Google Home or Google Pixel
Perspectives from subreddits

Sheng Jun
18 May 2020

AGENDA

CONTEXT (Slide 3)
1. PROBLEM STATEMENT
2. PROPOSED SOLUTION

DATA SOURCE
1. DATA

EDA
1. WORDART
2. VISUALS

MODEL & EVALUATE
1. MODEL WORKFLOW
2. PERFORMANCE SUMMARY

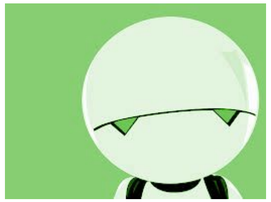
CONCLUSION (Slide 15)

CONTEXT

Background

Developing vision for next-gen IoT Devices

Trawl reddit posts for insights; understand User needs

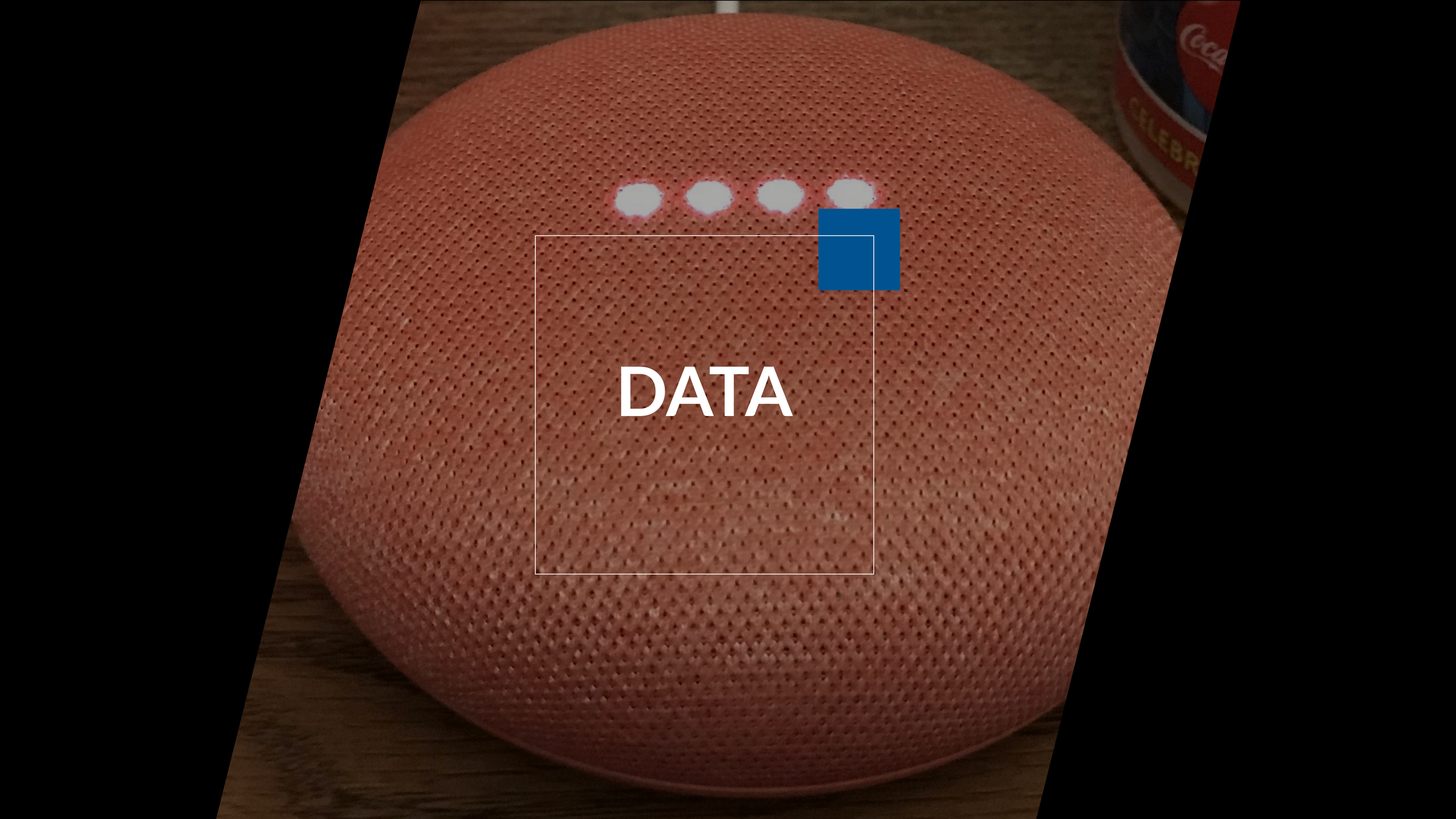


PROBLEM

Is there a better way to manual identification of trawled posts?

PROPOSED SOLUTION

NLP & ML; classify the posts


DATA

Data

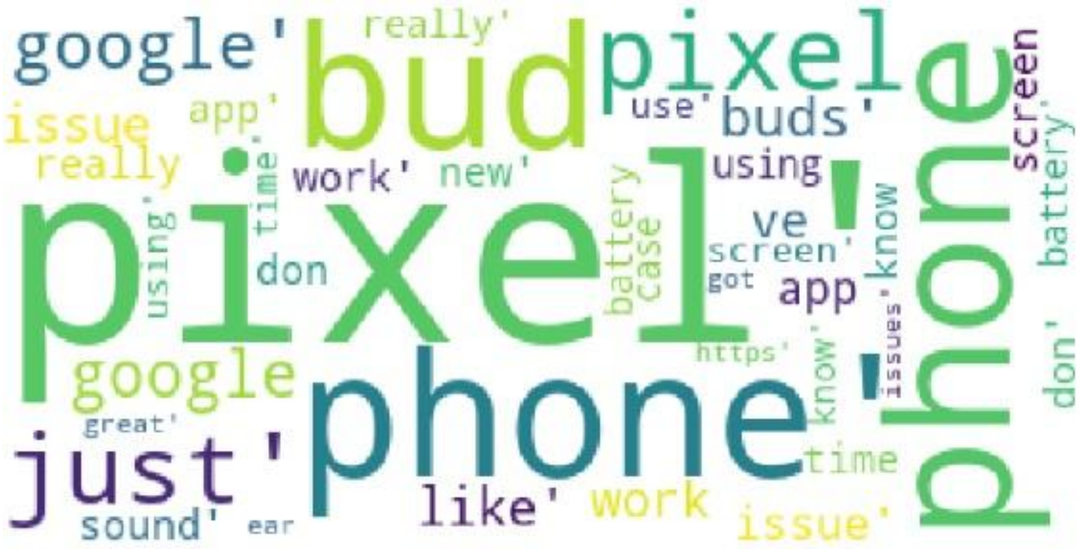
- **Relatively balanced classes.**
- **Baseline accuracy 50.68%.**

Count of Posts	Google Home	Google Pixel
Scraped	996	980
Removed duplicates	825	803
Drop Null values	824	802
Proportion (%)	**0.5068**	0.4932

