DSI Project 3

Which subreddit to make a given post in?

Presented by Yong Fah Aik

Table of Contents

1. Problem Statement
2. Methodology
3. Key Findings
4. Conclusion

Background

- Reddit is a network of communities where people can dive into their interests, hobbies and passions. Subreddits are user-created channels where discussion on the topic of interest, hobby or passion are organized.
- From Metrics For Reddit, there are over 3.2 million subreddits as of December 2021, with hundreds of subreddits being created every day.
- As there are many different subreddits on Reddit, and since interests, hobbies and passions can be similar, there are always various subreddits that are similar to each other. Without a doubt, anyone who is new to writing and posting to Reddit can be confused as to which subreddit to post to.

Problem Statement

In this project, the aim is to assist the new Reddit user in the decision of which subreddit to make the post in, through classification models based on the analysis of the posts from two subreddits:

- Success of the model are to be based on:
  - Accuracy (How accurate in determining which subreddit a post comes from)
  - Specificity (The proportion of the posts not from the target subreddit being predicted correctly)
- The identification of key root word features of the subreddits

Methodology: Web Scraping and Data Cleaning

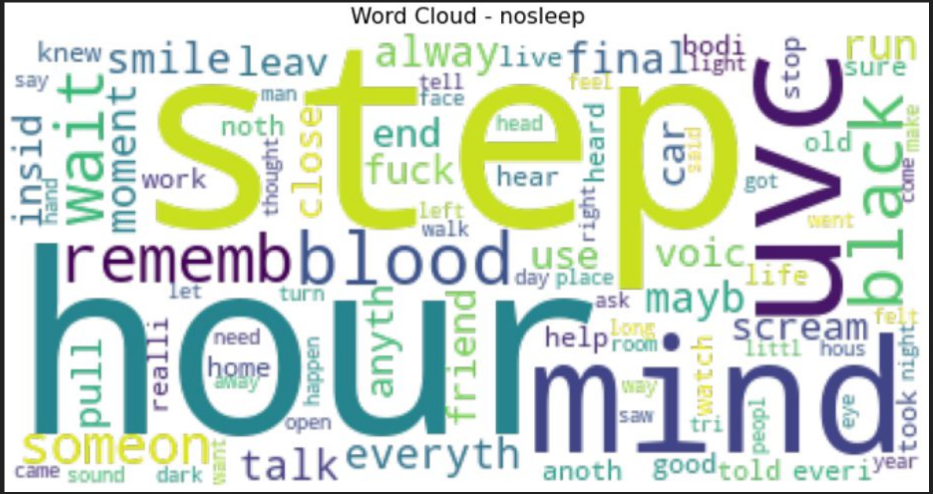
- The chosen subreddits are:
  - r/nosleep
  - r/Paranormal
- 1,000 posts from each subreddit are then scraped using the PushShift's API.
- Text for analysis to be generated from title & selftext of posts
- Posts with missing selftext, removed posts, duplicated posts, posts with duplicate title in selftext are dealt with accordingly.

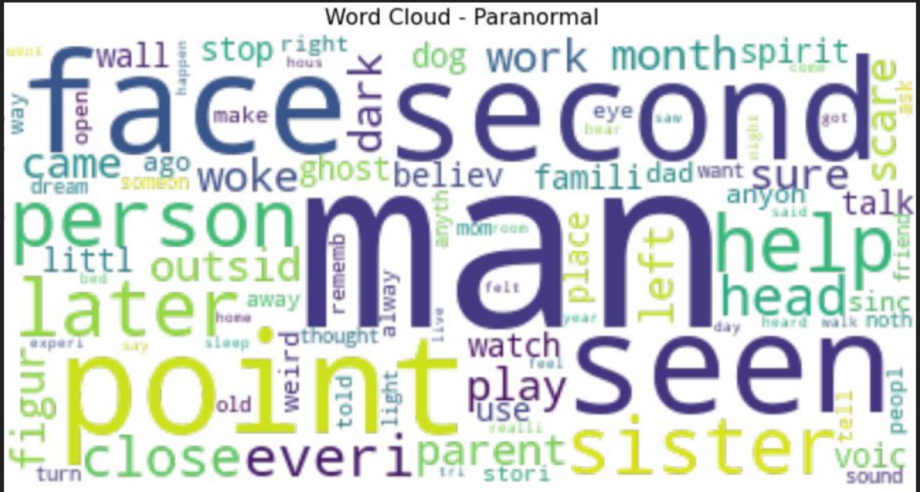
Methodology: Text Preprocessing

Steps of text preprocessing:

