Project 3: Subreddit classification (NLP)

Maybelle Auw
13-Aug-2022

Problem statement

Train 2 NLP (Natural Language Processing) classifiers using 2 subreddits (**Marvel** & **DC comics**), webscraped from Reddit.com via Pushshift.API.

One classifier must be Random Forest.

Goal: To classify between these 2 subreddits a given post came from.

Data

Subreddits

1. r/Marvel - 1008 posts
2. r/DCcomics - 1010 posts

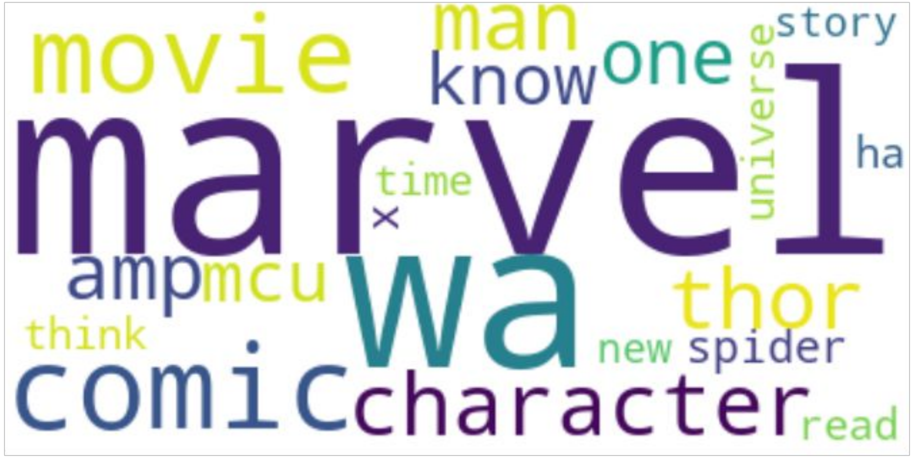
Webscrape method

- PushshiftAPI
- Filtered blank / deleted / removed posts

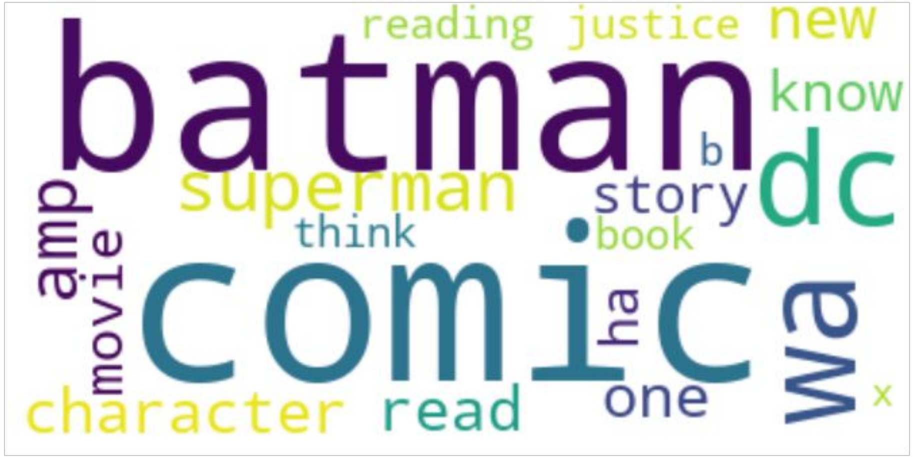
Feature engineering

- full_post = selftext + ‘ ’ + title

Marvel wordcloud top 20 from full_post



DCcomics wordcloud top 20 from full_post



Preprocessing

- **One-hot encode** target variable **(Marvel==1, DC comics==0)**
- **Train-test-split**
  - 0.2 test size
  - stratify by target variable
- **Tokenize for lemmatization**
  - NLTK module
  - RegexpTokenizer('[a-z]+', gaps=False) to capture alphabet characters / words only
  - WordNetLemmatizer() + stopwords removal
  - Rejoin words for pipeline

Pipelines

1. CountVectorizer + Random Forest Classifier
2. TfidfVectorizer + Random Forest Classifier
3. CountVectorizer + Multinomial Naive Bayes
4. TfidfVectorizer + Multinomial Naive Bayes

Tuning hyperparameters

Word Vectorizers (Count, Tfidf)