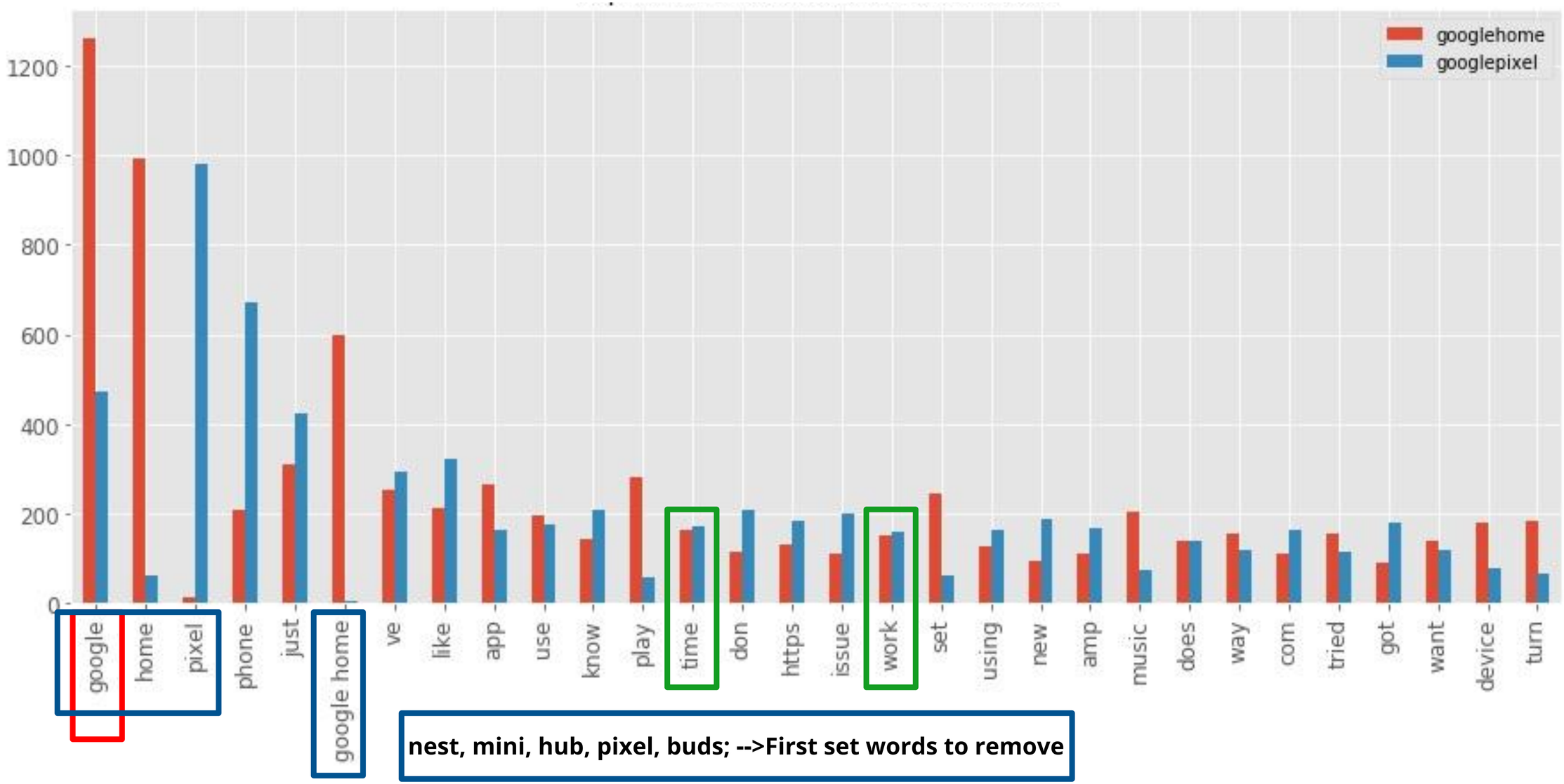

EDA

Word Art

- **Top 30 common words per sub-reddit**