- Remove Special Characters
  - Include removing urls and syntax words for spacing etc.
- Tokenizing
  - Turning the text string into smaller word tokens
- Lemmatizing/Stemming
  - To normalize the word tokens for easier analysis
- Stop Word Removal
  - Removal of commonly used words that provide little information for analysis, as well as removal of words that may result in bias of the model

Methodology: Top Occurring Words

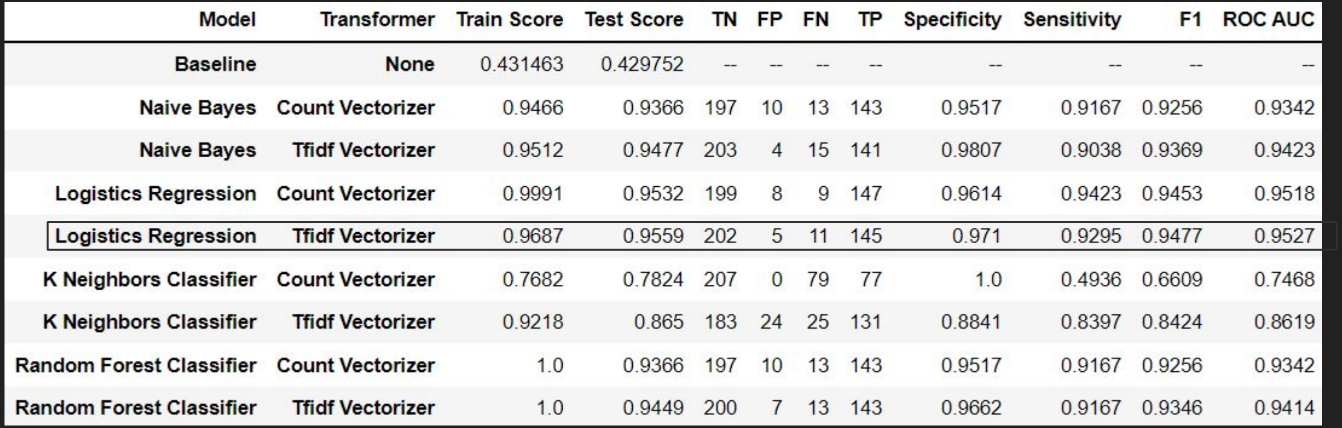




Word Clouds of the top occurring word features of both subreddits:
It should be noted that the words are generated after stemming, and as such may seem incomplete.

Key Findings (Model)

- Baseline made determining every post to be of the target subreddit.
- Models based on the combination of 2 transformers (Count Vectorizer & Tfidf Vectorizer) and 4 estimators (Naive Bayes Multinomial, Logistics Regression, K Neighbors Classifier & Random Forest Classifier)
- Evaluations are made using statistical metrics to show how well the model does.



