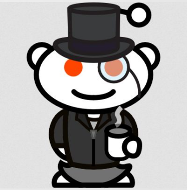
Let’s catch the killer

How accurately can we tag a subreddit to -> Sherlock or Poirot



Audience

VP of Marketing - Commission and budget approvals

Marketing Manager - Strategy and execution

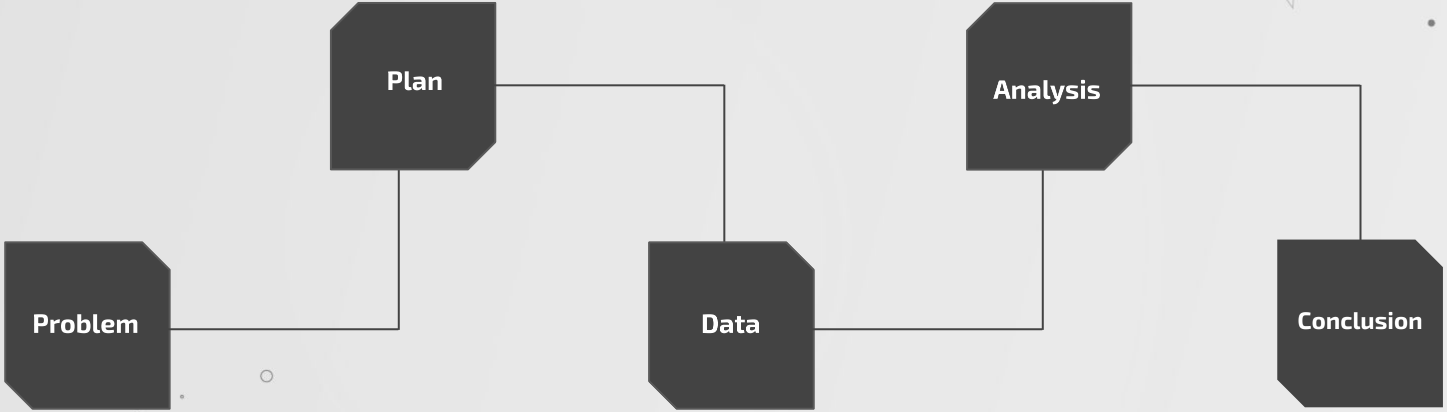
Marketing Operations Manager - Consume the data model downstream for Operations

Problem Statement

One of the large online book publishers wants to commission a fiction detective series and as a part of their larger marketing research (segmenting users) project want to analyse feedback from social media feeds on detective genre

As a part of this exercise, we are going to focus on Reddits comments on two detectives:……***.Mr Holmes and Mr Poirot***

Approach



Problem

Analyse feedback from social media feeds on detective genre

Plan

Focus on creating two data corpuses (of Subreddit ***Text***** for further analysis.

Data

Collect, Manage, Clean the data using PushShift API and Python Libraries

Analysis

Use NLP libraries, LangDetect to derive insights into Subreddit ***Text**.*** *Build a classification model to segregate incoming data.*

Conclusion

*** Provide insights into how useful the Subreddit ***Text***** is and deliver a classification model

*** **Text:** here indicates Subreddit 'Title' and 'SelfText' text data*

**** We may need to revisit the Problem / Plan after Analysis is complete*

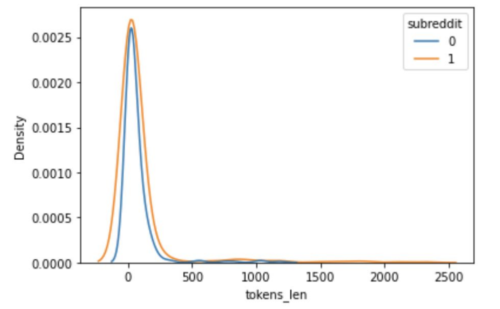
Birds Eye view

67% of Reddit users are Male ; and **64%** are below the age of 29.   This causes and inherent risk and bias in the analysis.

