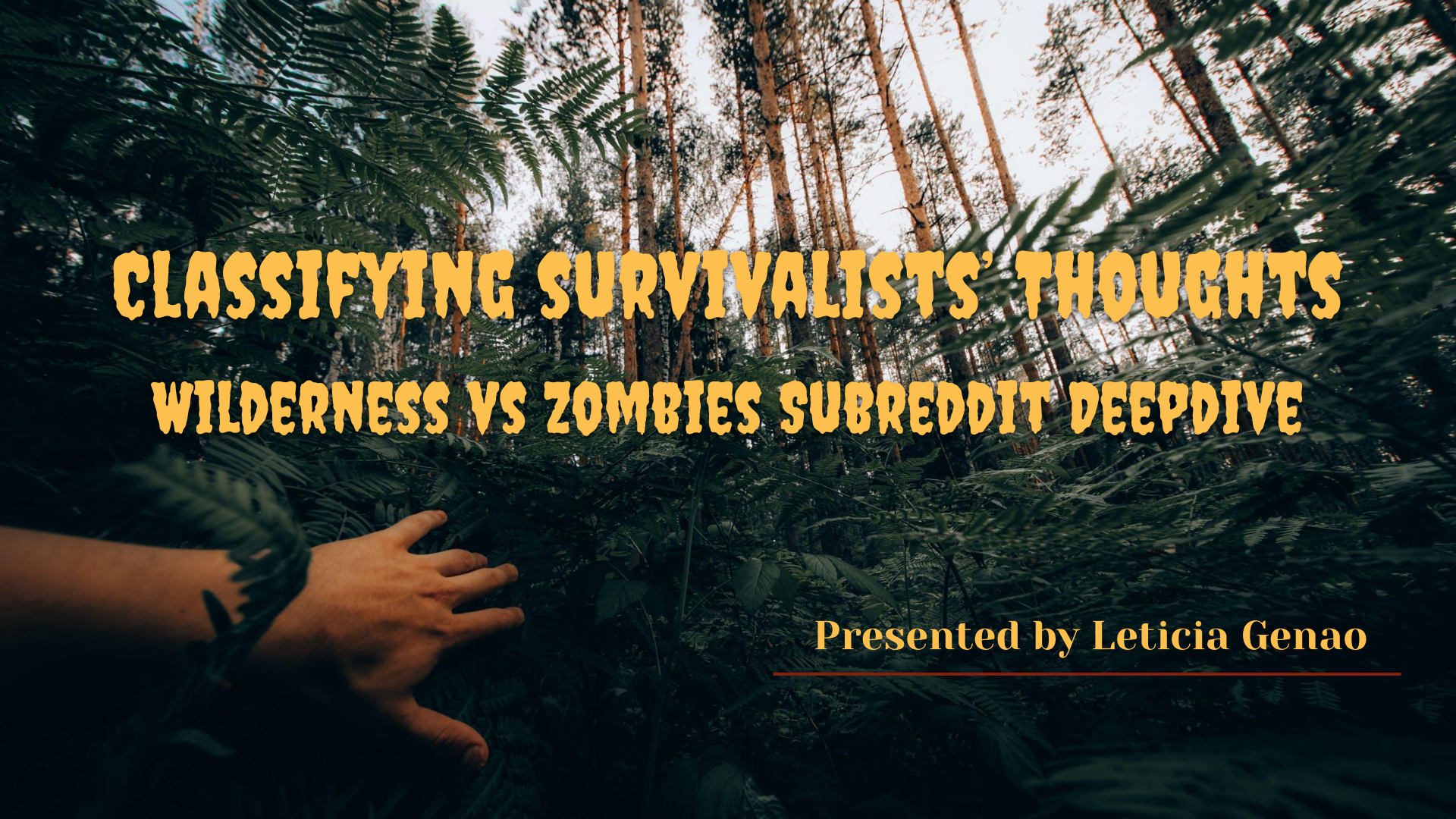


CLASSIFYING SURVIVALISTS' THOUGHTS

WILDERNESS VS ZOMBIES SUBREDDIT DEEPDIVE

Presented by Leticia Genao

PROBLEM STATEMENT

A popular forum social media, Reddit is a site where people can share posts, creates polls, and vote on good or bad posts by up and down voting them. Centered around sharing opinions, seeking advice, and engaging in discourse in specified subreddits on topics Reddit is a useful site to learn about people's thoughts regarding a news, products, shows, and other topics.

