Roche

Interview Exercise 2

Dawid Sitnik

1.Objective
The aim of this exercise is to analyze a dataset of headlines and classify them as sarcastic or not using binary classification model. I decided to solve the whole problem using only python.

2.Thinking Process
The classic steps in Data Science process are:
-acquire
-prepare
-analyze
-report
-act
In this particular case, data was already acquired, so the first thing which I did was the preparation of a dataset. After my analysis I found out, that it was already well-formatted – it was a JSON file consisted of two fields: 'headline' and 'is_sarcastic' which is 0 for not sarcastic headline and 1 for sarcastic one. To ensure that there aren't any inconsistencies in the dataset I got basic information about it.

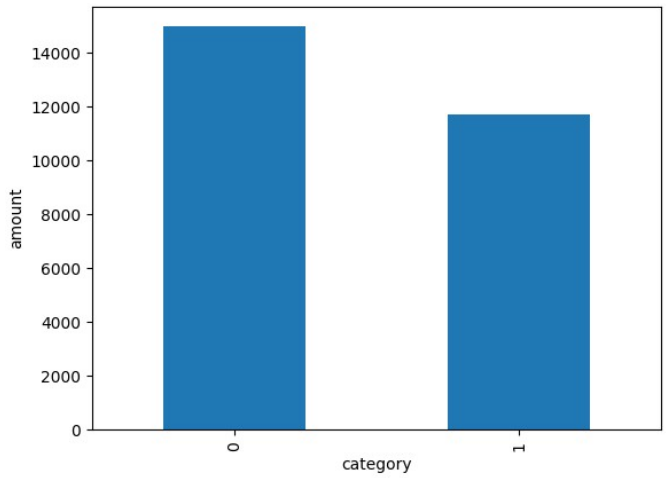
```
RangeIndex: 26709 entries, 0 to 26708
Data columns (total 2 columns):
headline        26709 non-null object
is_sarcastic    26709 non-null int64
dtypes: int64(1), object(1)
memory usage: 417.4+ KB
None
```

As we see there are no null values in the dataset, so the data can be processed.

I was also looking for any overlapping data. I couldn't find any so I started another data analyzing part.

To make this process more clear I printed some plots. Next, I tried out different models to check which performs the best. In the end, I drew a confusion matrix to visualize my classification result and tried to figure out why the model makes a decision it made printing weight of each word.

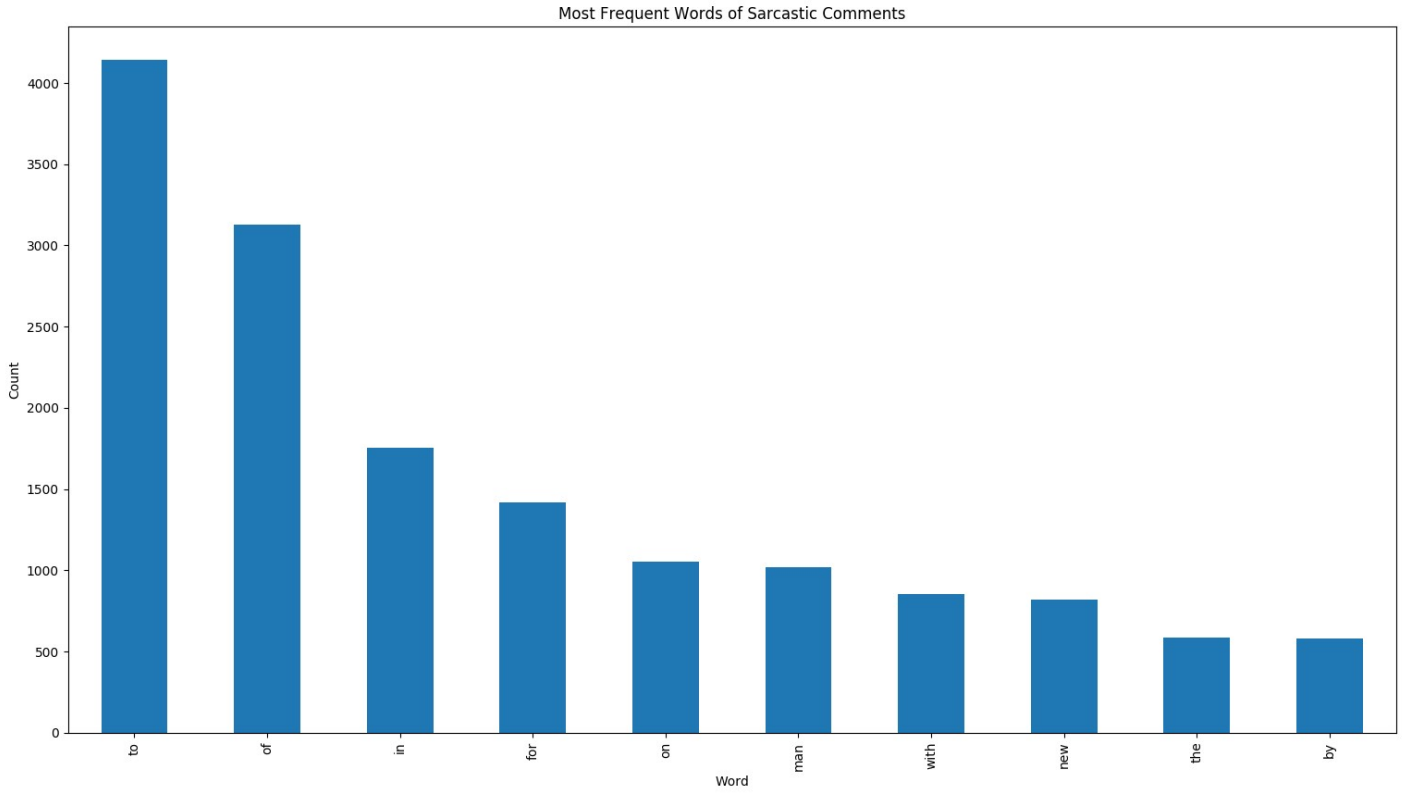
3. Data Visualization
The aspect with which I want to start my analysis is the amount of sarcastic and not sarcastic headlines.

