Amazon Alexa VS Google Home

Project3 : API & Text Classification

Pratch Luepuwapitakkul

U.S. Smart Speaker Market Share by Brand

January 2018 & 2019

2019

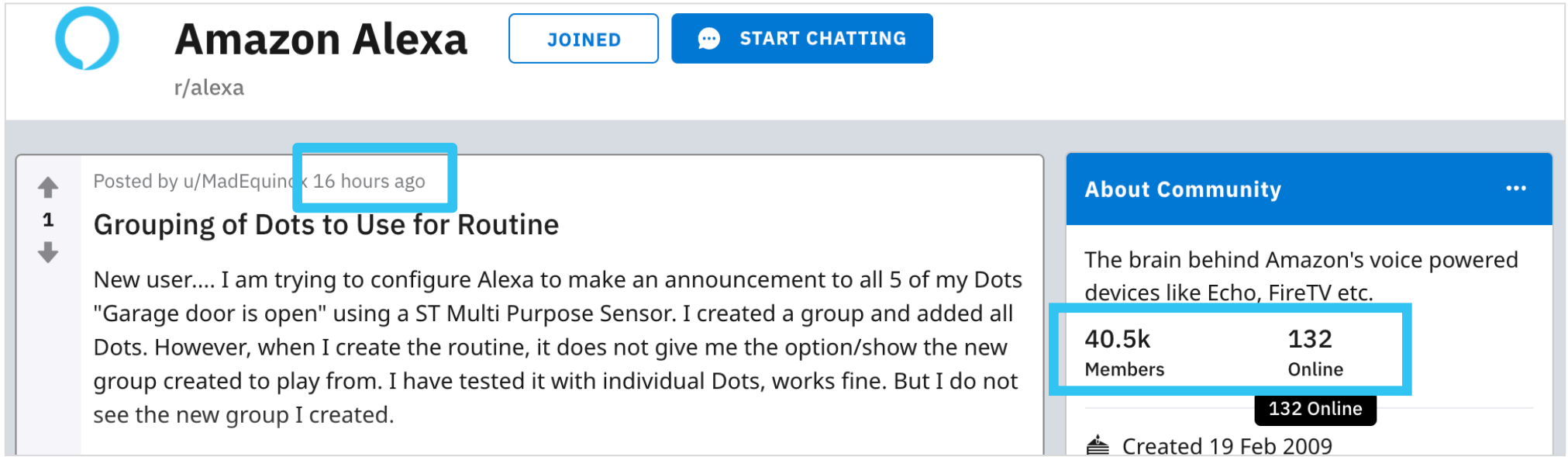
user-created topics *'subreddits'*

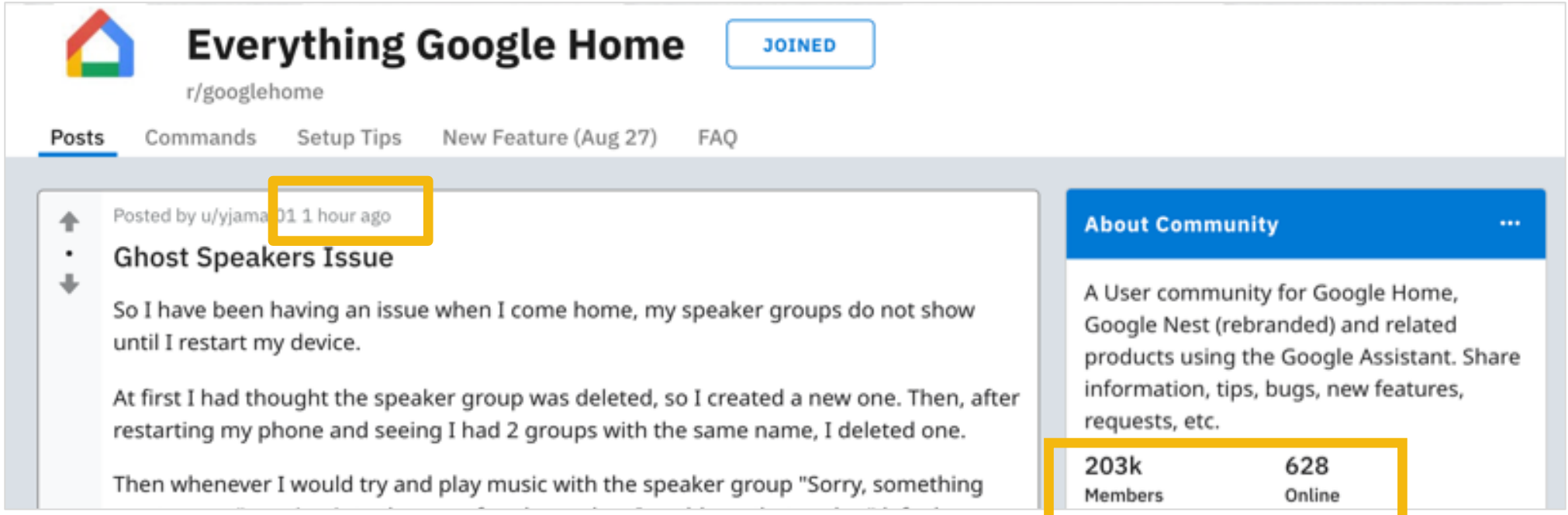
19th most-visited site in the U.S. and in the world *- May 2020*