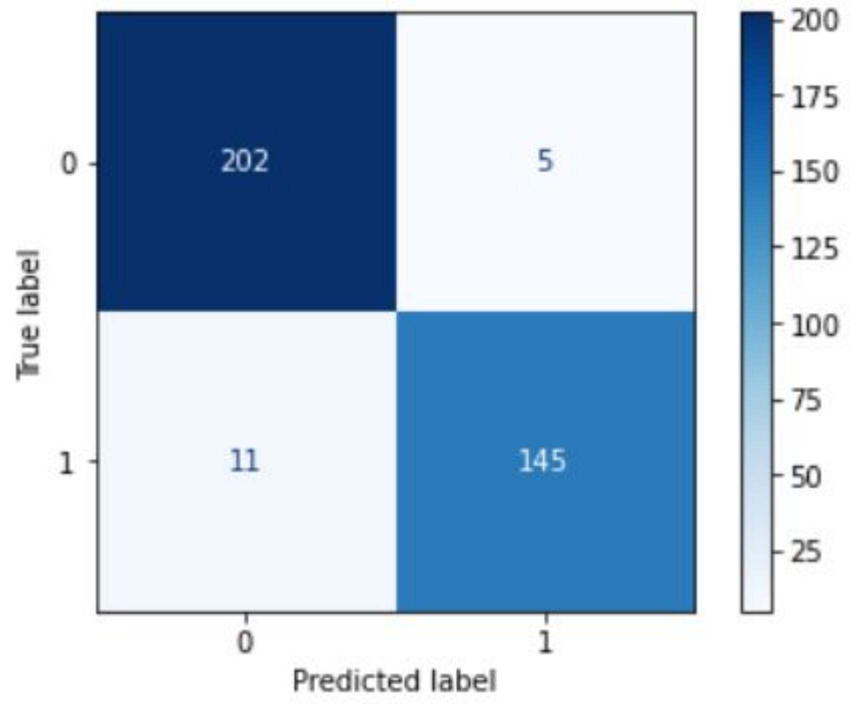
Model	Transformer	Train Score	Test Score	TN	FP	FN	TP	Specificity	Sensitivity	F1	ROC AUC
Baseline	**None**	0.431463	0.429752	--	--	--	--	--	--	--	--
Naive Bayes	**Count Vectorizer**	0.9466	0.9366	197	10	13	143	0.9517	0.9167	0.9256	0.9342
Naive Bayes	**Tfidf Vectorizer**	0.9512	0.9477	203	4	15	141	0.9807	0.9038	0.9369	0.9423
Logistics Regression	**Count Vectorizer**	0.9991	0.9532	199	8	9	147	0.9614	0.9423	0.9453	0.9518
Logistics Regression	**Tfidf Vectorizer**	0.9687	0.9559	202	5	11	145	0.971	0.9295	0.9477	0.9527
K Neighbors Classifier	**Count Vectorizer**	0.7682	0.7824	207	0	79	77	1.0	0.4936	0.6609	0.7468
K Neighbors Classifier	**Tfidf Vectorizer**	0.9218	0.865	183	24	25	131	0.8841	0.8397	0.8424	0.8619
Random Forest Classifier	**Count Vectorizer**	1.0	0.9366	197	10	13	143	0.9517	0.9167	0.9256	0.9342
Random Forest Classifier	**Tfidf Vectorizer**	1.0	0.9449	200	7	13	143	0.9662	0.9167	0.9346	0.9414

Key Findings (Final Model)

The Final Model is the Logistics Regression Model with Tf-idf Vectorizer.

GridSearchCV Hyperparameters:

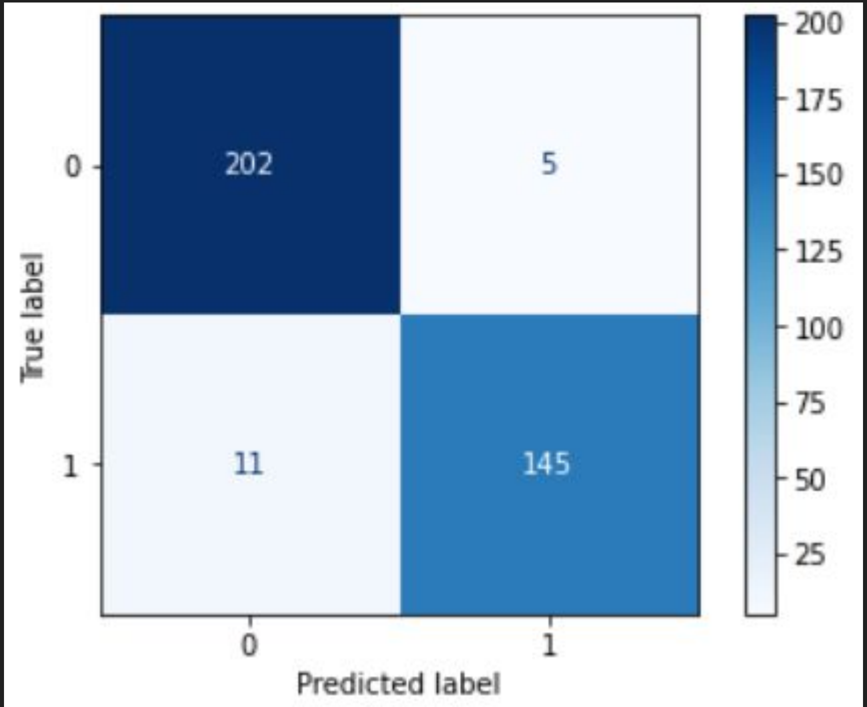
- ‘tvec__max_df': 0.7
- 'tvec__max_features': 2000
- 'tvec__min_df': 1
- 'tvec__ngram_range': (1, 1)



