     |        97 |       2 |
+----+------------------------+-------------+-----------+---------+
```

On this dataset we find **K-means** as best cluster and **RandomforestClassifier** as best ML Algorithm and pass complete data set for the training of application phase. After training we pass unseen data and get some predicted output from best model.

Conclusion:

Load all unlabeled data and convert this data into feature vector than use **K-means** for clustering and attach that clustering list with data as output labeled and train **RandomforestClassifier** on that clustered labeled data.