'Real Consumer Feedback'

community discussion

similar user profiles *(as 'Virtual Assistants Devices'* >40% US *+ UK/Canada/Germany*


Amazon Alexa
JOINED
START CHATTING
r/alexa
Posted by u/MadEquinox 16 hours ago
1
Grouping of Dots to Use for Routine
New user.... I am trying to configure Alexa to make an announcement to all 5 of my Dots "Garage door is open" using a ST Multi Purpose Sensor. I created a group and added all Dots. However, when I create the routine, it does not give me the option/show the new group created to play from. I have tested it with individual Dots, works fine. But I do not see the new group I created.
About Community
The brain behind Amazon's voice powered devices like Echo, FireTV etc.
40.5k
Members
132
Online
132 Online
Created 19 Feb 2009


Everything Google Home
JOINED
r/googlehome
Posts
Commands
Setup Tips
New Feature (Aug 27)
FAQ
Posted by u/yjama01 1 hour ago
Ghost Speakers Issue
So I have been having an issue when I come home, my speaker groups do not show until I restart my device.
At first I had thought the speaker group was deleted, so I created a new one. Then, after restarting my phone and seeing I had 2 groups with the same name, I deleted one.
Then whenever I would try and play music with the speaker group "Sorry, something
About Community
A User community for Google Home, Google Nest (rebranded) and related products using the Google Assistant. Share information, tips, bugs, new features, requests, etc.
203k
Members
628
Online

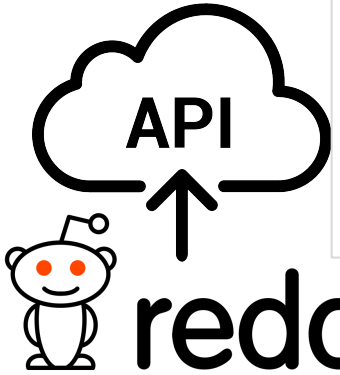
Missing labels in the data!!!

Problem Statement

Problem 1:
Classify the texts with missing label

- Text classification model
- Focus on correct classification for `alexa` (positive class)

Problem 2:
Identify areas of opportunities for market expansion to Asia



Gather Data:

Assign positive/negative class

Remove Duplicates

Drop N/A

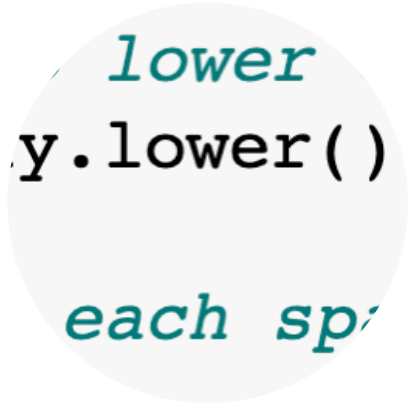
Convert Date/Time Object

	Googlehome 0	Alexa 1	Total
total	722	684	1406
train	541	513	1054
test	181	171	352

	subreddit	selftext	ups	title	created_utc	subreddit_class
0	alexa	title. \n\n\nthe actual devices are fine, but...	170	the alexa app is crap. it's slow, laggy, fully...	2020-06-21 17:23:42	1
1	alexa	There has been a persistent referral link spam...	99	[Announcement] Raising the minimum karma neede...	2020-04-16 06:02:27	1
2	alexa	"Alexa, I've put my red jacket in the storage ...	78	Reminders like this please.	2020-06-29 08:11:47	1
3	alexa	I know this might be a picky thing, but it ser...	72	Anyone Else Get Really Annoyed by Alexa's "By ...	2020-08-22 19:42:57	1
4	alexa	No karma whoring, likes, subscribes, whatever....	70	Looking to give away an Alexa Echo to someone ...	2020-08-20 01:49:08	1
...	...	...	...	...	...	...
1403	googlehome	Chromecast Audio hooked into the PC line-in wo...	0	Still looking for a way to yell at my Google H...	2020-08-12 03:48:10	0
1404	googlehome	We love having Google tell the corny jokes (da...	0	Turning off Santa jokes...	2020-08-11 12:24:44	0
1405	googlehome	Wouldn't it be good if you could change the tr...	0	Wish you could change the wakeup word	2020-08-11 10:26:08	0
1406	googlehome	I would like to get an answer of the current s...	0	What are few possible utterances to query a vo...	2020-08-11 07:52:59	0
1407	googlehome	Hi,\n\nI am developing Google smart home appli...	0	What are the utterance for "action.devices.com...	2020-08-11 07:41:36	0