- Features: 1000
- Min_df: 3
- Max_df: 0.6
- Ngrams: (1,2)
- Stop_words: english (redundant)
- Accent: unicode (redundant)

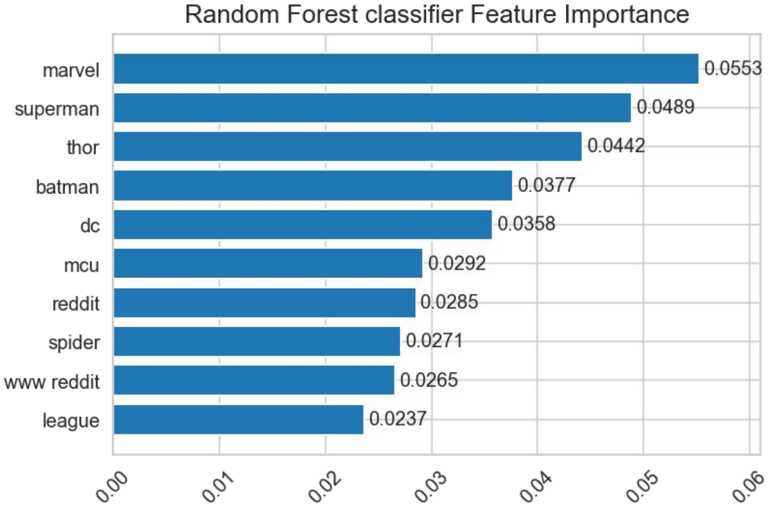
Random Forest Classifier

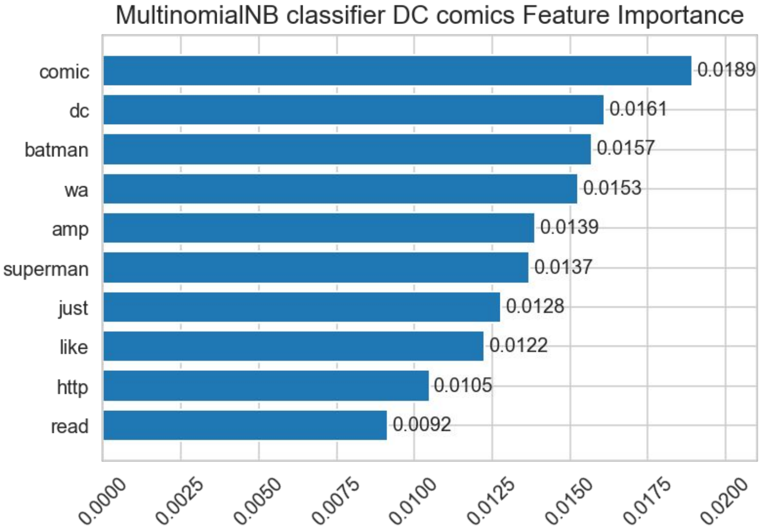
- Max_depth: 3
- Min_samples_split: 2
- N_estimators: 100
- N_jobs: -2
- Random_state: 3

