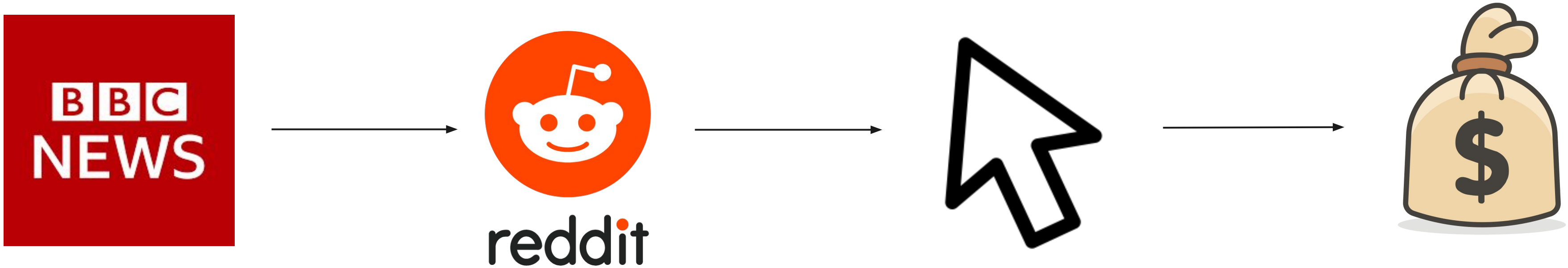
Subreddit Classifier

Problem Statement

- Imagine Reddit wants to venture into the news industry



- The goal is to train a classifier that would become the basis of the bot

Subreddits Chosen

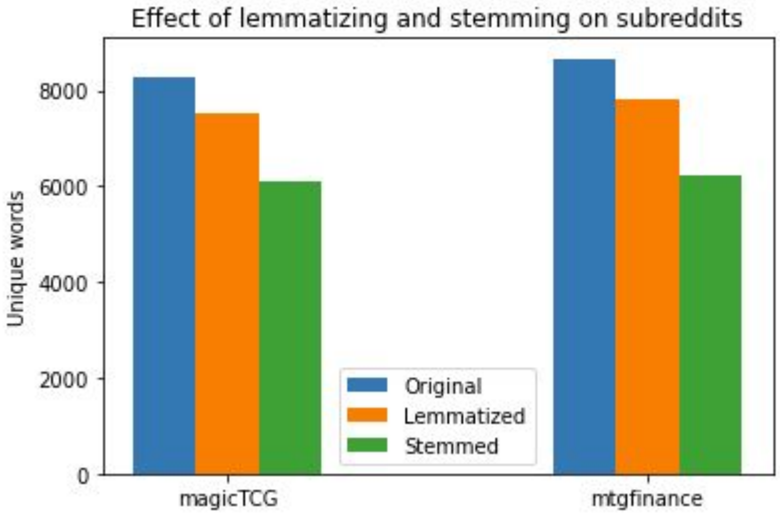
- /r/magicTCG and /r/mtgfinance
- Magic: the Gathering (MTG) is a trading card game (TCG)
- /r/magicTCG is all about playing the game
- /r/mtgfinance is all about playing the market of the game

- A typical /r/magicTCG post:
  - Does the combination of Darksteel Reactor and Tezzeret’s Gambit invoke rule 104.3f?
- A typical /r/mtgfinance post:
  - BGS 9.5 Alpha Black Lotus Ebay Auction closes at $166,100

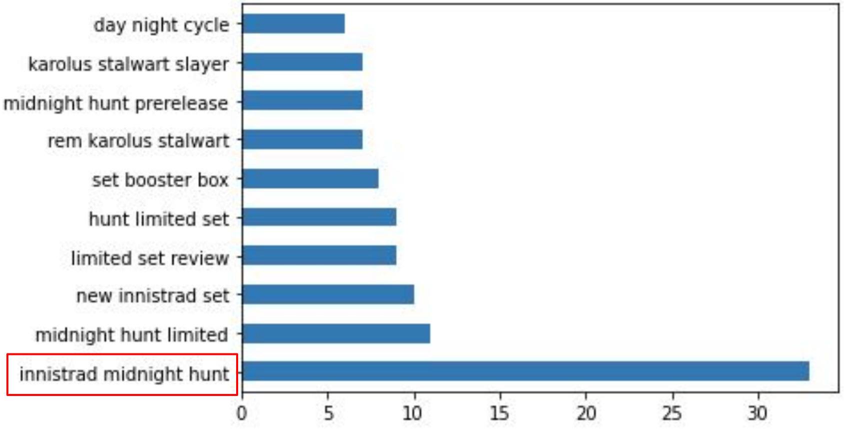
Web Scraping

- Used the Pushshift API
- Had issues with scraping enough data for /r/magicTCG
- Overall had to loop the API 20 times for 2000 posts
- /r/mtgfinance was mostly textual
- Looped 12 times for 1200 posts
- Ended with 922 posts for /r/magicTCG and 957 posts for /r/mtgfinance

