Author Classification Model

—

By Diprotiv Sarkar

The Problem Statement

The Problem Statement

An additional tool, built to filter the opinions regarding your product!

We have to categorize social media authors as Healthcare Professionals (HCPs) or Non-HCPs to gain great

clarity from the opinions mined in subsequent steps.

Data Collected

The following are the features available in the data-set:-

- **Twitter Handle:** Represents the unique Twitter profile ID belonging to the author
- **Name, Location:** Stores the fundamentals identification details for the author *(filled by himself, in the social media)*.
- **Description:** Stores the actual bio (details) regarding the author, and can be viewed as useful information.
- **Followers_count:** Number of followers the author has.
- **Following_count:** Number of other twitter users the author is following.

WE NEED TO PREDICT?

A categorization of the author, (this field is empty in the submission file, and our model will predict the probability of HCP) called the **HCP_Flag.**

- **A value of 1 indicates that the person is associated with Health-Care, as a professional.**
- **A value of 0 indicates otherwise.**

STEP 1:
Data Preprocessing and Analysis

Looking into the data-set, we divide the stages into following:-

(1) Removal of unnecessary columns.
(2) Translation of Data (Profile Description)
(3) Feature Engineering
(4) Final Analysis and Preprocessing

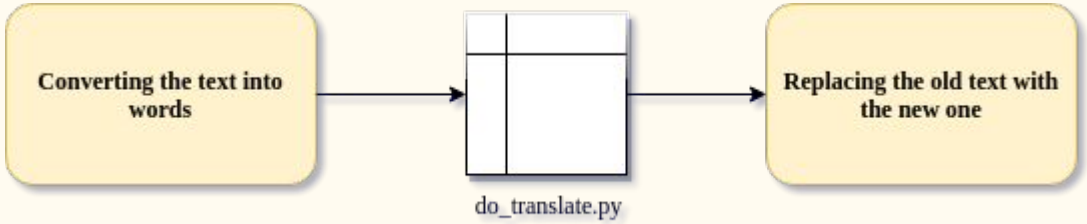
Removal of Unnecessary Columns

	UID	Handle	Name	Description	Loca	llowers_count	Following_count	HCP_flag
0	4778	oncoblogbulbul	Oncoblog	Oncólogo médico. Músico. Lector empedernido. P...	aN	1133	592	1
1	6542	ritrattosalute	Ritrattodellasalute	Il ritratto della salute è il primo progetto i...	NaN	700	903	0
2	13355	juan3punto14	juan 3.14	NaN	Earth	1999	1318	0
3	1764	rachaeyyy	Rachael Morrison	USCA '18	NaN	408	306	0
4	14082	allahbas9	علاء الشمري	گسرتْ اَلْدِنْيَآ گآ ِلْيَ فَاَقَعَدِتْتِيَ ...	NaN	1442	1441	0
5	10929	stephrstarr	Stephanie Starr MD	General pediatrician & medical educator (UME h...	Rochester, Minnesota	96	159	1
6	14632	penny_donnely	Penny Donnely	The OFFICIAL Twitter page of Actor, Producer, ...	San Diego, CA	1	34	0
7	875	karenychin		NaN	NaN	57	83	1
8	7390	joperezxx		.Mermaid.	Fearfully & Wonderfully Made	106	244	0
9	7011	maditini	Madi	NaN	NaN	22	43	0

These columns are not beneficial for our prediction model, they make no logical sense.

Translation of Data

We see that a lot of the description data is belonging to different languages! So training our model and pre-processing the text may get difficult! For this purpose we translate everything in the ‘Description’ column to English.



Feature Engineering

After translation, our data-set is looking clean, with a lot of NaN values, and rows of the data, containing garbage values like '**I**', '**am**', '**so**', etc.

- These values will make no difference in our model predictions, because they appear everywhere!
- We need to further refine the text-data so that all the alpha-numerica tokens and special characters are further reduced.

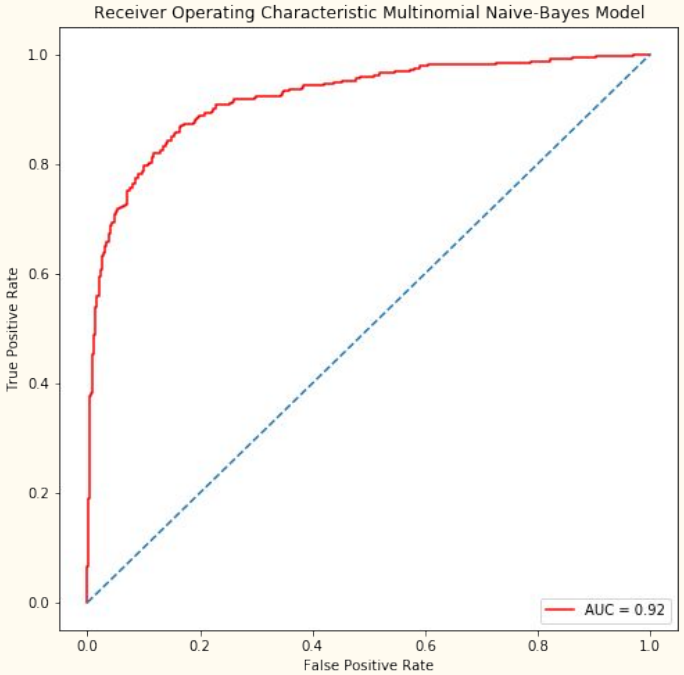
Steps Taken!