1408 rows × 6 columns

Text Data Cleaning:



Before:

['Someone requested an invitation to join my home']
['Hi reddit\n\nI have a Google Nest Mini connected to my Google home, and I was wondering if you could help me, since I can’t find anything on Google support. \n\nThree times I received emails about someone requested an invitation to join my home, and it’s people I don’t know. \n\nI thought people could only request access if they are on the same wifi as the Google Home? Should I be concerned about someone having access to my wifi, even though I changed the wifi password after I received the first email.\n\nI just decline the requests in the Google Home app, but I’m annoyed that someone in the first place is able to request access.\n\nSincerely, \n\nJeppe']

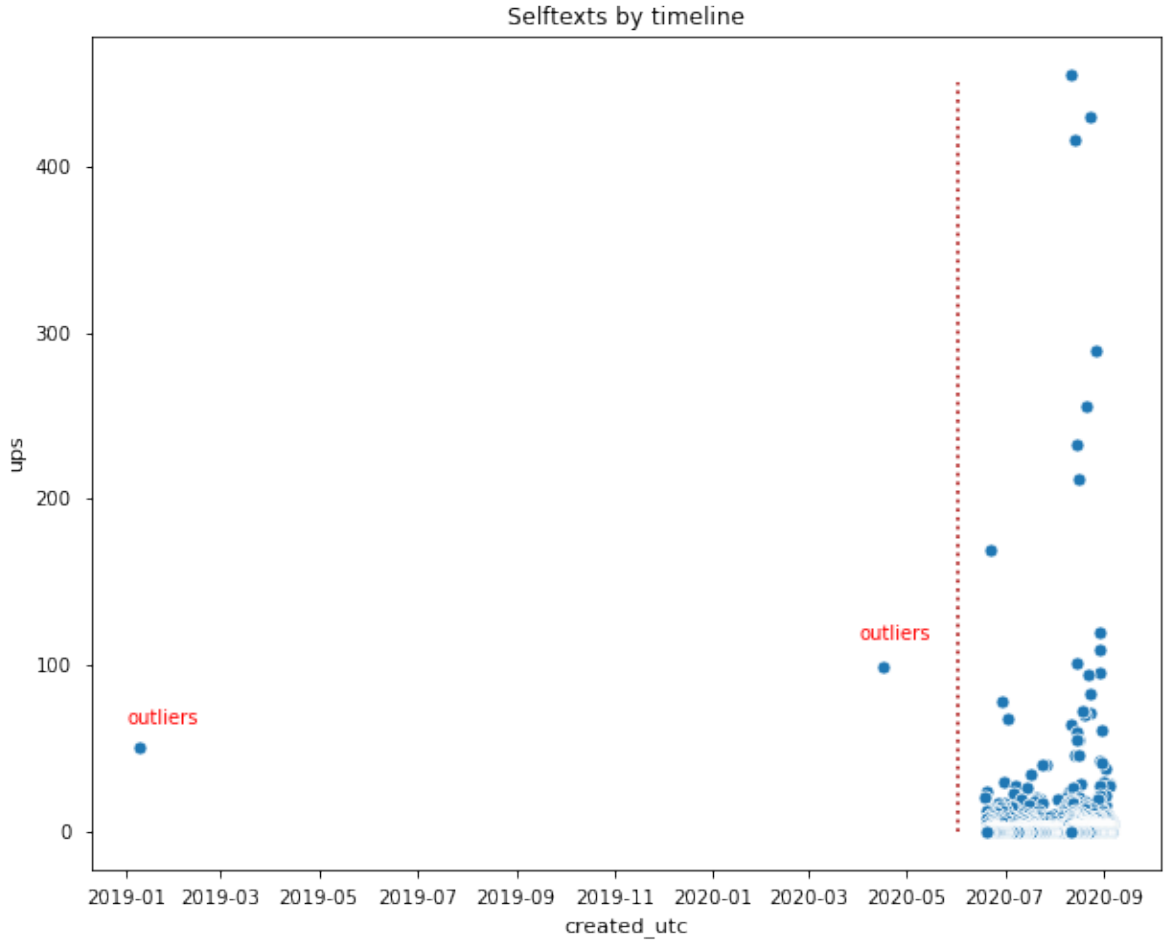
After:

['someon request invit join hi reddit mini connect wonder could help sinc find anyth support three time receiv email someon request invit join peopl know thought peopl could request access wifi concern someon access wifi even though change wifi password receiv first email declin request annoy someon first place abl request access sincer jepp']