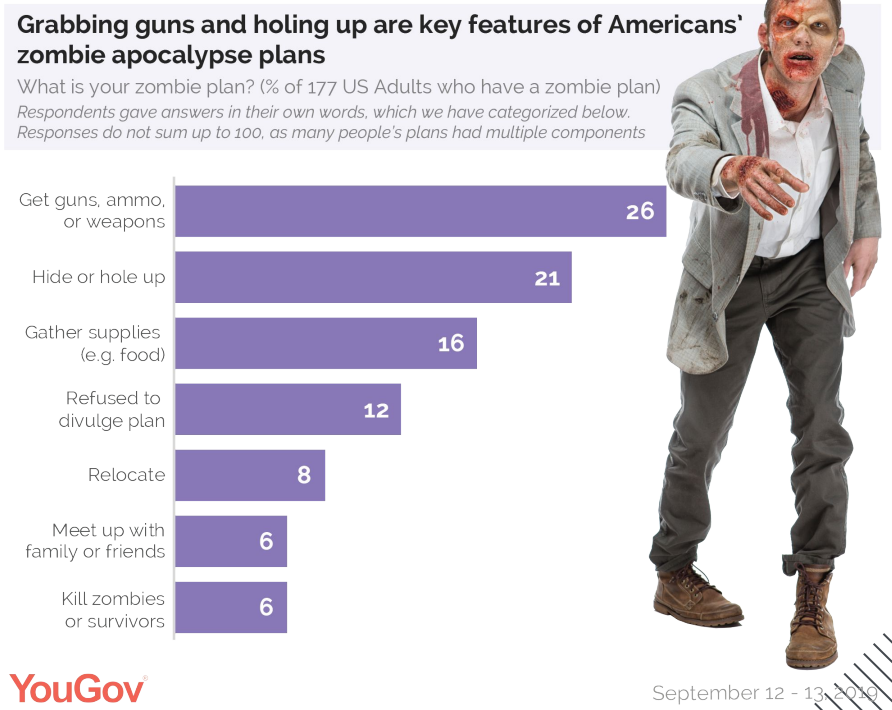
This project aims to develop classification models that can correctly specify which of two subreddits a post originated from based on the title and description of the post. The subreddit of choice are "Survival" which provides wilderness survival tips, and “Zombie Survival Tactics” which provide advice if there were ever a zombie apocalypse. This project aims to discover useful insights for the movie and television production market to see what survivalists and preppers discuss or desire.

BRIEF OVERVIEW OF DATA

Description	Reddit posts pulled from API and ran through NLP binary classification models
Data details	4800 rows (50/50 split from each subreddit). 10 total columns with all_text (title and descriptions) being the primary independent feature used.
Survival Reddit	A community of 1,645,218 dedicated to wilderness survival and prep
Zombie Survival Tactics	A community of 22,004 dedicated to zombie apocalypse survival

THE MARKET

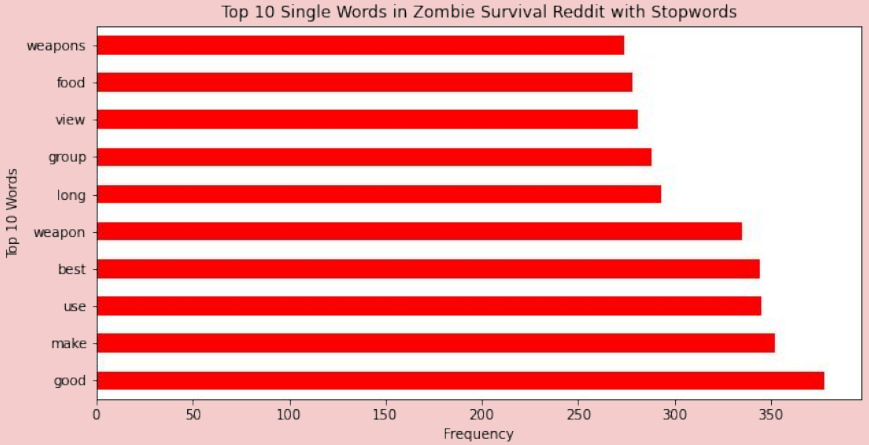
- 3.7 million Americans classify themselves as survivalists
- $107 billion dollar industry
- 14% of Americans have a zombie survival plan
- 24% of survivalists are millennials
- 15% are Generation X
- 6% are Boomers