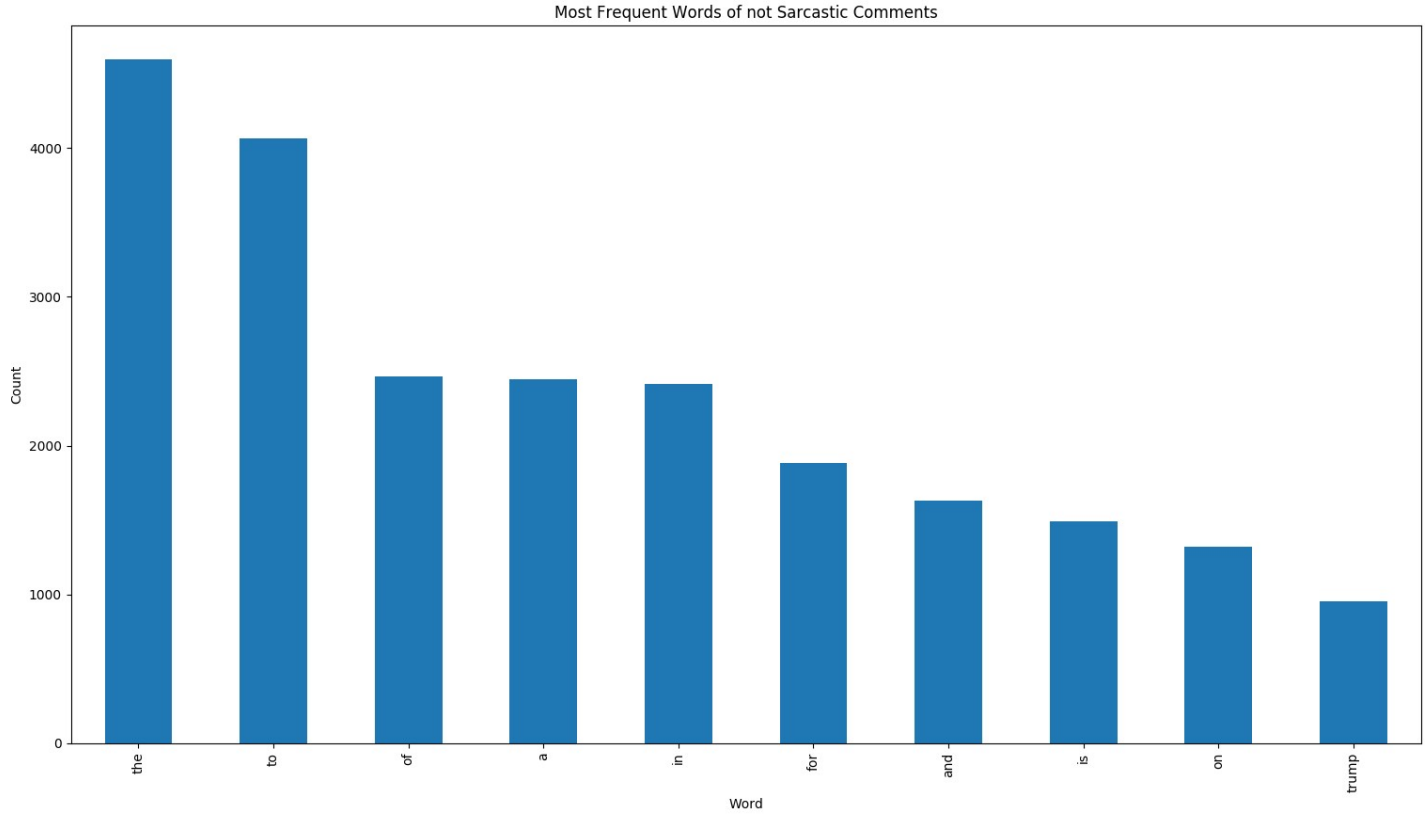


Amount of Sarcastic and Not Sarcastic Words

According to the picture their quantity is rather balanced, however, there is the small advantage of non-sarcastic headlines.

Then I count and display the most popular words for two groups





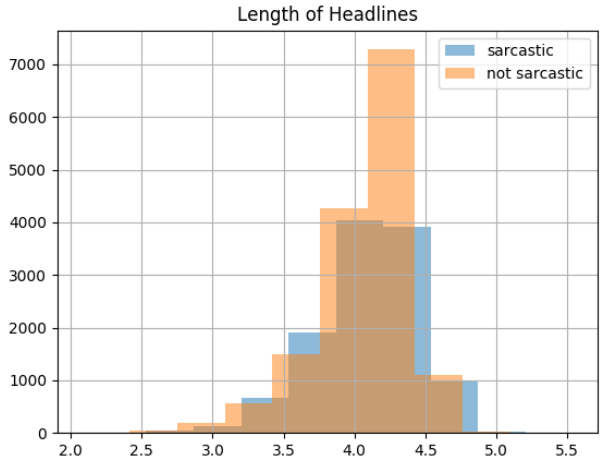
As we can see, the most popular words are mostly prepositions like to, the, of, etc.

To visualize the popularity of longer words I used word map and get those results:

For sarcastic headlines, words which definitely out stands are a report, new, man, woman, leave, braking, etc.

For non-sarcastic I got: button, time, trump, election, Cruz, new, government, etc.

In the end, I would like to show the distribution of lengths for sarcastic and normal headlines. They are quite similar. The only difference is between words of 4.5 lengths. There is an advantage of non-sarcastic words.



4. Modeling

4.1 Steps To Be followed When Applying an Algorithm

1. Split the dataset into training and testing dataset.
2. Select classification algorithm
3. Train algorithm on dataset
4. Pass the testing data to the trained algorithm to predict the outcome
5. Check the accuracy by passing the predicted outcome and the actual output to the model.

4.2 Split Dataset into a Training Set and a Testing Set

At the beginning of modeling process I started with splitting data into training and testing sets.

Advantages

- By splitting the dataset we can train using one set and test using another.
- This ensures that we won't use the same observations in both sets.
- More flexible and faster than creating a model using all of the dataset for training.

Disadvantages

- We can get different results depending on how the set was spitted
- We can solve this problem using k-fold-cross-validation. In have already used this method in previous exercise, so now I decided to skip this part.

4.3 Select classification algorithm

I decided to use different algorithms and see which one will give the best result. I used:

- Logistic Regression
- SVC
- Linear Discriminant Analysis
- GBoost
- Mnb

As there are many of them I am going to describe one.

Accuracy of my models is shown bellow:

```
Accuracy of Logistic Regression classifier on test set: 0.846
Accuracy of Random Forest Classifier classifier on test set: 0.808
Accuracy of SVC classifier on test set: 0.852
Accuracy of GBoost classifier on test set: 0.825
Accuracy of Mnb classifier on test set: 0.840
```