- **Technically not Data science**

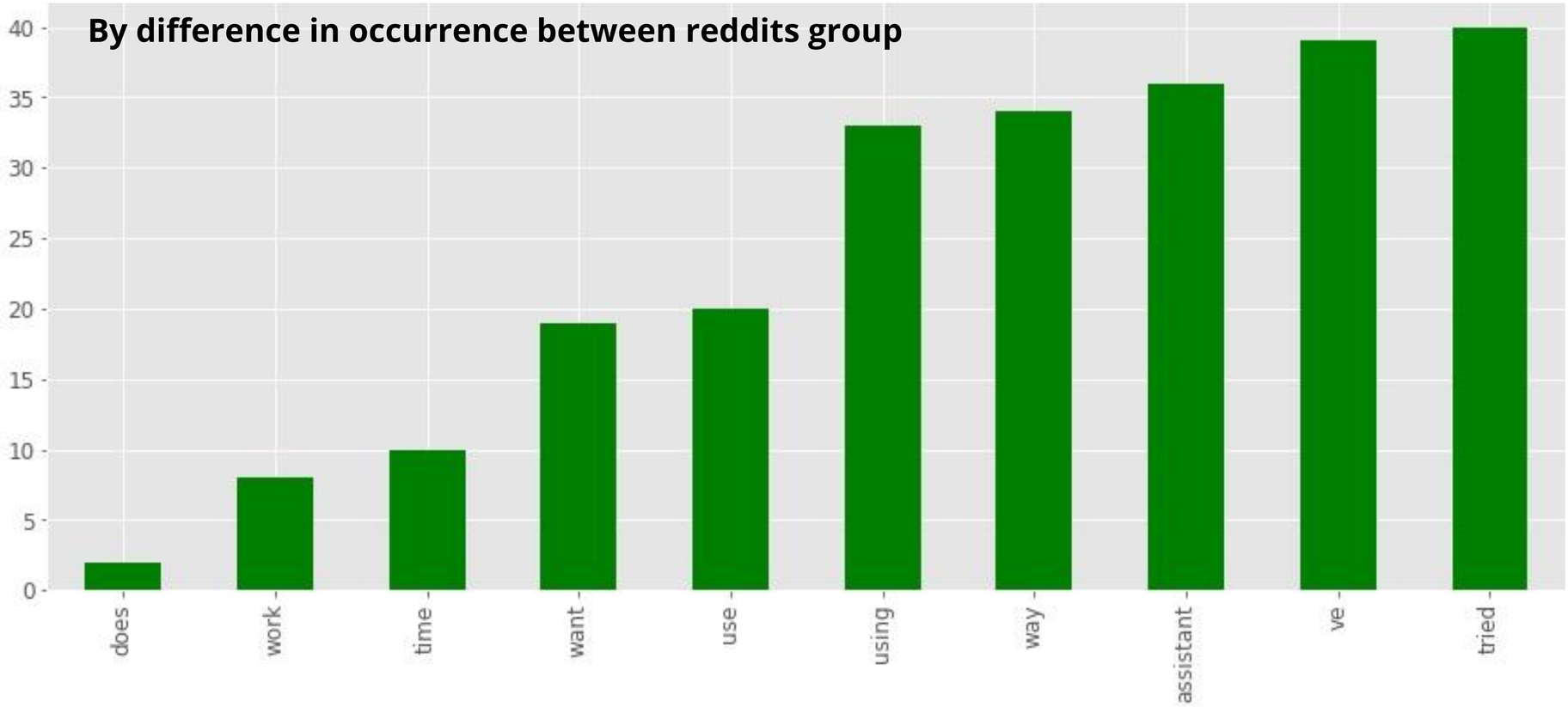


Top 30 Common Words (By Distribution)