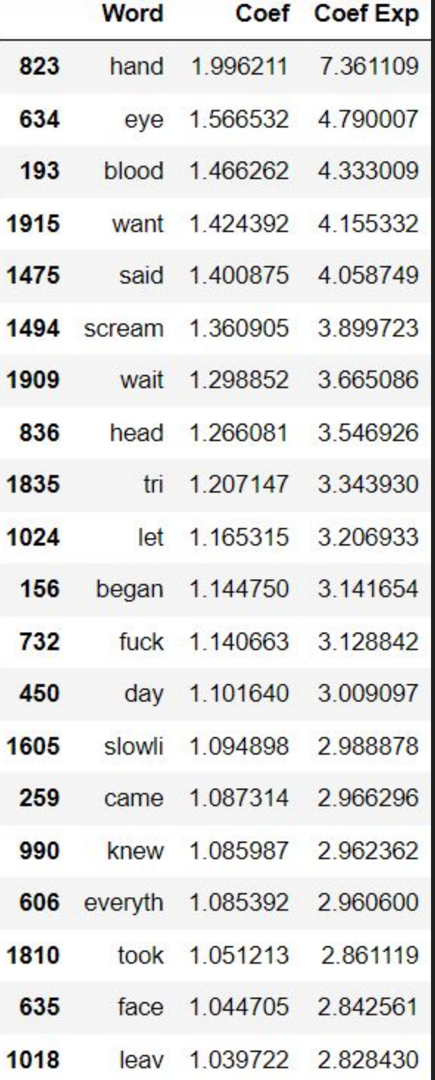
Key Findings (Final Model)

The Model Scores are as follows:

- Test Score: 95.59%
- Specificity: 97.1%
- Sensitivity: 92.95%
- F1 Score: 94.77%
- ROC/AUC: 0.9527
- Misclassification: 4.41%



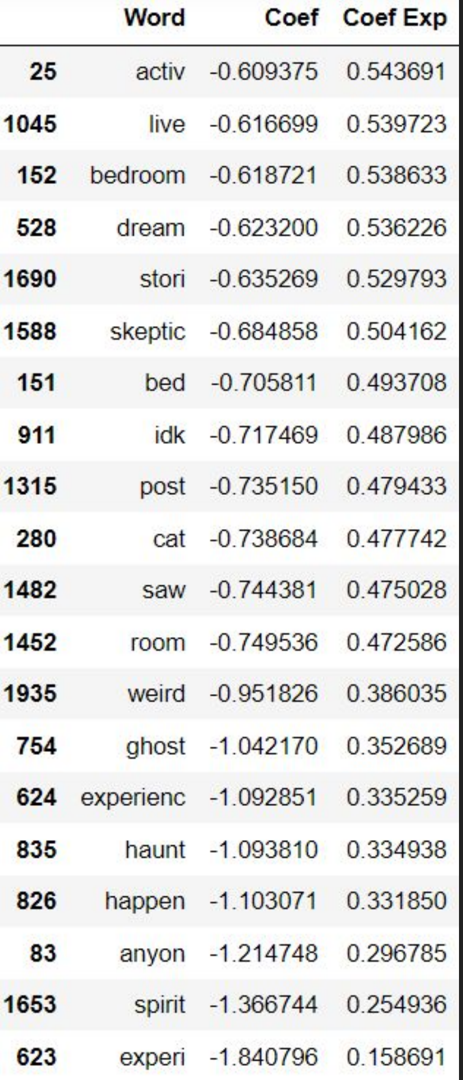
Key Findings (Final Model)