Data-Refining Along with Analysis

- Looking at Frequency of Words from both positive and negative results.
- Extracting the Stop-Words and Removing them from data.
- Use of Regex to only keep proper English words (size > 2)
- Using the TfidfVectorizer

STEP 2:
Model Preparation

MultinomialBias Classifier



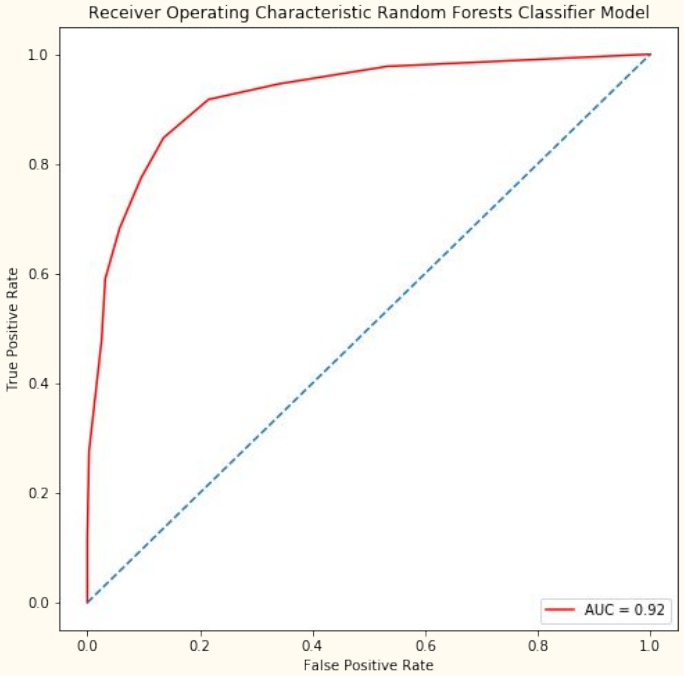

```
Confusion Matrix for Multinomial Naive-Bayes Model
[[594  33]
 [ 89 226]]
Classification Report for Multinomial Naive-Bayes
Model
              precision    recall  f1-score   support

           0   0.869693  0.947368  0.906870       627
           1   0.872587  0.717460  0.787456       315

    accuracy                       0.870488       942
   macro avg   0.871140  0.832414  0.847163       942
weighted avg   0.870660  0.870488  0.866939       942

Area under under ROC curve for Multinomial
Naive-Bayes Model
0.9207311207311208
```

RandomForest Classifier



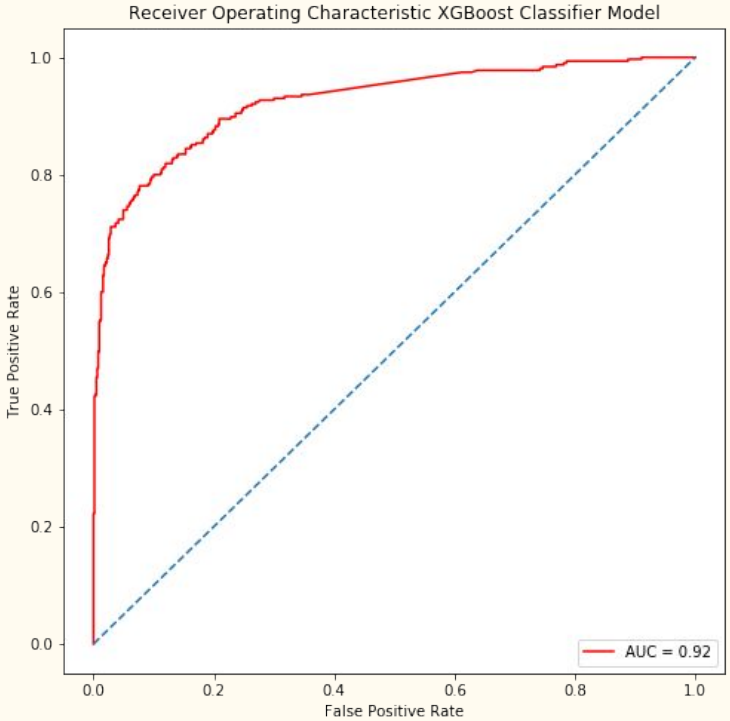

```
Confusion Matrix for Random Forests Classifier Model
[[594  33]
 [ 89 226]]
Classification Report for Random Forests Classifier
Model
              precision    recall  f1-score   support

           0   0.869693  0.947368  0.906870       627
           1   0.872587  0.717460  0.787456       315

    accuracy                       0.870488       942
   macro avg   0.871140  0.832414  0.847163       942
weighted avg   0.870660  0.870488  0.866939       942

Area under under ROC curve for Random Forests
Classifier Model
0.9229310650363283
```

XGBoost Classifier




```
Confusion Matrix for XGBoost Classifier Model
[[594  33]
 [ 89 226]]
Classification Report for XGBoost Classifier Model
              precision    recall  f1-score   support

           0   0.869693  0.947368  0.906870       627
           1   0.872587  0.717460  0.787456       315

    accuracy                       0.870488       942
   macro avg   0.871140  0.832414  0.847163       942
weighted avg   0.870660  0.870488  0.866939       942

Area under under ROC curve for XGBoost Classifier Model
0.924915824915825
```

STEP 3: Making the Final Predictions

Hypothesis support

Why did I not use CountVectorizer?

The ultimate aim to specifically predict all the rows which are belonging to Health-Care Professional.
Now, most of the description may contain rare terms like ‘Neurologist’, ‘Pediatrician’, which are otherwise not seen in normal languages. We have to give a high-weightage to these rare terms, and it has to get nearly-equal importance as terms like ‘health’, ‘doctor’

Choosing XGBoost as the final model

- This is better than a single-decision tree, or Logistic Regression, as far as regularization and overfitting is concerned.
- It consists of an army of predictors, (simple trees, logistic regression), and each one of the predictors are built sequentially, based on the mistakes of previous predictors.

Conclusion

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip.

Thank You!