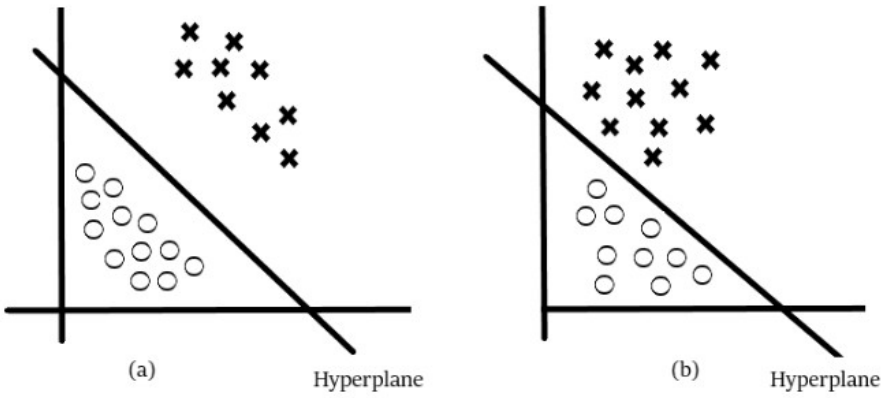
As we can see all the results are quite similar, but the best algorithm for that case was SVC with accuracy 0.852. The worst one was Random Forest Classifier with 0.808.

4.4. Best algorithm explanation:

What is SVM?
Support vector machines is a supervised algorithm used for classification and regression problems. It is used for the smaller dataset as it takes too long to process

The ideology behind SVM:
SVM is based on the idea of finding a hyperplane that best separates the features into different domains.

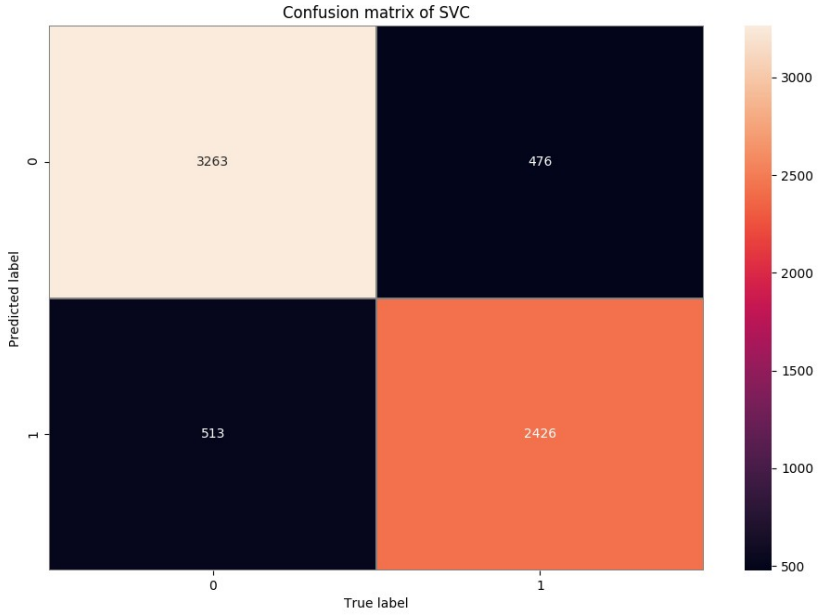
Intuition development:
Let's consider our situation:
We have headlines that are sarcastic or not and we want to design a function (hyperplane) which will clearly differentiate those two groups.



Consider those two figures with drawn hyperplane. As we can see the first one makes celearer classification so we would pick this one.

Basically, SVM bases on choosing an optimal hyperplane which will clearly classify the different classes(binary in this case).

To visualize my result I have also printed its confusion matrix.



As we can see, the total size of the testing data is 6678, which is more less 25% of the whole data. Looking at the matrix, we can see how many mistakes our algorithm did and in which situations. The number of proper classifications for testing set was 3263 + 2426 and the number of mistakes is the sum from black squares which is 476 + 513.

4.5. Words weights

There are two main ways to look at a classification model:
- inspect model parameters and try to figure out how the model works globally
- inspect an individual prediction of a model, try to figure out why the model makes the decision it makes.

I decided to follow the first proposition and used show_weights() function from the ELI5 library. This function prints us weights of each word, starting from the most important ones. The biggest score word got the biggest impact on the classification it has.

```
y=1 top features
 Weight  Feature
-------  -------
 +6.496  nation
 +6.170  man
 +5.896  area
 +5.104  report
 +4.274  of
 +3.606  local
 +3.552  only
 +3.132  he
... 14568 more positive ...
... 15161 more negative ...
 -3.112  donald
 -3.936  my
 -3.995  this
 -3.996  are
 -4.068  how
 -4.422  an
 -4.488  why
 -5.343  your
 -6.686  trump
 -7.471  and
 -7.546  is
-11.752  the
```

Words Weights

As we can see at the picture, the words which have the biggest impact on our classification were: nation, man and area, the least important ones were short words like the, and, is (the one which we could see on the plot as the most frequent)