Breaking Bad or Better Call Saul, That Is The Question!

Introduction

Ella Shafi
Data Scientist, GA

Table of Contents

01.
Problem Statement

Problem Statement:

- ➔ pushshift.io Reddit API was used to export data from two subreddits.
- ➔ Using NLP to train a classifier model to predict which subreddit a given post belongs to.
- ➔ The chosen subreddits are:

Vs.

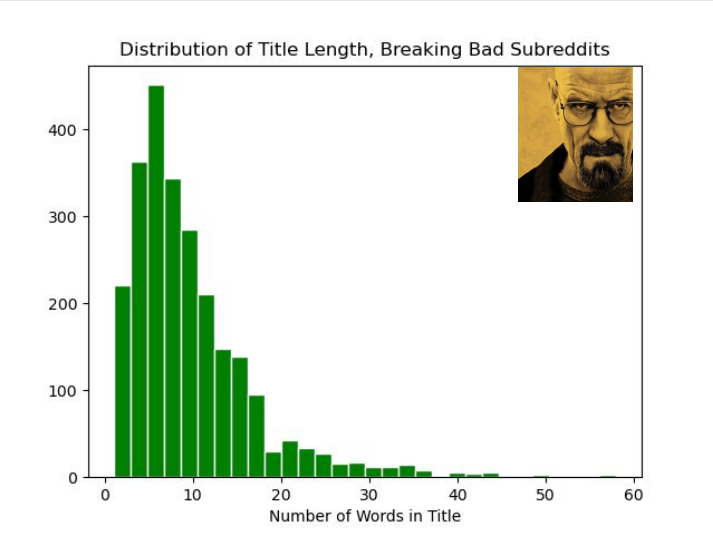
02.
Data Description

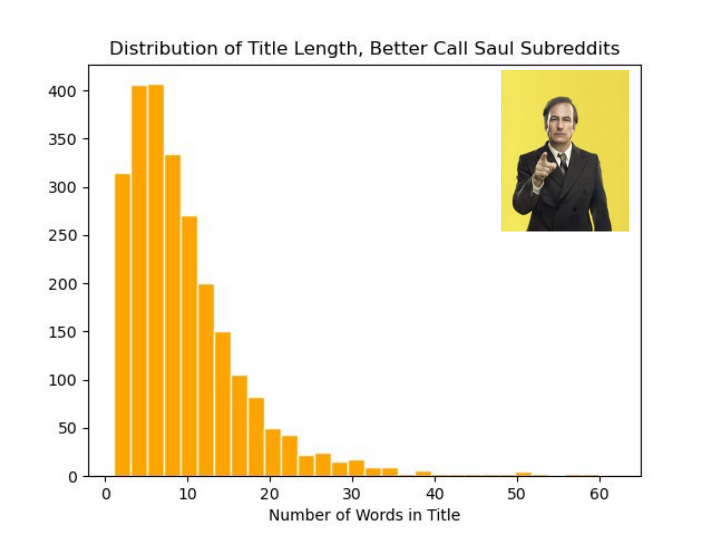
Data Description

- ➔ 2500 posts were exported from each subreddit using pushshift.io Reddit API
- ➔ The features of the dataset are:

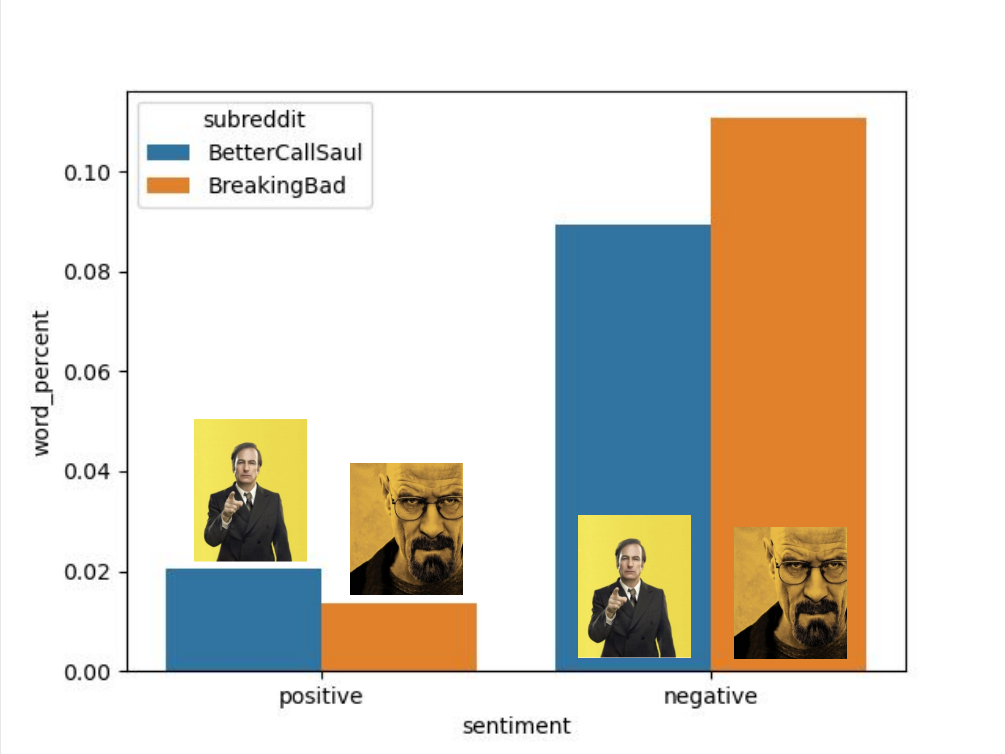
Feature	Type	Description
id	Intiger	Post ID
subreddit	Category	Subreddit Name
selftext	String	Post
title	String	Post Title

Title Length Distribution





Sentiment Analysis of Titles



03.
Modeling & Results

Model features

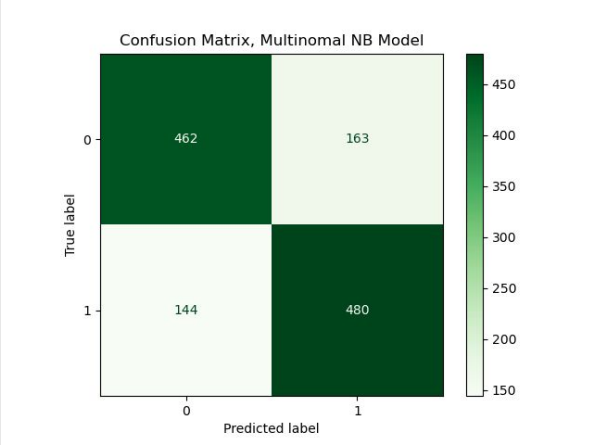
- ➔ Explanatory variable is “post title”.
- ➔ Titles with no words (length 0) were dropped.
- ➔ Test-train split: 0.2 : 0.8
- ➔ Countvectorizer transformer were used to convert text data into a structured, numeric data frame, with 3000 features.
- ➔ Stop word feature used to eliminate common words.

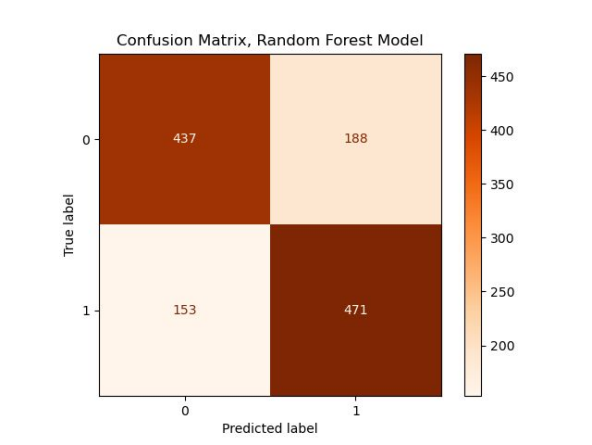
Modeling Results

- ➔ Multinomial Naive Bayes
- ➔ Random Forest

	Multinomial Naive Bayes	Random Forest
Train Score	0.868	0.978
Test Score	0.754	0.734

Classification Metrics





	Multinomial Naive Bayes	Random Forest
Specificity	0.739	0.736
Sensitivity	0.769	0.732
Accuracy	0.754	0.734

04. Conclusion

Conclusion

- ➔ Classifying Subreddits based on titles only (without using the text of the posts) was performed, with fairly small sample sizes.

- ➔ With larger data sizes and better hyperparameter tuning, even higher scores should be achievable.

- ➔ The two models used here are comparable, and neither one has an obvious advantage over the other

Dare To Ask Any Questions?