Natural Language Processing

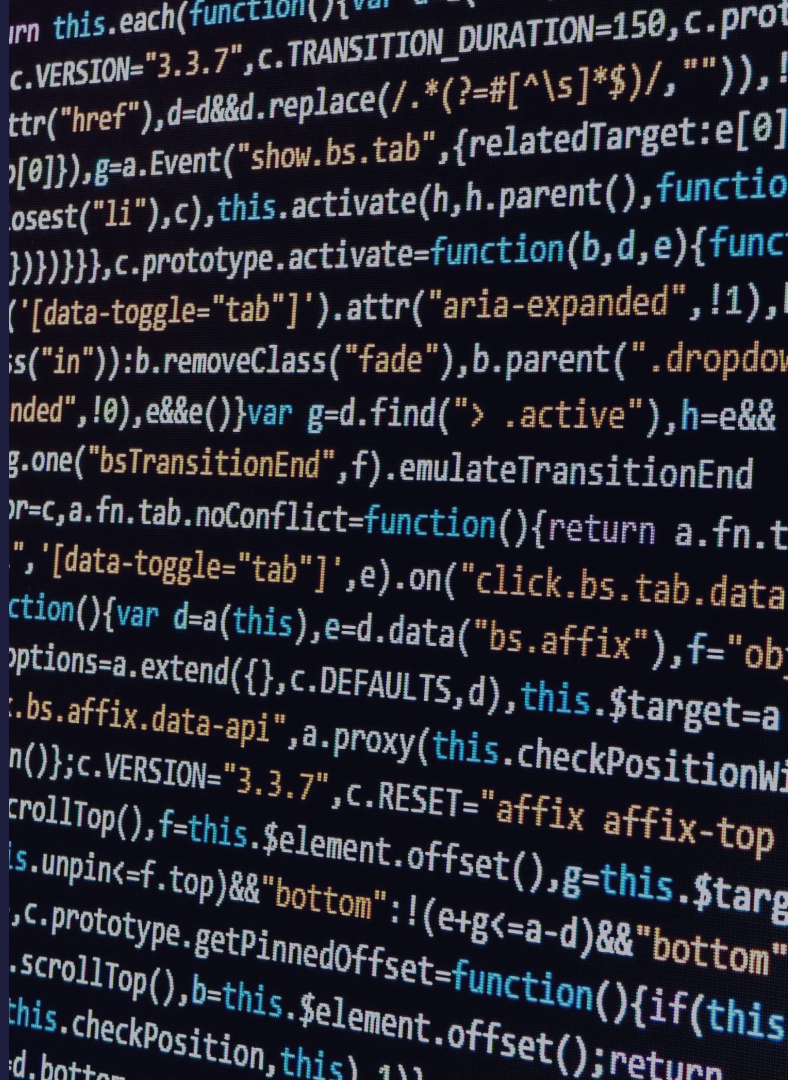


Table of Contents

01 Overview

Problem Statement

“Men are from Mars, women are from Venus”

As users from subreddit AskMen and AskWomen expresses themselves differently, how can we accurately differentiate them through the use of languages or words?

Datasets

AskMen

"The premier place to ask random strangers about the intricacies of the human condition"

01

AskWomen

"A subreddit dedicated to asking women questions about their thoughts, lives, and experiences; providing a place where all women can comfortably and candidly share their responses in a non-judgmental space."

02

02 Modeling Process

Modeling Process

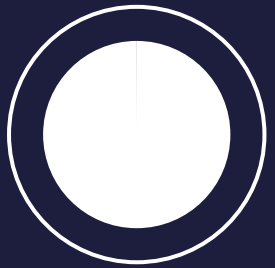
Data Collection

Data collection is done using Reddit API

925 Posts

AskMen

925 Posts



AskWomen

Pre-Processing

1292 posts

Training Set

555 posts

Testing Set

Pre-Processing

01
Binary Label

1 Askmen
0 AskWomen

02
Remove non-letter

Regular Expressions:
^a-zA-Z

04
Stop-words

- Default English stop-words
- askmen, askwomen
- did, was, ve

03
Lemmatizing

Rootform:
Part of Speech - Verbs

Modeling

Modeling was done using pipeline and gridsearch to get the best parameters

	Model	Text Vector	Train Score	Test Score
01	Logistic Regression	Count Vec	0.945	0.727
02	Logistic Regression	TF-IDF	0.900	0.758
03	Naive Bayes	TF-IDF	0.887	0.733

03 Model Inferential

Top 5 Features

Logistic Regression (Count Vec)	Askmen	men	surprised	girlfriend	girl	genuinely
	AskWomen	ladies	women	late	affect	positive
Logistic Regression (TF-IDF)	Askmen	men	guy	girl	girlfriend	advice
	AskWomen	women	ladies	partner	learn	story
Naive Bayes (TF-IDF)	Askmen	men	just	like	guy	make
	AskWomen	women	feel	partner	relationship	ladies