	Word	Coef	Coef Exp
823	hand	1.996211	7.361109
634	eye	1.566532	4.790007
193	blood	1.466262	4.333009
1915	want	1.424392	4.155332
1475	said	1.400875	4.058749
1494	scream	1.360905	3.899723
1909	wait	1.298852	3.665086
836	head	1.266081	3.546926
1835	tri	1.207147	3.343930
1024	let	1.165315	3.206933
156	began	1.144750	3.141654
732	fuck	1.140663	3.128842
450	day	1.101640	3.009097
1605	slowli	1.094898	2.988878
259	came	1.087314	2.966296
990	knew	1.085987	2.962362
606	everyth	1.085392	2.960600
1810	took	1.051213	2.861119
635	face	1.044705	2.842561
1018	leav	1.039722	2.828430

Top 20 root word features by r/nosleep:

- In order: hand, eye, blood, want, said, scream, wait, head, try, let, began, fuck, day, came, everything, knew, slowly, took, face, leave



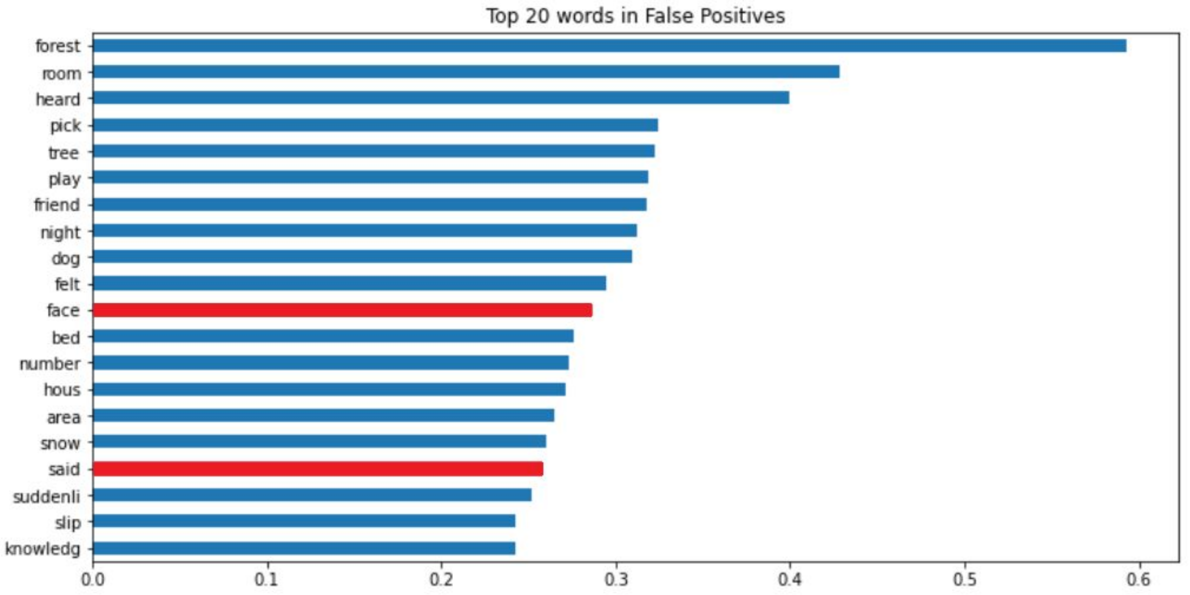
	Word	Coef	Coef Exp
25	activ	-0.609375	0.543691
1045	live	-0.616699	0.539723
152	bedroom	-0.618721	0.538633
528	dream	-0.623200	0.536226
1690	stori	-0.635269	0.529793
1588	skeptic	-0.684858	0.504162
151	bed	-0.705811	0.493708
911	idk	-0.717469	0.487986
1315	post	-0.735150	0.479433
280	cat	-0.738684	0.477742
1482	saw	-0.744381	0.475028
1452	room	-0.749536	0.472586
1935	weird	-0.951826	0.386035
754	ghost	-1.042170	0.352689
624	experienc	-1.092851	0.335259
835	haunt	-1.093810	0.334938
826	happen	-1.103071	0.331850
83	anyon	-1.214748	0.296785
1653	spirit	-1.366744	0.254936
623	experi	-1.840796	0.158691