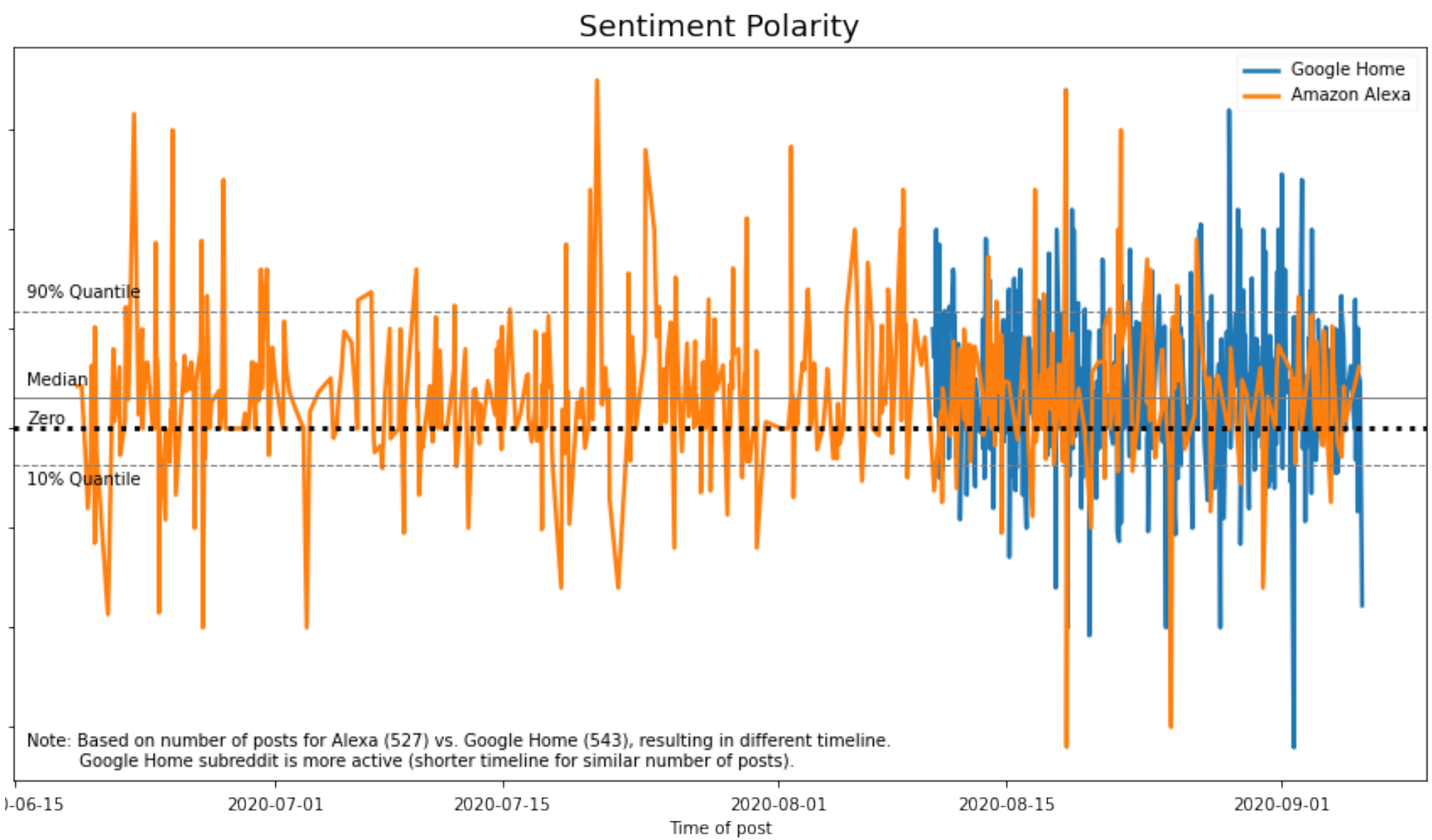
Virtual Home Assistant Devices


volum
make
two
alway
question
turn
spotifi
smart
sound
song
right
via
someth
live room
think
recent
fix
also
call
connect
seem
light
control
enabl
happen
time
thing
figur
speaker
realli
phone
playlist
voic
said
everyth
go
chang
routin
hous
day
problem
option
video
switch
music
fine
announc
stop
answer
help
possibl
list
start
need
creat
abl
keep
current
tell
open
see
anyth
well
dot
way
even
command
without
know
respond
screen
name
new
look
idea
sure
still
back
skill
minut
specif
group
could
tv
differ
got
alarm
first
show
tri
account
find
anyon
room
thank
ad
issu
remind
add
recent
realli
call
setup
routin
still
speaker
see
hous
anyon els
help
hub
look
control
room
volum
anyon
happen
issu
right
sound
bulb
updat
chang
live room
know
anyth
back
tri
option
mini
account
well
go
router
light
alarm
find
also
thank
howev
start
command
two
new
keep
etc
day
think
make
fine
with
way
via
respond
got
possibl
turn
switch
wifi
idea
everi
need
song
sinc
gh
time
add
problem
video
stop
show
tell
voic command
differ
fix
someth
music
name
hey
cast
ad
sure
tv
voic
seem
first
even
open
hi
noth
phone
could
chromecast