VISUALIZATIONS - TOP 10 SINGLE WORDS

Zombie Survival Tactic

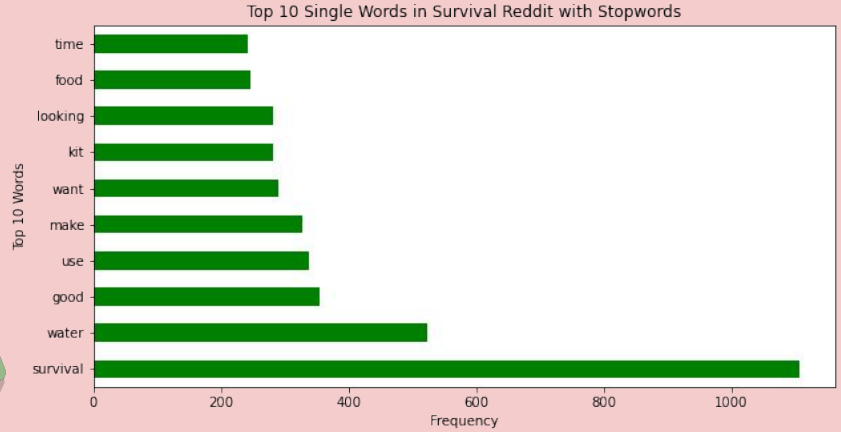
"Make", "use", "weapon", and "best" seem to indicate those in the ZST use language revolving inquiring about the best options available or to make to survive an outbreak.



Survival

In contrast to the ZST, the survival reddit seems to be centered around usefulness regarding words as "good", "kit", and potentially even "time" depending on the context.

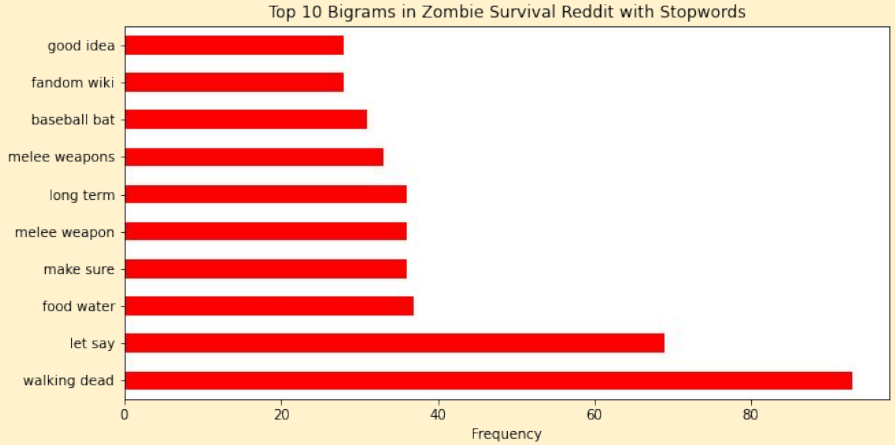
"Water" seems like the biggest concern for this group whereas ZST had "weapon" being the most insightful word.



VISUALIZATIONS - TOP 10 BIGRAMS

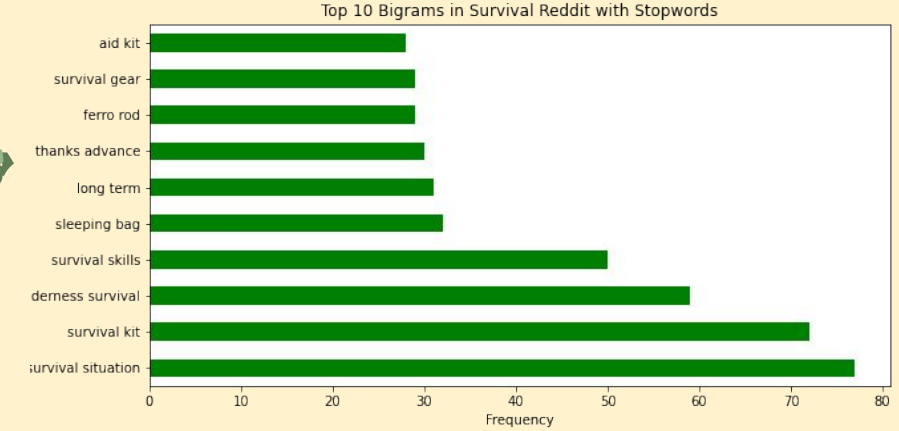
Zombie Survival Tactic

"Walking dead", "food water", "melee weapon", "baseball bat" are insightful bigrams. “Baseball bat” but with a popular walking dead wielding a baseball bat the buzz makes sense.



Survival

This by far is the most informative graph as we see key pairings in the subreddit such as "survival situation", "survival kit", and "wilderness survival" leading, but also key words such as "sleeping bag", "survival kit", and even "survival gear"