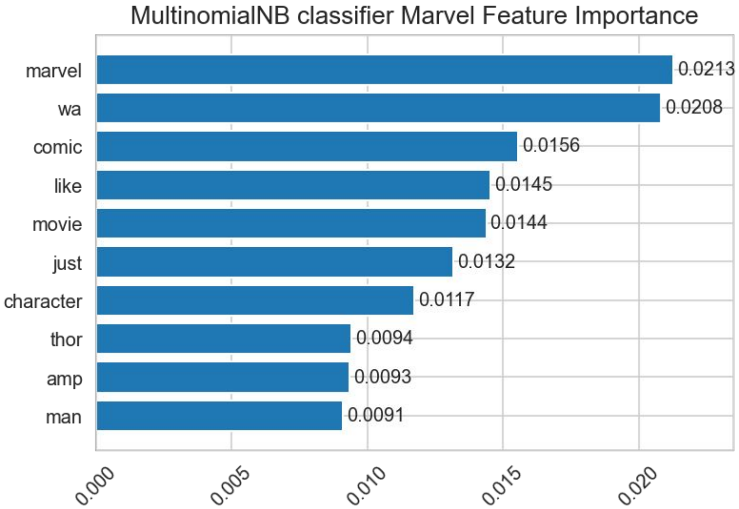
Multinomial Naive Bayes

- Alpha: 0.5

Feature importances

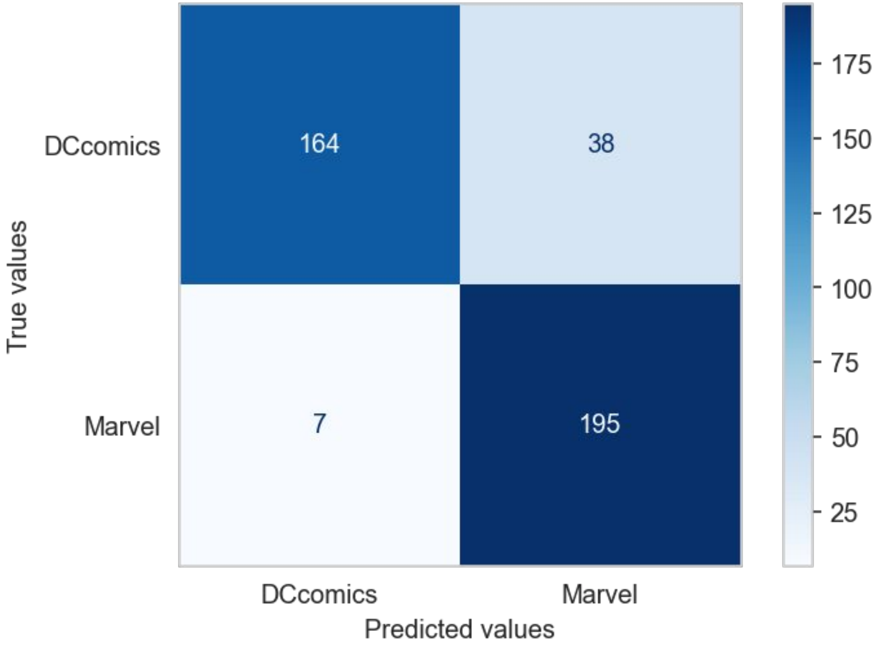

Random Forest classifier Feature Importance


MultinomialNB classifier DC comics Feature Importance