****** Pew Research from 2016*** Pew Research from 2016***

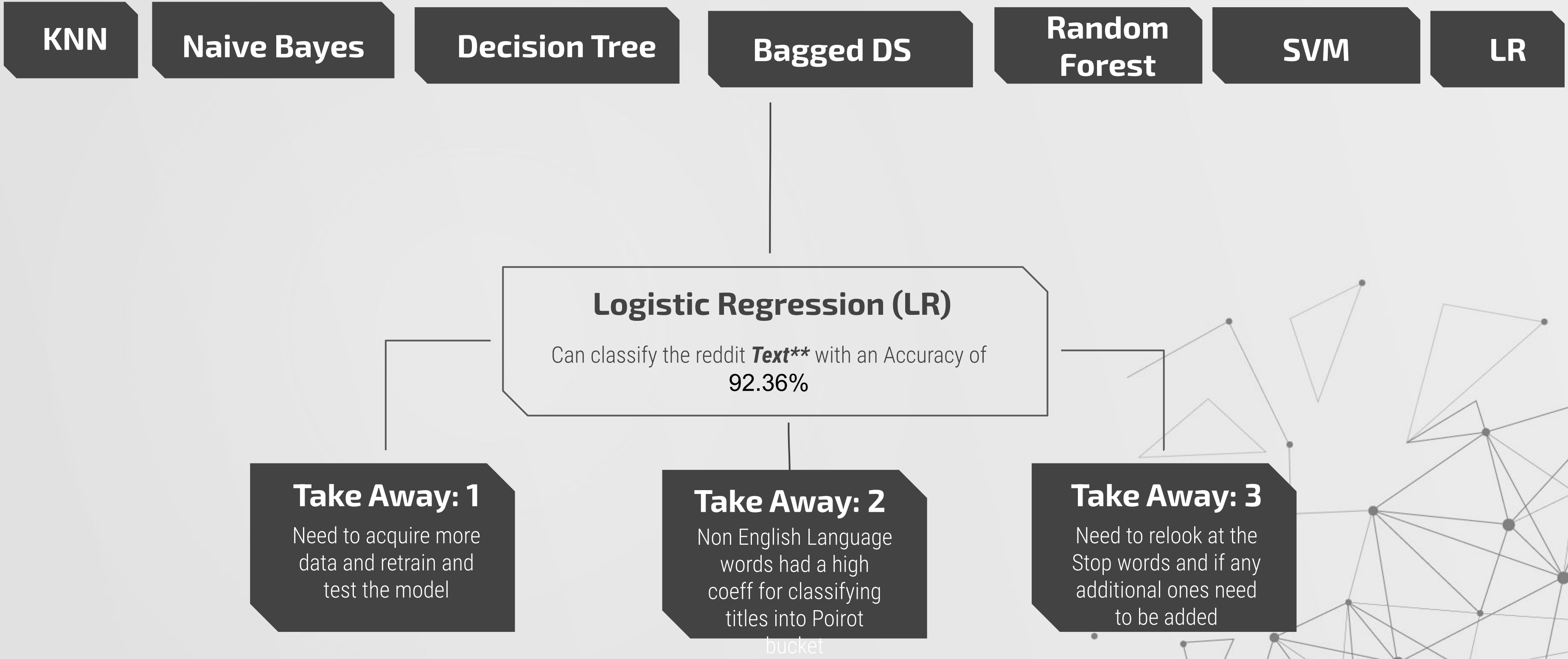
The amount of data for Sherlock and Poirot character ***'book'*** threads is low. Going back to 2012(Poirot) and 2009(Sherlock) , we have collected a total of 768 clean subreddit titles and selftexts.

The average words in a Title of a Subreddit is around 10 words (for both Sherlock and Poirot Subreddits .) However, Sherlock Subreddit threads has more number of 10 word title than Poirot. Could indicate Sherlock Subreddit users are more ***verbose***.



The ***Top*** most repeated Parts of Speech of Text is ***Noun.*** Interestingly, the other Parts of Speech are different for for Sherlock & Poirot Subreddit .

Modelling Approach



KNN

Naive Bayes

Decision Tree

Bagged DS

Random Forest

SVM

LR

Logistic Regression (LR)

Can classify the reddit ***Text***** with an Accuracy of 92.36%

Take Away: 1

Need to acquire more data and retrain and test the model

Take Away: 2

Non English Language words had a high coeff for classifying titles into Poirot bucket

Take Away: 3

Need to relook at the Stop words and if any additional ones need to be added

Citations

Citations:

https://www.journalism.org/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/

http://www.primarydigit.com/blog/analysis-of-reddit-structure-of-the-social-network-and-the-comment-threads

https://github.com/reddit-archive/reddit/wiki/API

Questions?