Drop Outliers – admin posts



title	created_utc
[Announcement] Raising the minimum karma neede...	2020-04-16 06:02:27
FAQ: Please read the subreddit FAQ before post...	2019-01-09 19:35:58


Sentiment Polarity
Google Home
Amazon Alexa
90% Quantile
Median
Zero
10% Quantile
Note: Based on number of posts for Alexa (527) vs. Google Home (543), resulting in different timeline.
Google Home subreddit is more active (shorter timeline for similar number of posts).
-06-15
2020-07-01
2020-07-15
2020-08-01
2020-08-15
2020-09-01
Time of post

CountVectorizer Pipeline:

```
cv_pipe = Pipeline([
    ('cvec', CountVectorizer()),
    ('nb', MultinomialNB())])
```

```
cv_param_grid = {'nb__alpha': [0.3, 0.4,0.5],
                 'cvec__ngram_range' : [(1,1), (1,2),(1,3)],
                 'cvec__max_features' : [2000, 2500, 2800],
                 'cvec__min_df' : [1,2,3],
                 'cvec__max_df' : [0.2, 0.3,0.4]
                }
```

```
gs_cv = GridSearchCV(cv_pipe, cv_param_grid, cv=5, verbose=1, n_jobs=-1)
best_model = gs_cv.fit(clean_train_selftexts, y_train)
```

```
Fitting 5 folds for each of 243 candidates, totalling 1215 fits
```

```
[Parallel(n_jobs=-1)]: Using backend LokyBackend with 8 concurrent workers.
[Parallel(n_jobs=-1)]: Done  34 tasks       | elapsed:    2.2s
[Parallel(n_jobs=-1)]: Done 184 tasks       | elapsed:    6.3s
[Parallel(n_jobs=-1)]: Done 434 tasks       | elapsed:   12.7s
[Parallel(n_jobs=-1)]: Done 784 tasks       | elapsed:   22.1s
[Parallel(n_jobs=-1)]: Done 1215 out of 1215 | elapsed:   33.9s finished
```

TfidfVectorizer Pipeline:

```
tv_pipe = Pipeline([
    ('tvec', TfidfVectorizer()),
    ('nb', MultinomialNB())])
```

```
tv_param_grid = {'nb__alpha': [0.6,0.7,0.8],
                 'tvec__ngram_range' : [(1,2),(1,3),(1.4)],
                 'tvec__max_features' : [2500, 2800, 3000],
                 'tvec__min_df' : [1,2,3],
                 'tvec__max_df' : [0.1, 0.2, 0.3,0.4]
                }
```

```
gs_tv = GridSearchCV(tv_pipe, tv_param_grid, cv=5, verbose=1, n_jobs=-1)
best_model = gs_tv.fit(clean_train_selftexts, y_train)
```

```
[Parallel(n_jobs=-1)]: Using backend LokyBackend with 8 concurrent workers.
```

```
Fitting 5 folds for each of 324 candidates, totalling 1620 fits
```

```
[Parallel(n_jobs=-1)]: Done  34 tasks       | elapsed:    1.0s
[Parallel(n_jobs=-1)]: Done 488 tasks       | elapsed:   12.2s
[Parallel(n_jobs=-1)]: Done 1488 tasks      | elapsed:   36.4s
[Parallel(n_jobs=-1)]: Done 1605 out of 1620 | elapsed:   39.2s remaining:    0.4s
[Parallel(n_jobs=-1)]: Done 1620 out of 1620 | elapsed:   39.6s finished
```