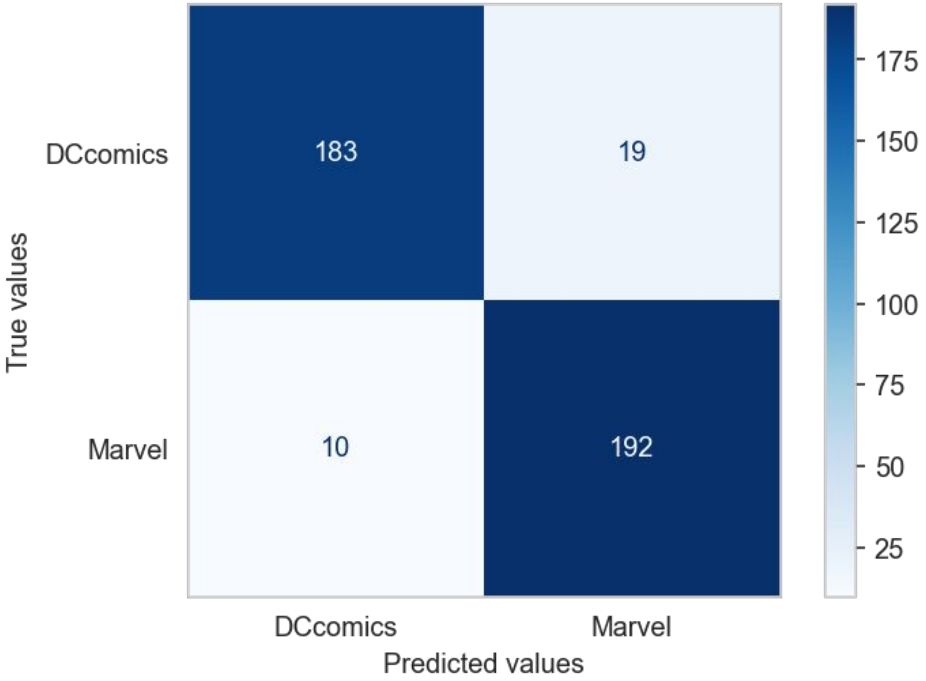

MultinomialNB classifier Marvel Feature Importance

Confusion matrix

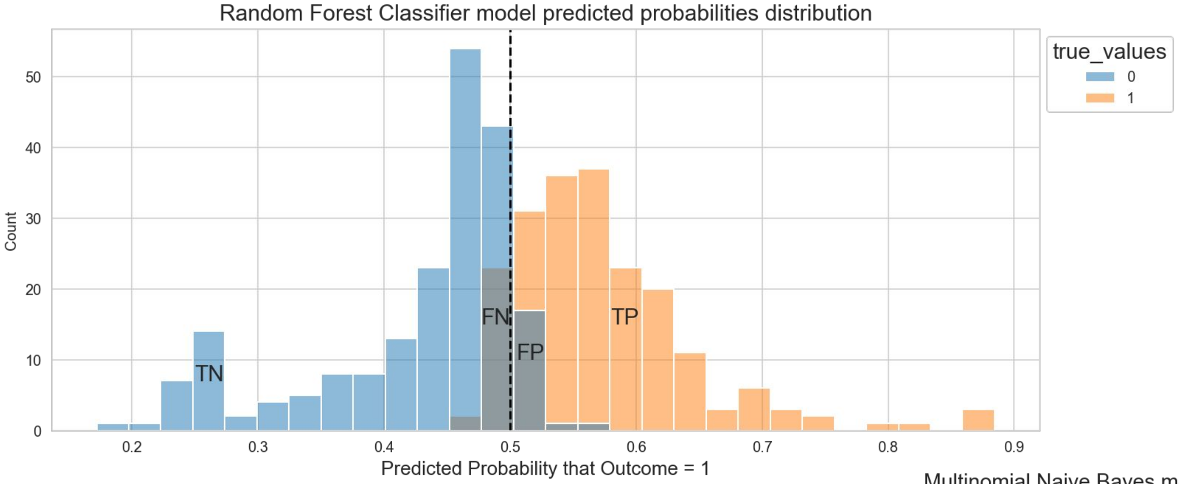
Random Forest classifier confusion matrix

