Tweets classification

CONTENTS

1. Problem Statement:

Build Classification model to identify sarcasm. Given dataset includes tweets, tweet id and target variable (sarcastic/non- sarcastic). We are trying to predict the sentiment of tweets if they are sarcastic or non-sarcastic. We are given a dataset which includes tweets and sentiment of tweets.

1.1 Business Understanding:

1 billion tweets generates every day. If we can predict the sentiment of tweets this can help us in many ways. Like predicting movie review. Nowadays Peoples tweet about their experience of movie. So if we can predict the sentiment of tweet we will be able to predict the review of the movie. We can also predict the review of some product. We can predict that what peoples are thinking about the latest decision issues by some authority or government. In the same way finding sentiment of tweets can help us in many other ways also.

2. Understanding of data

We have 3 column/variable in our dataset. First is tweets id that represent the ID of tweets. Second is tweets for that we have to predict if tweet is sarcastic or non-sarcastic using the third column which is our target variable

2.1 Variable Importance

Our data is in the form of text. Main component of our data is tweet. We are building our model on top of it. We can remove the tweet id from our dataset as we are not using tweet id to build our model. Our target variable contain only 2 type of output, sarcastic and non-sarcastic, So we can convert that to factor with labels 0 and 1. We will take the tweets out of dataset and do all the pre-processing steps on this data only. We have 90,000 rows in our dataset so to increase the processing time and model building time I choose 20,000 rows. Because RAM of my system is little less and R do all the processing in RAM memory. So to avoid this problem I choose 20k rows.

3 Exploratory data analysis

Exploratory data Analysis is the most imp step in any model building. We have to clean and pre-process our data. As our data is in the form of text. And we are using only tweets as our data, so first we have to make corpus and then do the required pre-process. And convert it into document Term matrix. And then convert it into data frame.

3.1 Pre-processing steps.

1. Uppercase to lower case
2. Remove numbers
3. Remove punctuation
4. Remove stopwords
5. Stemming
6. Remove white spaces

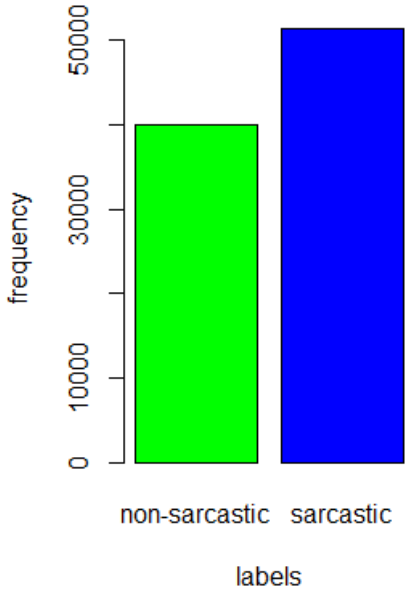
4 Insight of target variable

4.1 frequency of target variable (sarcastic/non-sarcastic)

table (dataset$label)

non-sarcastic	sarcastic
39998	51300

4.1 Bar plot of frequency of target variable



4.2 Word Cloud:

5 Model building

Data is divide into train data and test data in the ratio 70:30

I used two machine learning algorithms to build classification model.

Naïve Bayes

Model was built on training data and then tested on test data. Accuracy was very less for naïve bayes. It was around 60%

Confusion Matrix

	y_pred	
	0	1
0	2176	463
1	1995	1366

Random Forest

Random Forest was built to predict the sarcasm in tweet. Accuracy or random forest is much high as compare to naïve bayes. Accuracy of Random Forest is around 80%.

confusion matrix

	y_pred_rm	
	0	1
0	2036	603
1	624	2737

6. Conclusion

we are able to predict the sarcasm in tweets with 80% accuracy using random forest