Key Findings (Final Model)

Top 20 root word features by r/Paranormal:

- In order: experiment, spirit, anyone, haunt, happen, experience, ghost, weird, room, saw, cat, idk, post, skeptic, bed, bedroom, story, active, dream, live

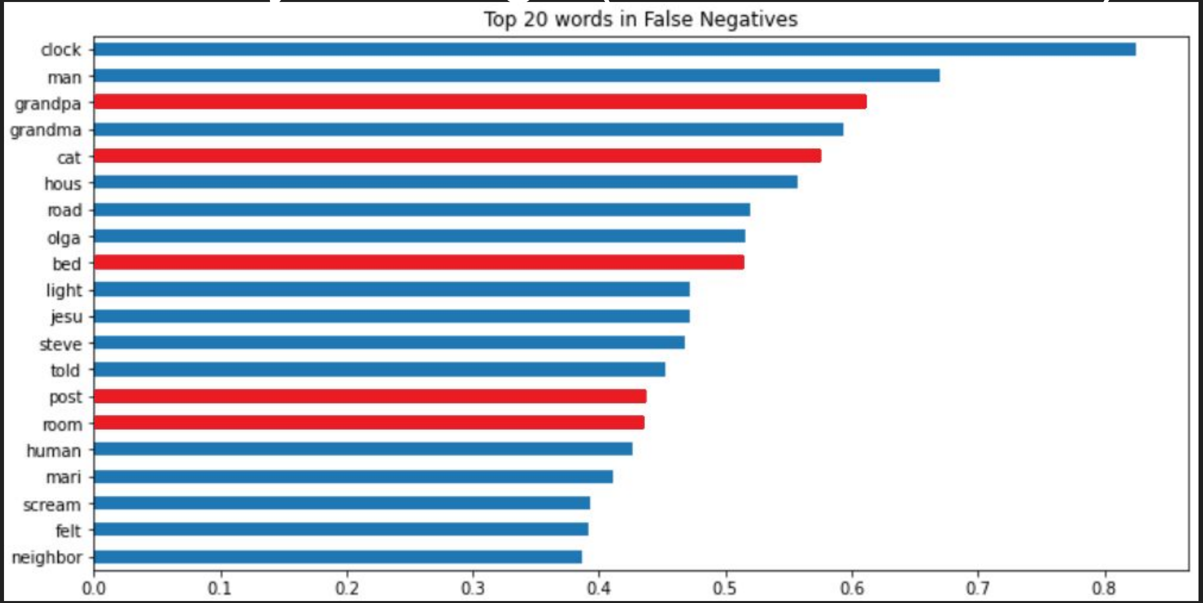
Key Findings (Misclassification)



False Positives (Type I Errors) (5 posts)

- r/Paranormal posts wrongly predicted as r/nosleep
- Words related to r/nosleep used: said, face

Key Findings (Misclassification)



False Negatives (Type II Errors) (11 posts)

- r/nosleep posts wrongly predicted as r/Paranormal
- Words related to r/Paranormal used: room, post, bed, cat, grandpa

Conclusion

- GridSearchCV only searched through the hyperparameters of the transformer, and the default model of the estimator is used. Refinement to the model can be made through the additions of hyperparameters of the estimator, though this is a trade-off between time taken to model due to the exponential increase of fittings needed.
- Text analysis is based on the text after stemming is performed. Further analysis can be counted on other forms of preprocessing of text like lemmatization or even without any preprocessing.
- Inherent issues with the choices of subreddits: Text-heavy subreddits and the topic of interest is slightly different.