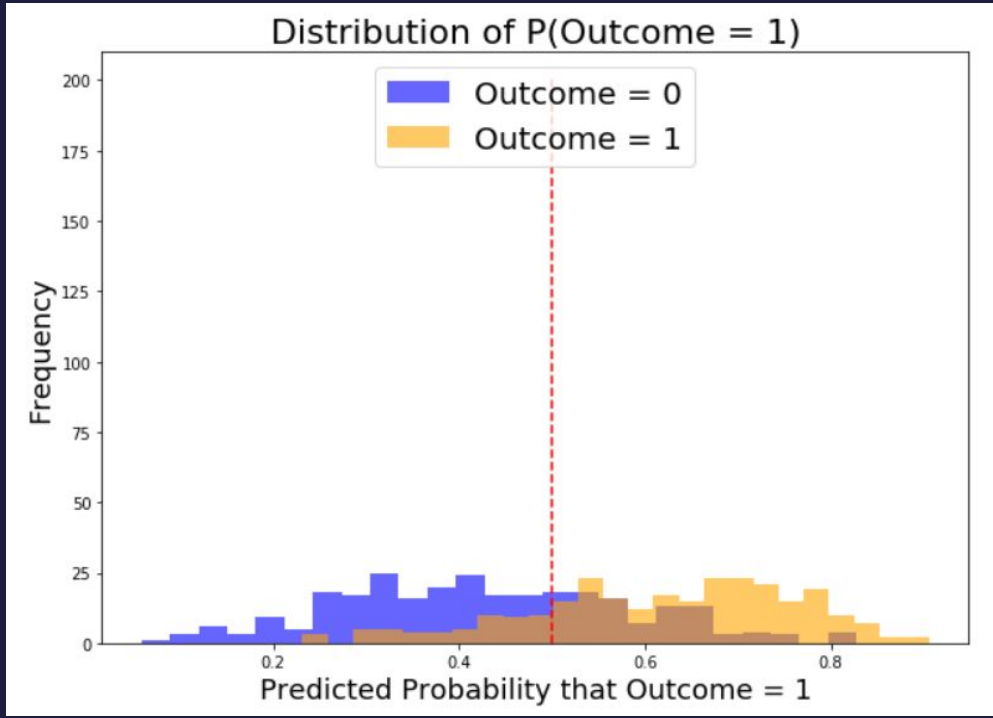
Evaluation Metric

	Sensitivity	Specificity	Precision	ROC Score	TN	FP	FN	TP
Logistic Regression (TF-IDF)	0.7818	0.7357	0.7439	0.845	206	74	60	215
Naive Bayes (TF-IDF)	0.8036	0.6643	0.7016	0.8129	186	94	54	221

Distribution of Positives

AskWomen

AskMen



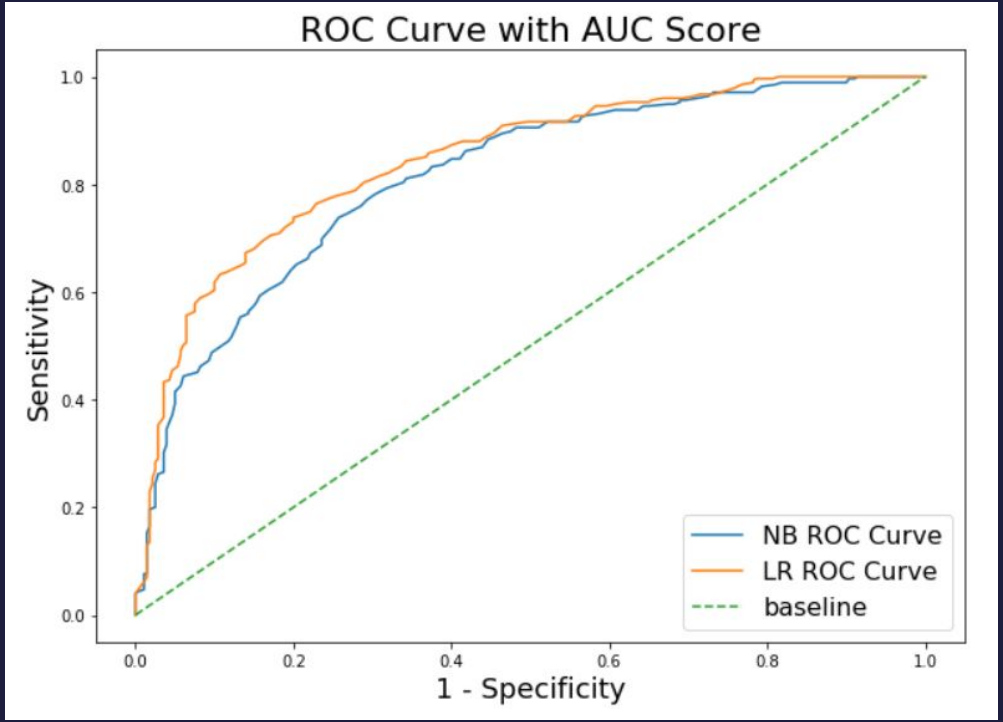
ROC AUC Score

Naive Bayes

ROC with AUC score of 0.813

Logistic Regression

ROC with AUC score of 0.845



04 Conclusion

Factors to Consider

Assumptions

Assumption that majority of the AskWomen post is contributed by the women. Likewise for AskMen's posts.

Phrases

Phrases provides more context than individual words

Recommendations

More data

Scrape more data to drive variance down

Refine Stop-Words

Add in more language stop-words to the list (e.g. helping verbs) to reduce features

Eliminate features

Words with extremely low frequency

Bibliography

1. Reddit.com. 2020. *Askwomen: Questions About Women's Thoughts, Lives, And Experiences*. [online] Available at: <https://www.reddit.com/r/AskWomen/> [Accessed 17 May 2020].

2. Reddit.com. 2020. *IT's TIME TO STOP*. [online] Available at: <https://www.reddit.com/r/AskMen/> [Accessed 17 May 2020].

End of Presentation