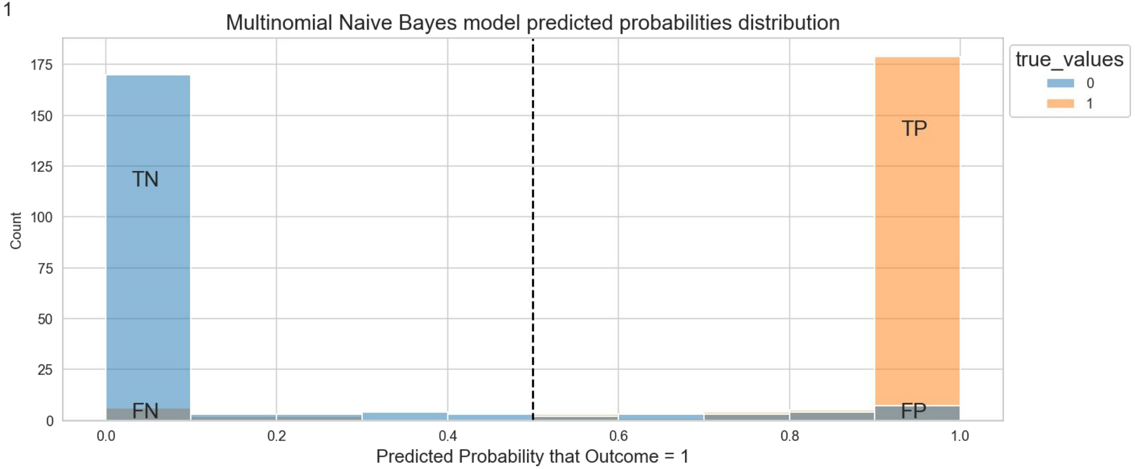


Multinomial **Naive Bayes** confusion matrix

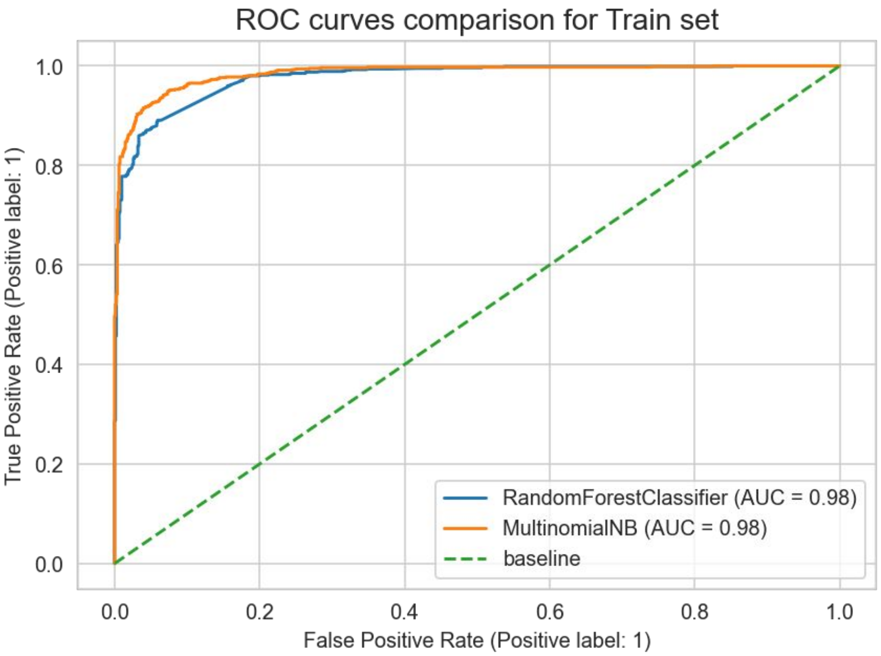


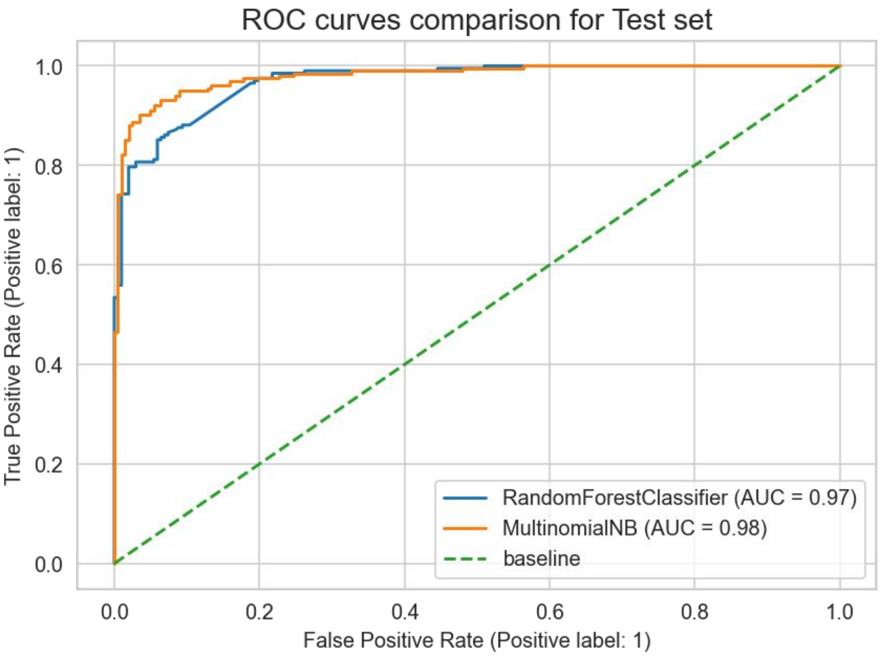
Probabilities distribution



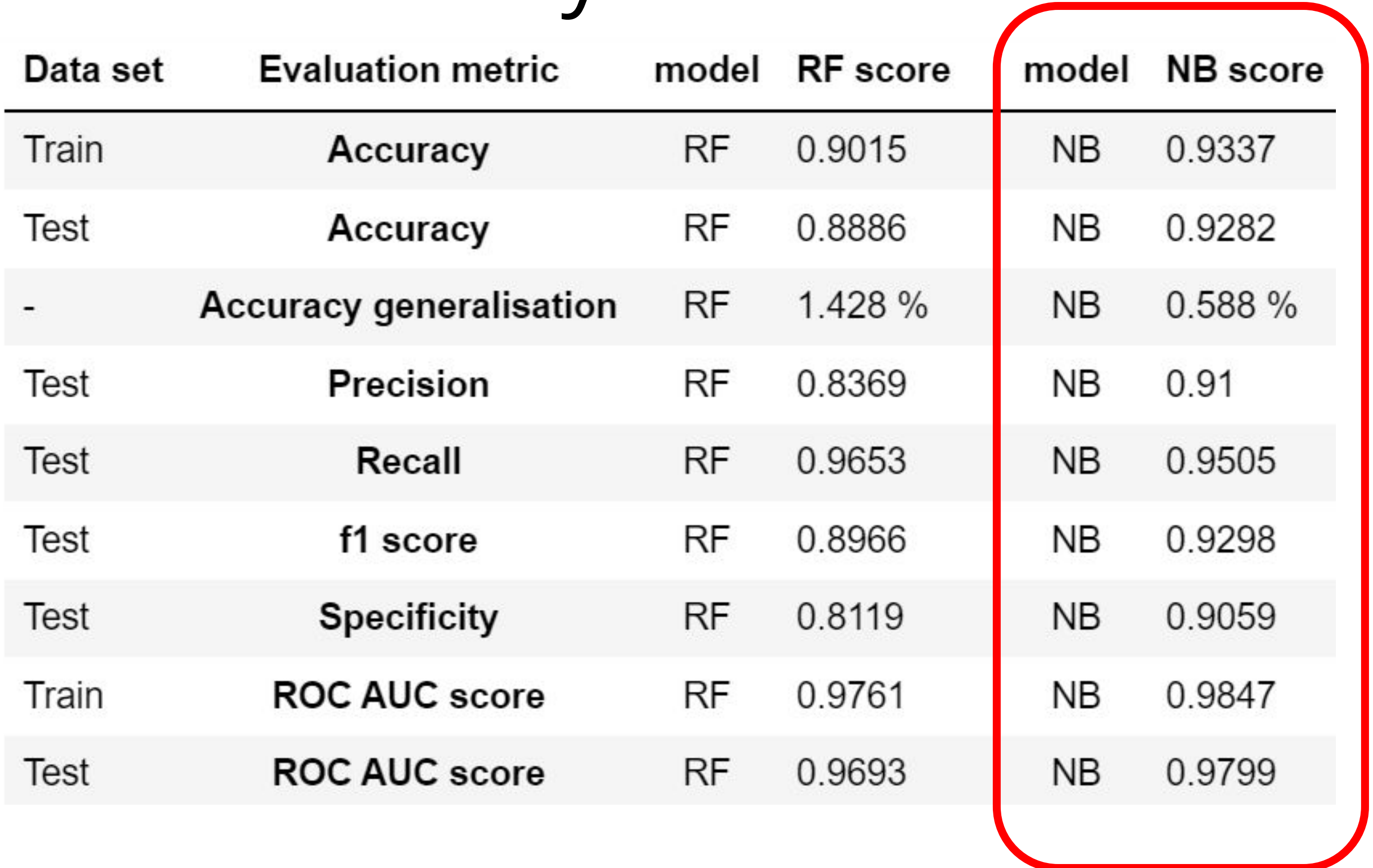


ROC curves


ROC curves comparison for Train set
True Positive Rate (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0
RandomForestClassifier (AUC = 0.98)
MultinomialNB (AUC = 0.98)
baseline


ROC curves comparison for Test set
True Positive Rate (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0
RandomForestClassifier (AUC = 0.97)
MultinomialNB (AUC = 0.98)
baseline

Evaluation summary



Data set	Evaluation metric	model	RF score	model	NB score
Train	**Accuracy**	RF	0.9015	NB	0.9337
Test	**Accuracy**	RF	0.8886	NB	0.9282
-	**Accuracy generalisation**	RF	1.428 %	NB	0.588 %
Test	**Precision**	RF	0.8369	NB	0.91
Test	**Recall**	RF	0.9653	NB	0.9505
Test	**f1 score**	RF	0.8966	NB	0.9298
Test	**Specificity**	RF	0.8119	NB	0.9059
Train	**ROC AUC score**	RF	0.9761	NB	0.9847
Test	**ROC AUC score**	RF	0.9693	NB	0.9799

Conclusion

- Based on ROC AUC score, CountVectorizer with Multinomial Naive Bayes is my chosen model for deployment for this particular use case and subreddit pair
- Random Forest Classifier model performance is sensitive to threshold change, while Multinomial Naive Bayes model is more resilient

Limitations

- Model is limited to classifying between Marvel and DC comics subreddits
- New posts with only out-of-vocabulary words may be wrongly classified