Naïve-Bayes

	model	classifier	vectorizer	cv_scores
0	gs_cv	naive_bayes	countvectorizer	0.773279
1	gs_tv	naive_bayes	tfidfvectorizer	0.758086

```
class_param_grid = [{'classifier': [MultinomialNB()],
                     'classifier__alpha': [0.5, 0.7]
                    },
                    {'classifier': [LogisticRegression()],
                     'classifier__solver': ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga']
                    },
                    {'classifier': [KNeighborsClassifier()],
                     'classifier__n_neighbors' : [5, 8, 12],
                     'classifier__metric': ['euclidean'],
                     'classifier__weights' : ['uniform', 'distance']
                    },
                    {'classifier': [RandomForestClassifier()],
                     'classifier__random_state' : [42],
                     'classifier__max_depth' : [10, 13, 15],
                     'classifier__n_estimators' : [300,350],
                     'classifier__min_samples_split' : [4,7,10],
                     'classifier__min_samples_leaf' : [2,3],
                     'classifier__ccp_alpha' : [0.001, 0.002]
                    },
                    {'classifier': [SVC()],
                     'classifier__C' : [0.1, 0.5, 1],
                     'classifier__kernel' : ['linear', 'rbf']
                    },
                    {'classifier': [AdaBoostClassifier()],
                     'classifier__base_estimator' : [DecisionTreeClassifier()],
                     'classifier__n_estimators' : [75,100, 150],
                     'classifier__base_estimator__max_depth' : [1,3],
                     'classifier__learning_rate' : [0.9]
                     }]
```

```
gs_c = GridSearchCV(class_pipe, class_param_grid, cv=5, verbose=1, n_jobs=-1)
best_model = gs_c.fit(cvec_train, y_train)
```

```
Fitting 5 folds for each of 97 candidates, totalling 485 fits
```

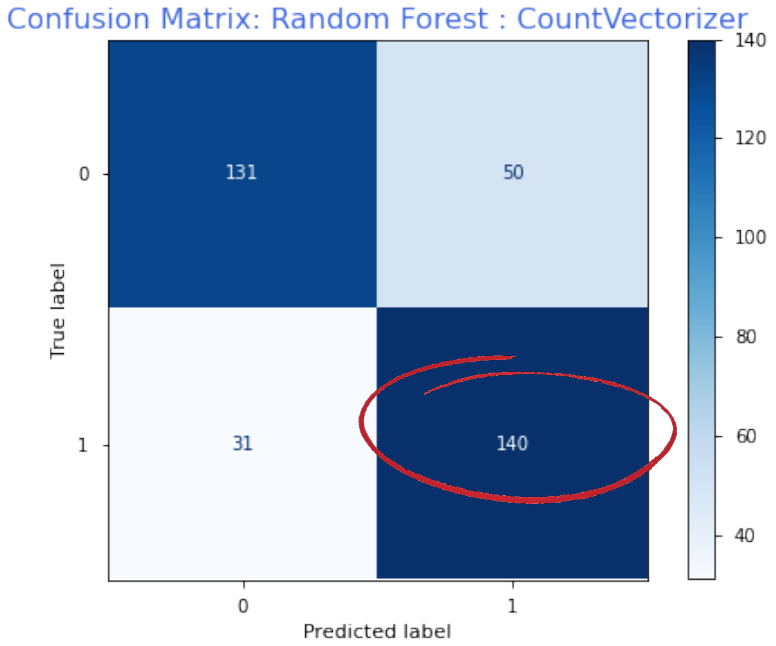
```
[Parallel(n_jobs=-1)]: Using backend LokyBackend with 8 concurrent workers.
[Parallel(n_jobs=-1)]: Done  56 tasks      | elapsed:    0.4s
[Parallel(n_jobs=-1)]: Done 360 tasks      | elapsed:   37.4s
[Parallel(n_jobs=-1)]: Done 485 out of 485 | elapsed:   49.5s finished
```

