Subreddits sentiments on brand wars:
Whole Foods or Trader Joes?

Yadan Tang

Outline

Background

Problem Statement

Data collection and EDA

NPL and Classification model screening(4 classification models)

Voting on the two comparable best performing model

Conclusion

Whole Foods vs. Trader Joe's

Trader Joe's and Wholefoods are two popular grocery chains who boast organic and sustainable products. Trader Joe's owned by ALDI, and Whole Foods are brought by Amazon in 2017.

They are often considered as competitors, **but they are very different animals in terms of:**
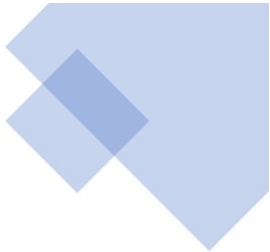- Store size
- Product mix
- Private labels

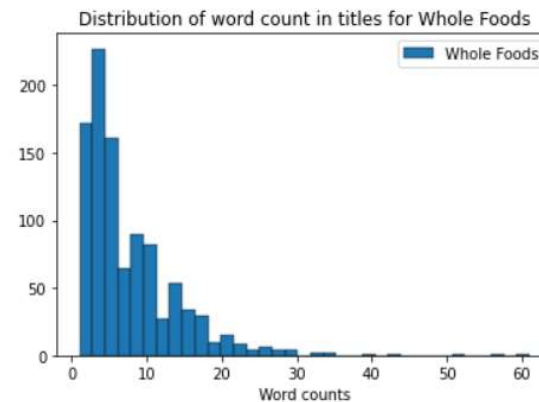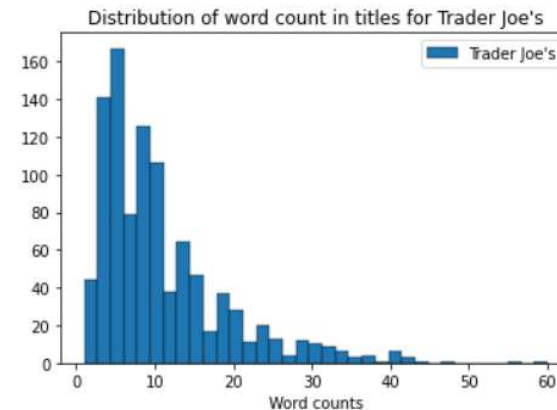**However, they have 95% overall in customers that shop both stores.**

Trader Joe's traffic recovered but Whole Foods hasn't since the onset of Pandemic

- Pandemic has a major hit on economic, both stores had witnessed a loss of in-store visits during the pandemic.

- Trader Joe's had won back traffic whereas Whole Foods still lag far behind. Despite of the advantage of win the customer by online orders and deliveries to amazon prime member, Whole Foods still cannot make up the loss in in-store visit.

- The most debate is the price. Whole Foods overly price the produce on average.

- **How is the sentiments on subreddits on the two stores? Are we able to predict the posts from subreddits for future study in customer's churn.**
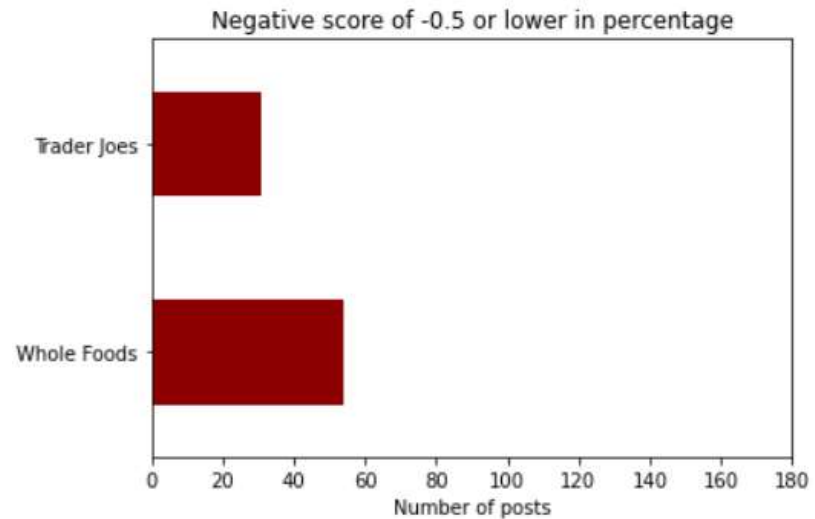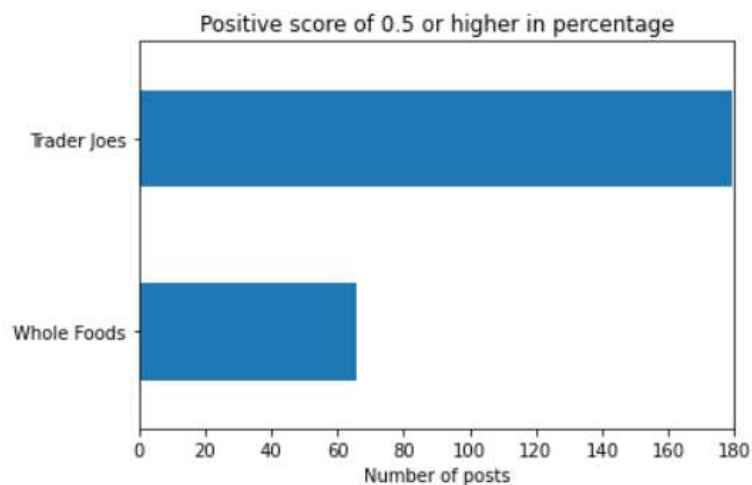
Data collection and EDA