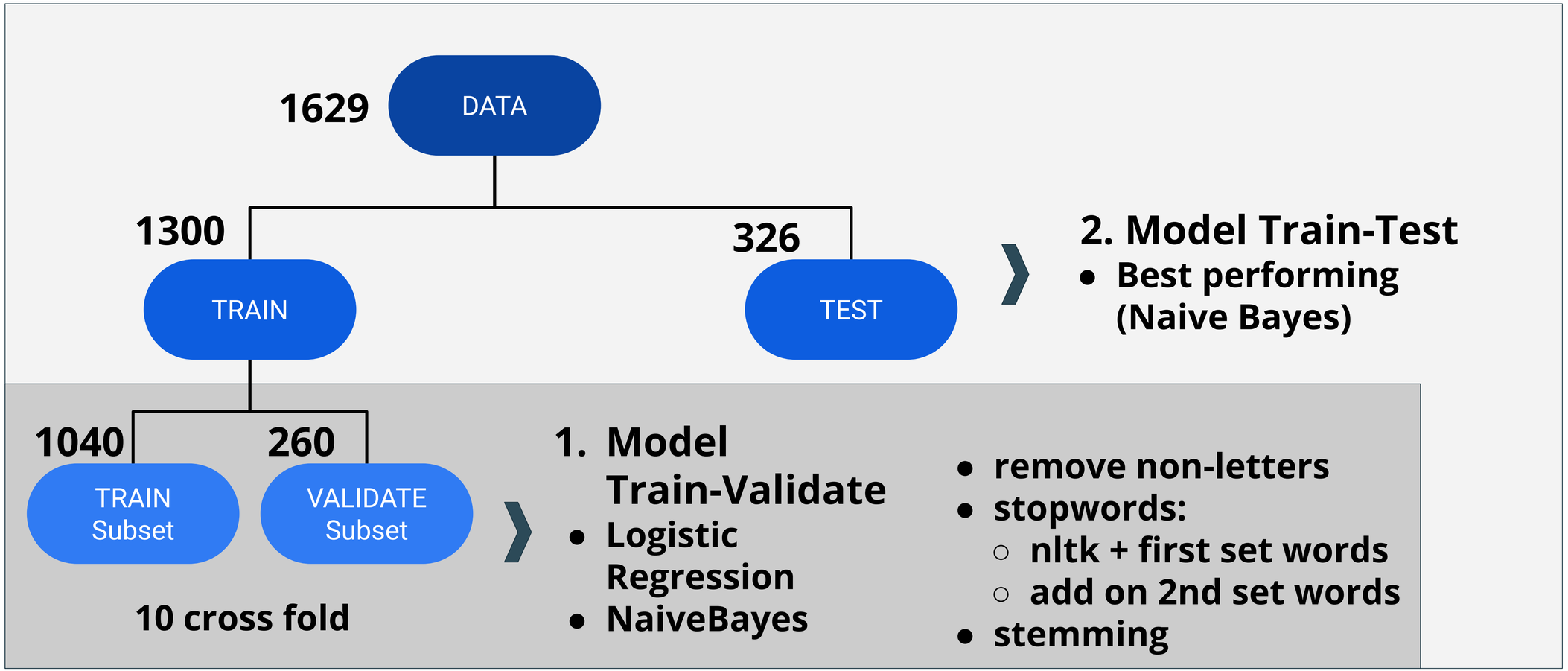
Top 10 Prevalent, Common Words

- **Minimal difference in occurrence between reddits;**
- **Occur more than 200 times**
- **Second set of words to remove**



MODEL & EVALUATE

Model WorkFlow



Performance Summary

1. Model Train-Validate

	LogReg_1	LogReg_2	NB_1	NB_2
accuracy	0.9308	0.9308	0.9423	0.9423
specificity	0.9297	0.9297	0.9062	0.9062
sensitivity	0.9318	0.9318	0.9773	0.9773
roc_auc	0.9910	0.9910	0.9924	0.9924

2. Model Train-Test

Confusion Matrix

	pred googlepixel	pred googlehome
Actual googlepixel	153	8
Actual googlehome	7	158

	Whole Train set	Test set
accuracy	0.9423	0.9540
specificity	0.9062	0.9503
sensitivity	0.9773	0.9576
roc_auc	0.9924	0.9840

- **Best Parameters**
  CountVectorizer(ngram_range=(1,2),max_df=0.9,min_df=3,max_features=4000)

- **Model does not appear to be overfitted on the whole train dataset**
- **Sensitivity and roc_auc above 95%**

Digging Deeper..

Top 50 words for positive class (Google Home)

```
['damn' 'as soon' 'the headphon' 'the iphon' 'charger and' 'chat with'
 'pattern' 'patch' 'the may' 'check for' 'charger' 'panel' 'the pair'
 'paid' 'the pixel' 'packag' 'overal' 'com galleri' 'com googlepixel'
 'other headphon' 'aptx' 'pair them' 'the replac' 'the galaxi' 'charg the'
 'cellular' 'chanc' 'the batteri' 'persist' 'the bud' 'the buzz'
 'percentag' 'charg and' 'the charg' 'the full' 'the charger' 'the comput'
 'the cord' 'the design' 'the ear' 'the earbud' 'the experi' 'charg it'
 'the fit' 'the flagship' 'charg case' 'the seri' 'comfort' 'optim'
 'thi phone']
```

Conclusion

- **Naive Bayes Model selected as production model.**
  Sensitivity > 95%
  ROC AUC >98%

- **Potentially, further tuningemented post-deployment by removing the words attributing to false classifications**
- **Caution to prevent over-tuning (Model loses its generalizability).**

- **Reassess the removed words when re-deployed to classifycompetitor's reddit posts**

Google Home or Google Pixel
Perspectives from subreddits

Thank you!
Q & A