MODEL COMPARISON: WHICH IDENTIFIED THE SUBREDDITS THE BEST?

```
~~~~~~~~~~~~~~~~~~
Logistic Regression with CVEC
Best Params: {'cv__max_df': 0.65, 'cv__max_features': 3500, 'cv__min_df': 2, 'cv__ngram_range': (1, 2), 'lr__C': 1, 'lr__class_weight': None, 'lr__penalty': 'l2', 'lr__solver': 'libline
r'}
Best Training Score: 0.9704601990049752
Best Testing Score: 0.8516414141414141
Accuracy Score : 0.8516414141414141
Precision Score : 0.8862690707350902
Sensitivity Score (Recall) : 0.8068181818181818
Specificity: 0.898989898989899
F1 Score: 0.8446794448116326
~~~~~~~~~~~~~~~~~~
MNB with TF-IDF
Best Params: {'mnb__alpha': 1, 'tvec__max_df': 0.65, 'tvec__max_features': 4400, 'tvec__min_df': 2, 'tvec__ngram_range': (1, 1)}
Best Training Score: 0.9328358208955224
Best Testing Score: 0.8768939393939394
Accuracy Score : 0.8768939393939394
Precision Score : 0.8698884758364313
Sensitivity Score (Recall) : 0.8863636363636364
F1 Score: 0.8780487804878049
~~~~~~~~~~~~~~~~~~
RF with CVEC
Best Params: {'cvec__max_features': 3500, 'cvec__min_df': 3, 'cvec__ngram_range': (1, 1), 'rf__ccp_alpha': 0.0001, 'rf__class_weight': 'balanced', 'rf__max_depth': 20, 'rf__min_samples_
af': 1, 'rf__min_samples_split': 2, 'rf__n_estimators': 100, 'tfidf__use_idf': False}
Best Training Score: 0.888681592039801
Best Testing Score: 0.803030303030303
Accuracy Score : 0.803030303030303
Precision Score : 0.875
Sensitivity Score : 0.7070707070707071
F1 Score: 0.7821229050279329
```

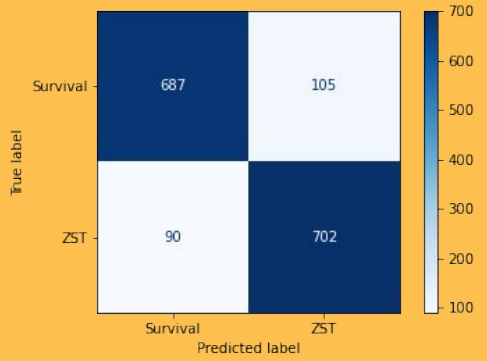
SUBREDDIT PREDICTION MATRIX

MNB with Stopwords

Training Score: 93%

Testing Score: 87%

Accuracy Score : 87%
Precision Score : 86%
Sensitivity Score (Recall) : 88%
F1 Score: 87%

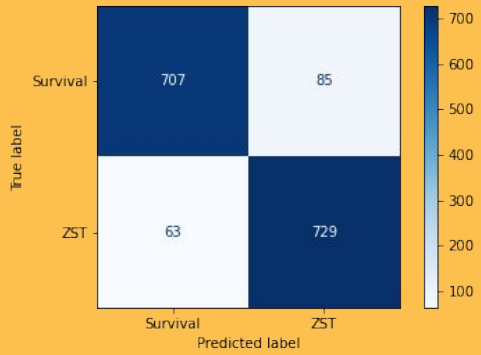


MNB without Stopwords

Training Score: 94%

Testing Score: 90%

Accuracy Score : 90%
Precision Score : 89%
Sensitivity Score (Recall) : 92%
F1 Score: 90%



CONCLUSION OF FINDINGS

- Multinomial Naive Bayes with TD-IDF is the best performing model to properly classify subreddit post origins. It had the highest accuracy and the best sensitivity score across the board minimizing false negatives
- If strong identifying words like "zombie" and "apocalypse were to be removed the training and test scores jump to 90% and 94% respectively with the highest metric scores
- As my goal was to optimize in accuracy and reducing misclassifications the MNB fits the goals wells. This could be because mnb classifiers are more suitable for text classification problems.

RECOMMENDATIONS

PERFECT PLACE FOR NEW PRODUCT IDEAS

Both subreddits showed language surrounding “useful” and “best” tools for survival

JOIN THE DISCUSSION!

Run polls , stimulate discourse, and use MNB to identify topics the market is passionate about

THE PREPPERS

Adding the r/Preppers community to a future study could provide a greater insight and boost ROI

THANK YOU!

Does anyone have any questions?

REFERENCES

- https://www.reddit.com/r/Survival/
- https://www.reddit.com/r/ZombieSurvivalTactics/
- https://thinkgrowth.org/the-doom-boom-inside-the-survival-industry-s-explosive-growth-2fece1f6cd6c
- https://trends.google.com/trends/explore?date=today%205-y&geo=US&q=survival,zombie
- https://today.yougov.com/topics/entertainment/articles-reports/2019/10/01/zombie-apocalypse-plan
- https://thehustle.co/coronavirus-prepping-doomsday-business/
- https://www.alliedmarketresearch.com/incident-and-emergency-management-market