- Data:
  - 1000 post on traderjoes and 1000 posts on wholefoods since 09/30/2021
- EDA
  - Title word counts for both stores showed very similar distribution.
    - Trader Joe's has ~90% of the posts below 20 words on both stores.
    - Whole Foods has about 95% of the posts below 20 words.

Distribution of word count in titles for Trader Joe's

Distribution of word count in titles for Whole Foods

EDA-Sentiments

In terms of the positive comments battle, Trader Joe's win over Whole Foods by a large margin:

* Trader Joe's does win over Whole Foods by a large margin in positive sentimental score
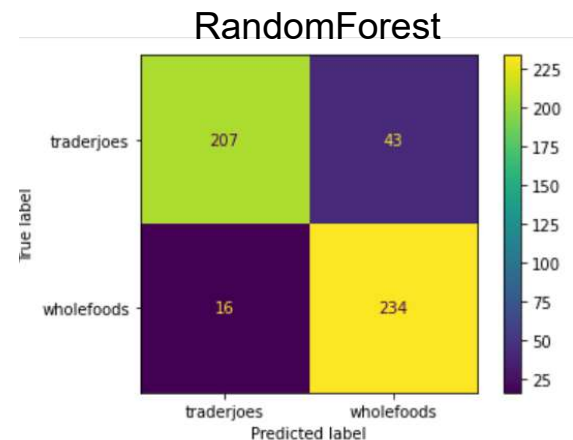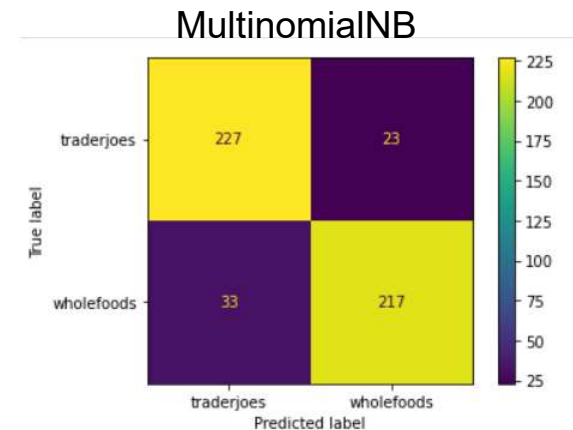* Trader Joe's has slightly fewer negative posts than Whole Foods

NPL and Classification Model

NPL: CountVectorizer()
Classification models:

Classification models	Test score	Train score
MultinomialNB	**0.89**	**0.98**
RandomForest Classifier	**0.88**	**0.99**
SVM	0.86	0.99
AdaBoostClassifier	0.84	0.96
Null model	0.5	0.5

- All models outperforms null model
- MultinomialNB and RandomForest have highest accuracy score
- MultinomialNB has a higher recall score for trader joes(if traderjoes is pos) whereas RandomForest has a higher recall score for wholefoods(if wholefood is pos)
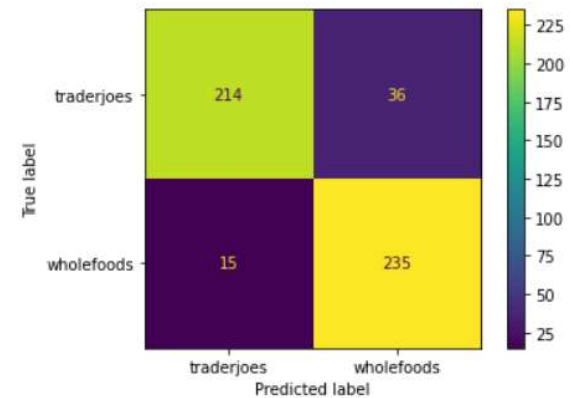

MultinomialNB

RandomForest

Final model recommendation

- VotingClassifier further improved the accuracy with a little compromise on recall.

- If the customer cares about overall prediction accuracy, voting classifier model is recommended.

- If the customer wants to have higher recall score for Trader Joes, Multinomial NB model is recommended.

	Accuracy	Precision	Recall (pos=traderjoes)
VotingClassifier	0.90	0.93	0.86
MultinomialNB	0.89	0.87	0.91
RandomForest Classifier	0.88	0.93	0.83

Votingclassifier confusion matrix

Conclusion

- The winner for the brand war on subreddit's sentiment is:

- We have established a model with 90% accuracy and 93% precision assuming correctly predicting trader joe's is the positive.

- If customer cares about a higher prediction on trader joe's posts, we will recommend an established model based on MultinomialNB.