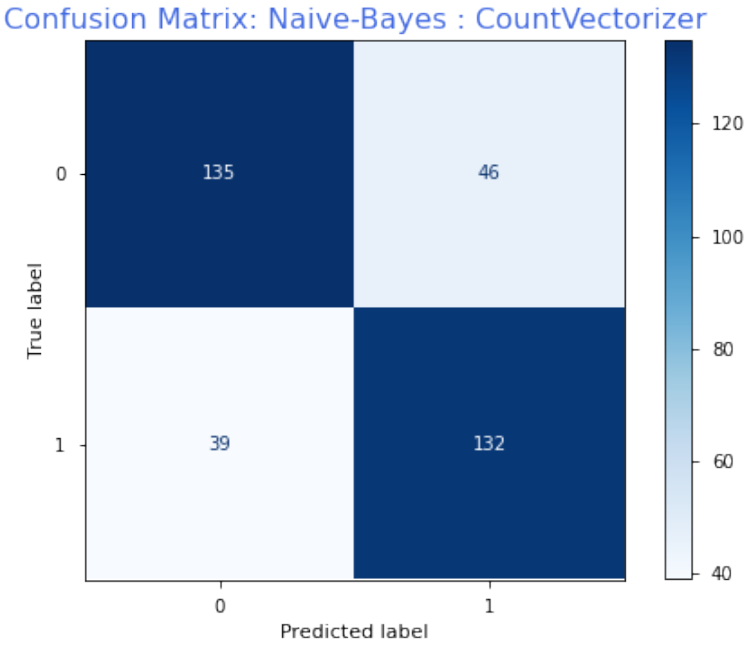
5 Classifiers & CountVectorizer

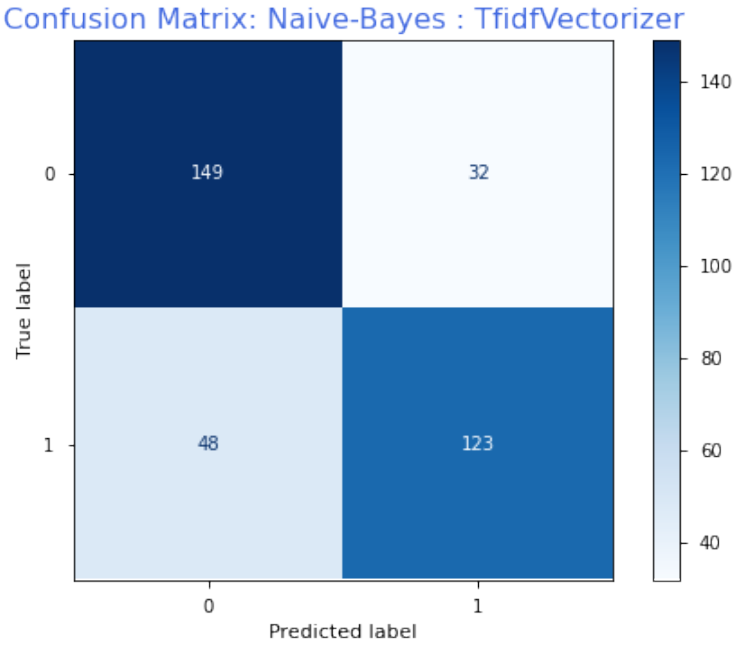
	params	mean_test_score	rank_test_score
72	{'classifier': RandomForestClassifier(ccp_alpha=0.002, max_depth=13, min_samples_leaf=3, min_samples_split=10, n_estimators=350, random_state=42), 'classifier__ccp_alpha': 0.002, 'classifier__max_depth': 13, 'classifier__min_samples_leaf': 3, 'classifier__min_samples_split': 10, 'classifier__n_estimators': 350, 'classifier__random_state': 42}	0.803606	1
92	{'classifier': AdaBoostClassifier(), 'classifier__base_estimator': DecisionTreeClassifier(), 'classifier__base_estimator__max_depth': 1, 'classifier__learning_rate': 0.9, 'classifier__n_estimators': 100}	0.774209	73
6	{'classifier': LogisticRegression(), 'classifier__solver': 'saga'}	0.774177	74
1	{'classifier': MultinomialNB(), 'classifier__alpha': 0.7}	0.754299	81
90	{'classifier': SVC(), 'classifier__C': 1, 'classifier__kernel': 'rbf'}	0.747601	84
10	{'classifier': KNeighborsClassifier(), 'classifier__metric': 'euclidean', 'classifier__n_neighbors': 8, 'classifier__weights': 'distance'}	0.607195	91

Random Forest is best at classifying positive class "Alexa"

	model	classifier	vectorizer	cv_scores	train_score	test_score	accuracy_score	misclassification_rate	sensitivity	specificity	precision
3	gs_rf	random_forest	countvectorizer	0.804550	0.891841	0.769886	0.769886	0.230114	0.818713	0.723757	0.736842
0	gs_cv	naive_bayes	countvectorizer	0.773279	0.914611	0.758523	0.758523	0.241477	0.771930	0.745856	0.741573
1	gs_tv	naive_bayes	tfidfvectorizer	0.758086	0.936433	0.772727	0.772727	0.227273	0.719298	0.823204	0.793548