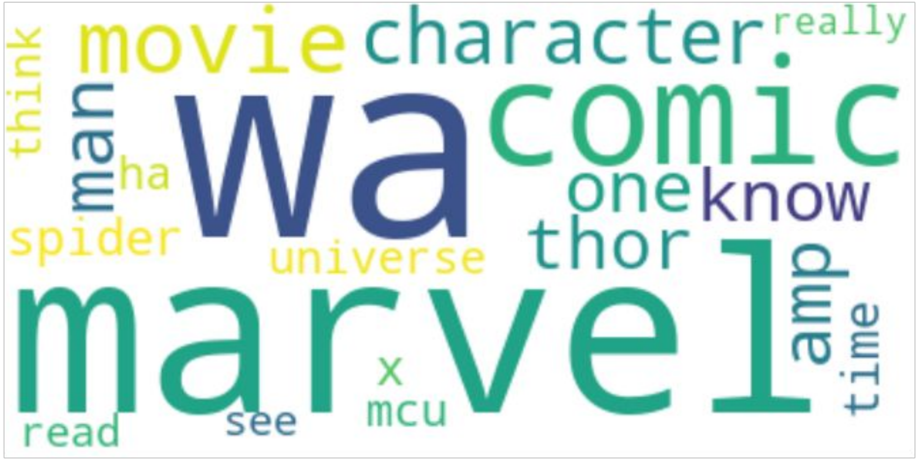
Recommendations

- Deep dive into posts that are wrongly classified, and use the findings to improve the model
- Try word embeddings for text processing e.g. Word2Vec, GLoVe, ELMo, BERT
- Try other models e.g. Voting, Stacking, XGBoost, AdaBoost, Gradient Boosting, CatBoost

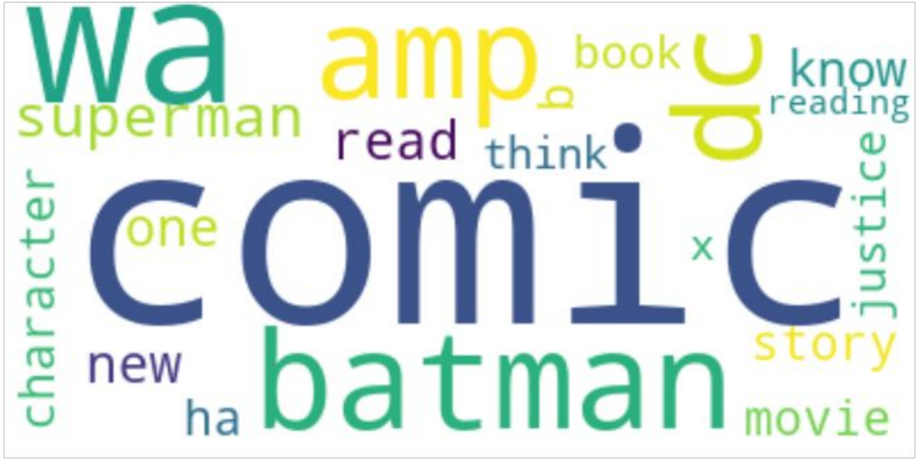
EDA: selftext (post)