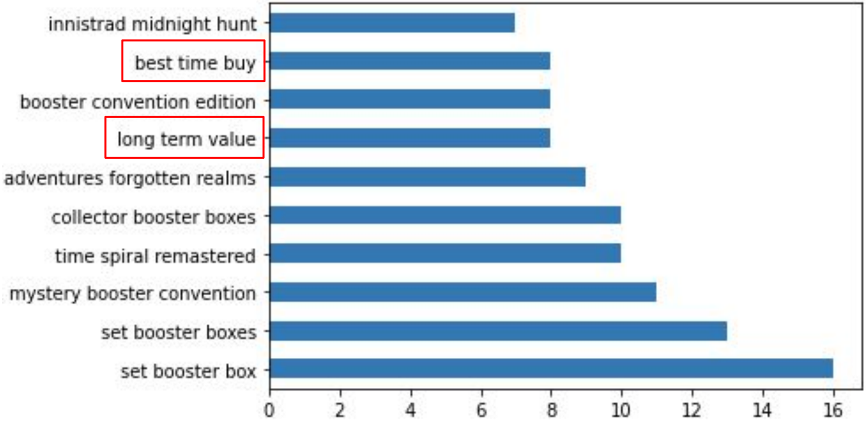
EDA



magicTCG



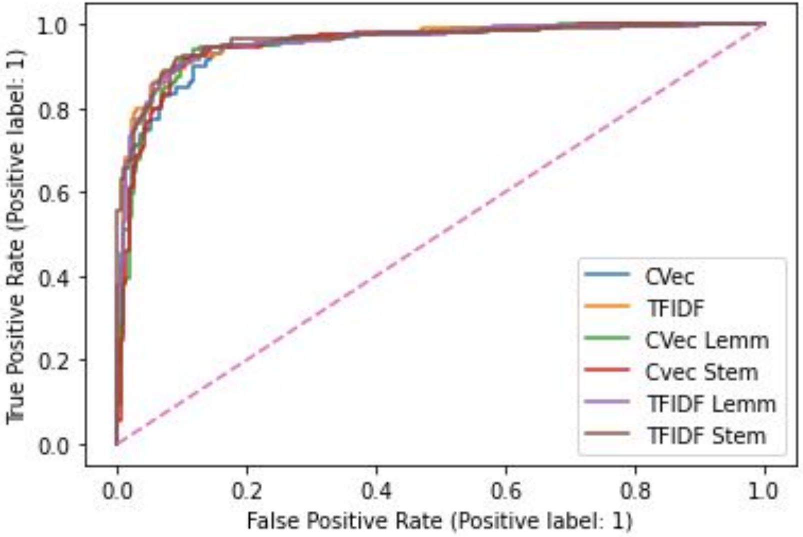
mtgfinance



Modelling - Comparing Lemmatization and Stemming

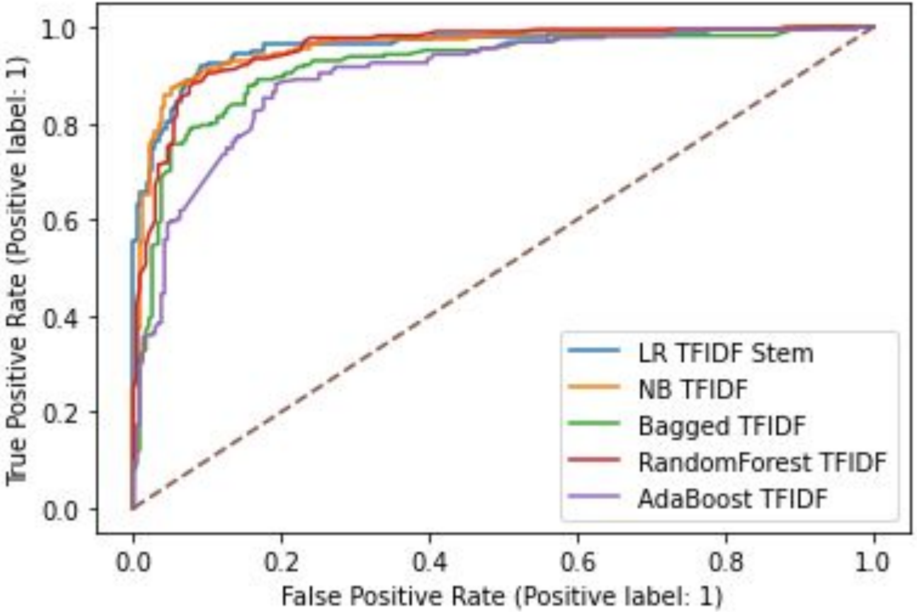
- Used logistic regression as base model
- Permutations of lemmatized, stemmed, neither and TF-IDF, CountVectorizer (CVec)