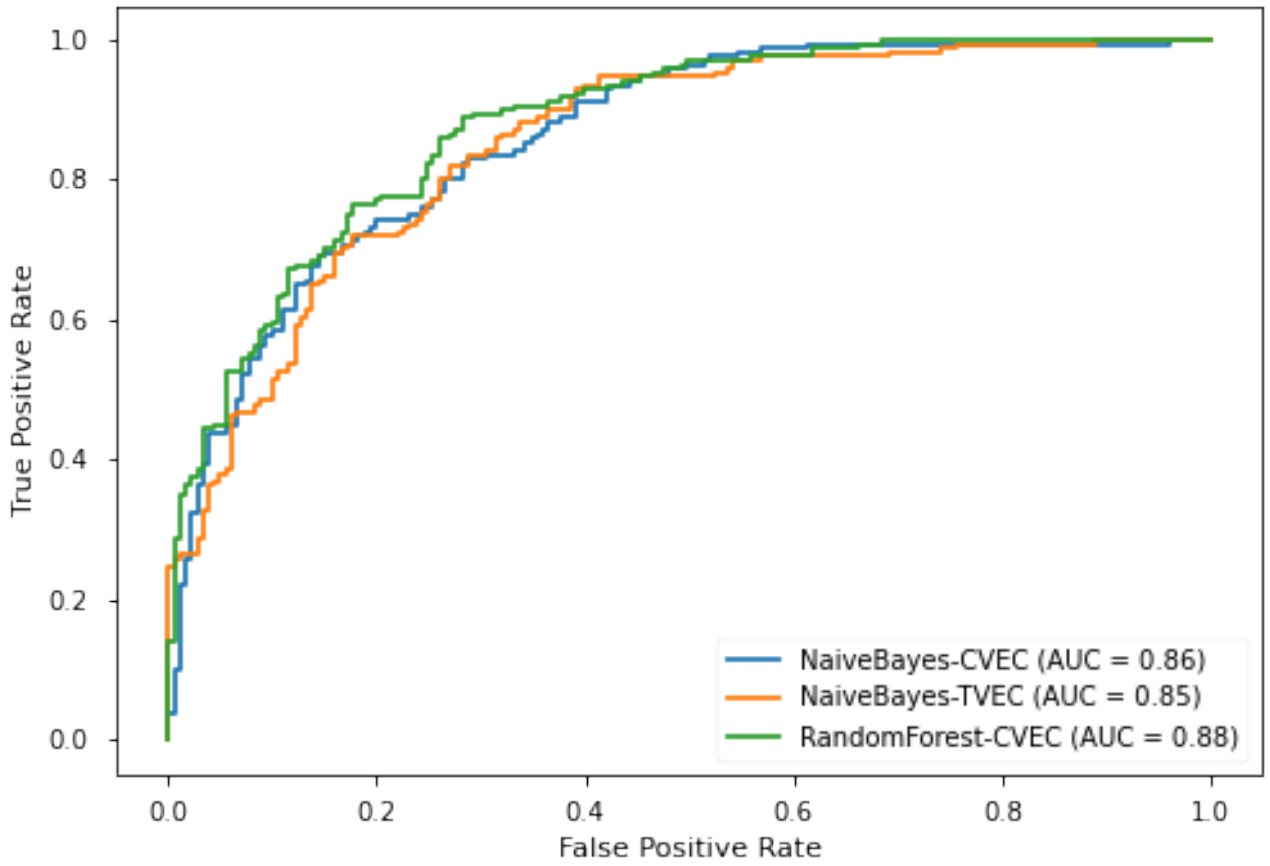






Highest AUC

- Correctly classify positive class 'Alexa'

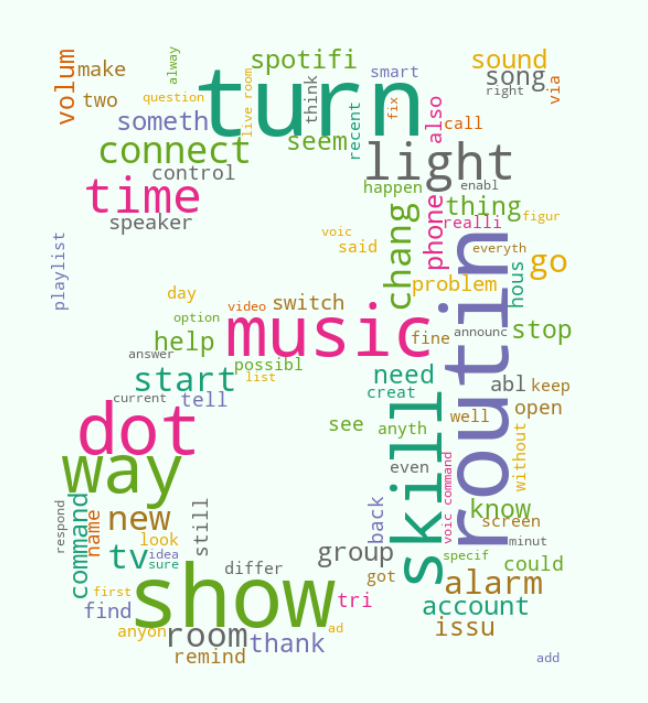


Conclusion 1:

- Model with RandomForest & CountVectorizer correctly classify texts with:
  - Accuracy 80.4%
  - Sensitivity 81.2% - correctly predicts positive class.

- Limitations & Area of Improvement:
  - More data for positive class to better train the model

Conclusion 2:



- This project provides a starting point for Marketing Analytics. Further analysis is needed.
- Most frequent words could be what users are interested in.
  - Echo Dot (small format) is most talked about and could be potential for market expansion.
  - Further sentiment analysis
  - Topic Modelling – esp related to Alexa features, i.e. routine, skill.
- Expand to other forums beyond Reddit to capture other profiles.