GA DSI Project 3

(*Web API & Reddit Classification*)

- By *Solomon*

Contents:

1. Overview

- **Proof-of-Concept** Classification Model
- The scope of this project will be limited to 2 subreddits: ***Romance*** & ***Horror***
- Prevent the wrong contents from being in the wrong subreddit and causing unnecessary unease of the subreddit followers

2. Problem Statement

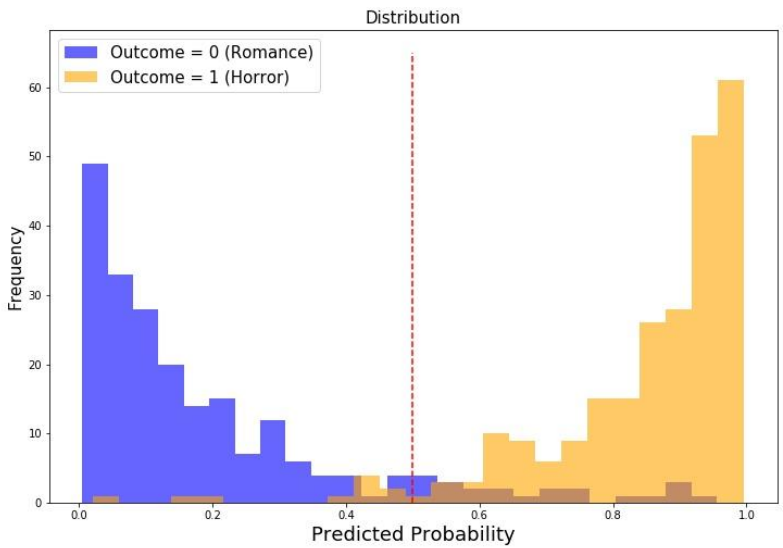
I. Based on the 2 datasets extracted from Reddit, are we able to *build a classification model which is able to accurately classify the text submitted*.

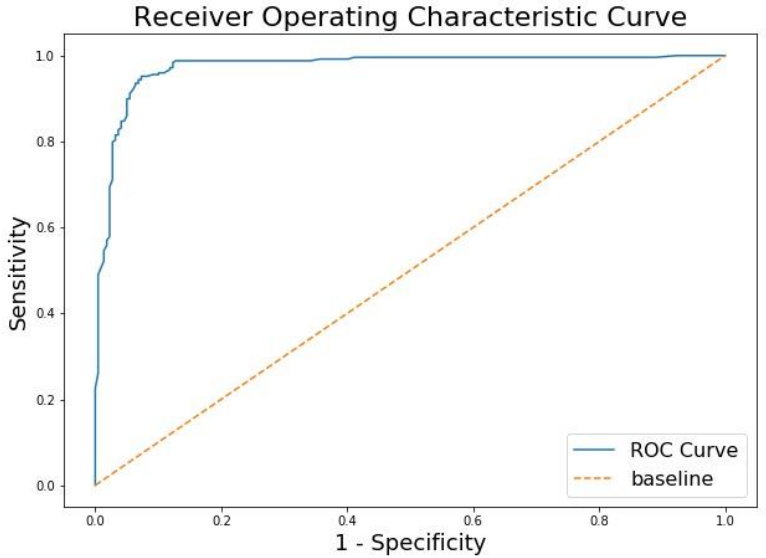
II. What are the *features/words that has the most impact* on the text being classified as a category

III. Using our model as a proof-of-concept, ***how do we proceed from here to implement a model to classify between the correct subreddit topic vs everything else.***

3. **Model Evaluation** (*MultinomialNB Classifier*)

Accuracy: 93.2% **Sensitivity:** 91.2% **Specificity** 90.9%





Comments:

- Based on the objective of our project the best metric in this instance would be Sensitivity. Erring to the safer side and reducing the False Negatives will help us capture more horror entries which are not supposed to be in the Romance subreddit section. This will help us further prevent readers in the Romance subreddit from reading texts which might make them feel uncomfortable.

- From our AUC-ROC Curve & the Predicted Probability Distribution, we can conclude that the features leading to the classification of these 2 categories do not overlap by much. This is also evident in the concurrently high sensitivity & specificity.

4. Possible Inferences

Rank	Word	Log Probability
1	eye	-6.099409
2	door	-6.150998
3	man	-6.313227
4	night	-6.337655
5	know	-6.399682
6	room	-6.408069
7	day	-6.416506
8	thing	-6.463643
9	didn	-6.498178
10	people	6.502723

From the list of top 10 words, it certainly does seem like those are words which someone would be unlikely to use when writing a text requesting romance advice. ***However, it also seem to me that the words can be considered generic when it comes to texts on topics other than romance and horror***.

Note: *The higher the Log Probability of the feature(word), the higher the influence it has on the classification of our text as horror.*

5. Executive Summary.

Bearing in mind that this project is a proof-of-concept and we are supposed to be proceeding from here to implement a model to classify between the correct subreddit topic vs everything else. The potential impact of our study for this 2 topics is:

- **Our model being build on seemingly generic words**

  This "phenomenon", I believe is a **clear indication of the limitation of our model**.

  Assuming we move on from here to start classifying a singular topic vs everything else (instead of the current topic vs topic), I believe there would be a drastic drop in our model performance simply due to the fact that with "everything else" comes a larger mix of generic words. We would then have a lot of **overlapping features/words leading to a significantly harder job for our model**.

5. Executive Summary.

Plausible Solution:

So far our preprocessing was done using TFIDF which is a normalized version of the CountVectorizer. This method is computationally cheap and easy to build, but its **limitation will only allow us to go so far**. I believe in order to move on to the next stage of classifying singular topic vs everything else problems, we would **need to use more advanced preprocessing techniques** like ***Word Vectorizers*** (BOW & Skip Grams) and run our models on the aggregated vectors of each text. By such, we ***move beyond the simple counting of words to capture relative meaning*** behind groups of words, helping us ease the problem of overlapping features/words.

This solution would however be computationally more expensive due to the use of Deep Learning models to handle the preprocessing.