Hidden

Marvel wordcloud top 20 from post

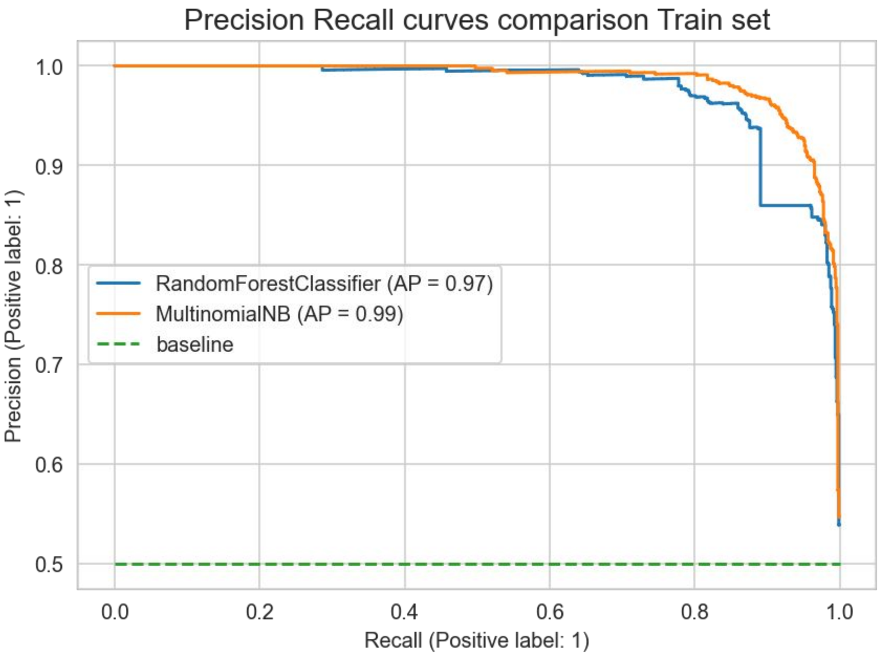


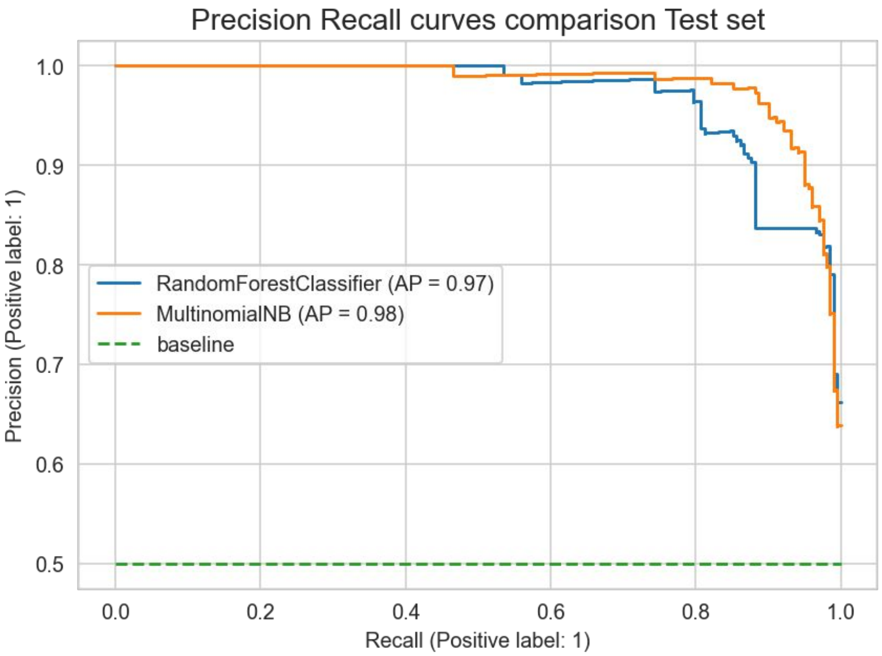
DCcomics wordcloud top 20 from post



Precision recall curves

Hidden

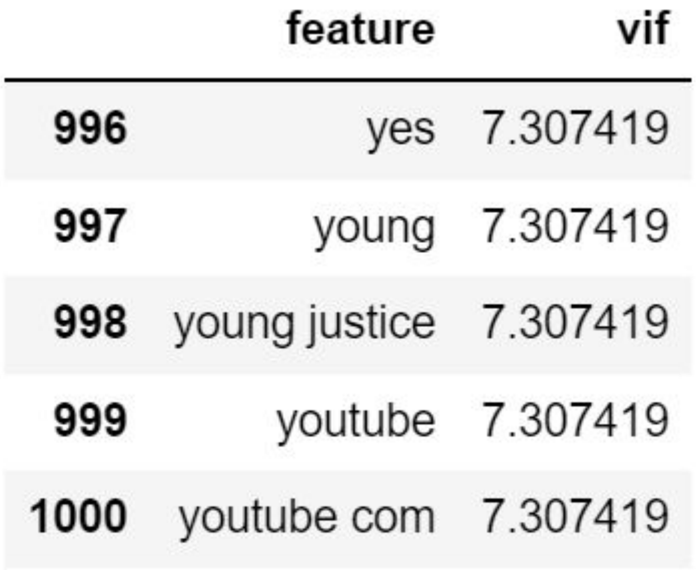




Hidden

Multicollinearity check

- Not useful exercise
- Unable to find other methods online for NLP problem



	feature	vif
996	yes	7.307419
997	young	7.307419
998	young justice	7.307419
999	youtube	7.307419
1000	youtube com	7.307419