ROC Area Under Curve

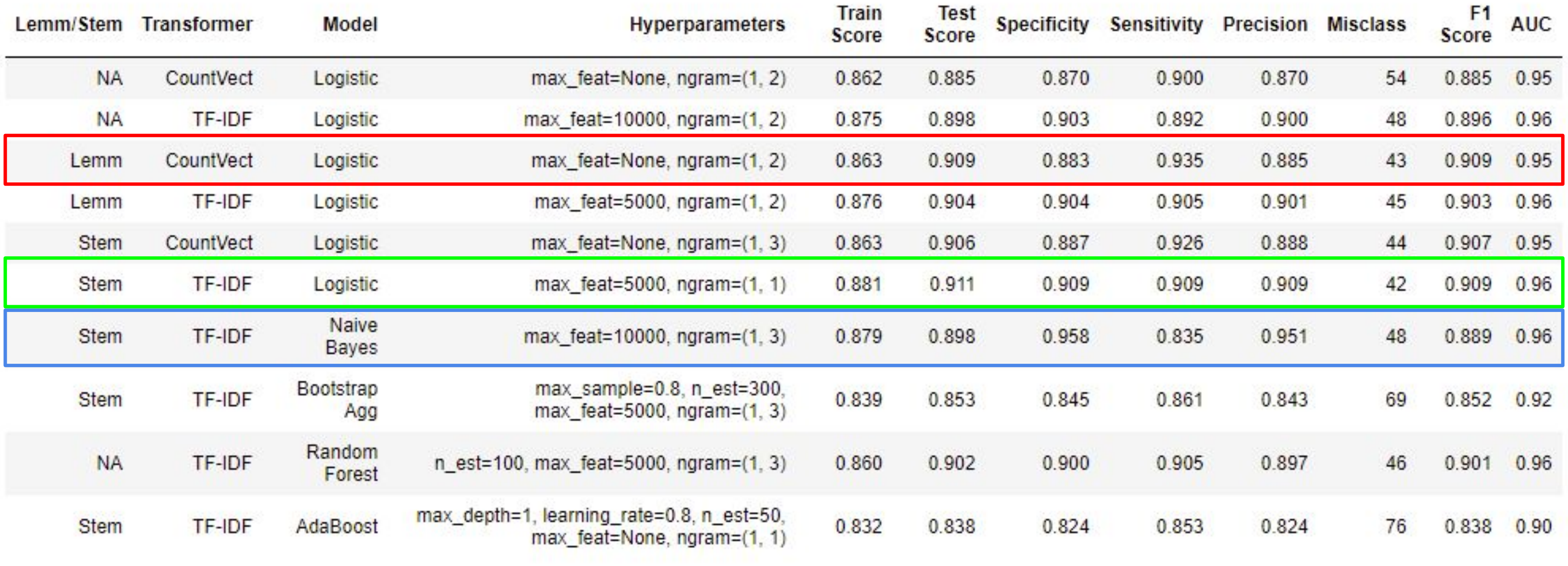


Modelling - Other Models

ROC Area Under Curve



Results and Conclusions



Lemm/Stem	Transformer	Model	Hyperparameters	Train Score	Test Score	Specificity	Sensitivity	Precision	Misclass	F1 Score	AUC
NA	CountVect	Logistic	max_feat=None, ngram=(1, 2)	0.862	0.885	0.870	0.900	0.870	54	0.885	0.95
NA	TF-IDF	Logistic	max_feat=10000, ngram=(1, 2)	0.875	0.898	0.903	0.892	0.900	48	0.896	0.96
Lemm	CountVect	Logistic	max_feat=None, ngram=(1, 2)	0.863	0.909	0.883	0.935	0.885	43	0.909	0.95
Lemm	TF-IDF	Logistic	max_feat=5000, ngram=(1, 2)	0.876	0.904	0.904	0.905	0.901	45	0.903	0.96
Stem	CountVect	Logistic	max_feat=None, ngram=(1, 3)	0.863	0.906	0.887	0.926	0.888	44	0.907	0.95
Stem	TF-IDF	Logistic	max_feat=5000, ngram=(1, 1)	0.881	0.911	0.909	0.909	0.909	42	0.909	0.96
Stem	TF-IDF	Naive Bayes	max_feat=10000, ngram=(1, 3)	0.879	0.898	0.958	0.835	0.951	48	0.889	0.96
Stem	TF-IDF	Bootstrap Agg	max_sample=0.8, n_est=300, max_feat=5000, ngram=(1, 3)	0.839	0.853	0.845	0.861	0.843	69	0.852	0.92
NA	TF-IDF	Random Forest	n_est=100, max_feat=5000, ngram=(1, 3)	0.860	0.902	0.900	0.905	0.897	46	0.901	0.96
Stem	TF-IDF	AdaBoost	max_depth=1, learning_rate=0.8, n_est=50, max_feat=None, ngram=(1, 1)	0.832	0.838	0.824	0.853	0.824	76	0.838	0.90

Limitations and Improvements

- 0.911 accuracy is not good enough for production
- The real product has to be able to perform multiclassification
- Only took into account textual data and not pictorial data