Twitter Brand Support

A DS201X Final Project by ***James Taylor***

Stage 1

Ask a Question

The Dataset

“Customer Support on Twitter”

Over 3 million tweets and replies from the biggest brands on Twitter

Available on Kaggle:
https://www.kaggle.com/thoughtvector/customer-support-on-twitter

The Dataset Content

tweet_id - unique, anonymized ID for tweet

author_id - unique, anonymized user ID

inbound - whether tweet was sent to company

created_at - date/time when tweet sent

text - tweet content

response_tweet_id - IDs of tweets that are responses to this tweet

in_response_to_tweet_id - ID of tweet this was in response to, if any

Questions/Answers & Patterns Expected

- Q: Can we figure out, with some high certainty, what company a person was talking about just based on the tweet - disregarding the mention of the actual company twitter account

- I expect that this can be done very well, and I expect can do it to some extent. Based on brand names, product names, and services and products offered I think it should be relatively effective. (if someone mentions ios11 or iphone, it was probably meant for Apple)

Benefits to the Organization

We can use this to automatically track down problems people are having on Twitter when they don't specifically @ our support team. If it is a close enough match to previous tweets that were @ our team, we can have a human review to see if it is actually about our service/product and needs addressing.

Stage 2

Get the Data

Obtaining the Data

Data is premade from Kaggle:

https://www.kaggle.com/thoughtvector/customer-support-on-twitter

Clean & Prepare Data

Text needs a lot of preprocessing, so here is what I did - more could certainly be done, and it could be done better.

1. Converted tweets to conversations based on tweet_id and in_response_to_tweet_id
2. Dropped all columns except tweet from user, company who replied, and company's reply (unsure if using last part)
3. Dropped any conversation excluding the company replying
4. Removed all @mentions and links
5. Make everything lowercase
6. Remove all punctuation
    - !"#$%&'()*+,-./:;<=>?@[\]^_`{|}~'“”

Clean & Prepare Data (cont.)

7. Remove stopwords (words with little meaning, used to make English ‘flow’). Used nltk library of english stopwords and filtered out every tweet

STOPWORDS:['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]

Clean & Prepare Data (cont.)

8. Found and removed 100 most common words from all tweets (after stop words removed) - those that would not help the model differentiate.

10 most common: us, please, dm, help, hi, sorry, get, thanks, im, like

9. Stemming?

- Break words down to just their stem so the root and meaning of the word remain, while the part of speech is removed
- Had at one point, but was actually hurting my results across the board
- More research needed

Need to acquire more related data?

- Dataset started with 3M+ Tweets
- After filtering and cleaning, 650K Tweets
- Still a very large data set, can’t imagine any more data would help

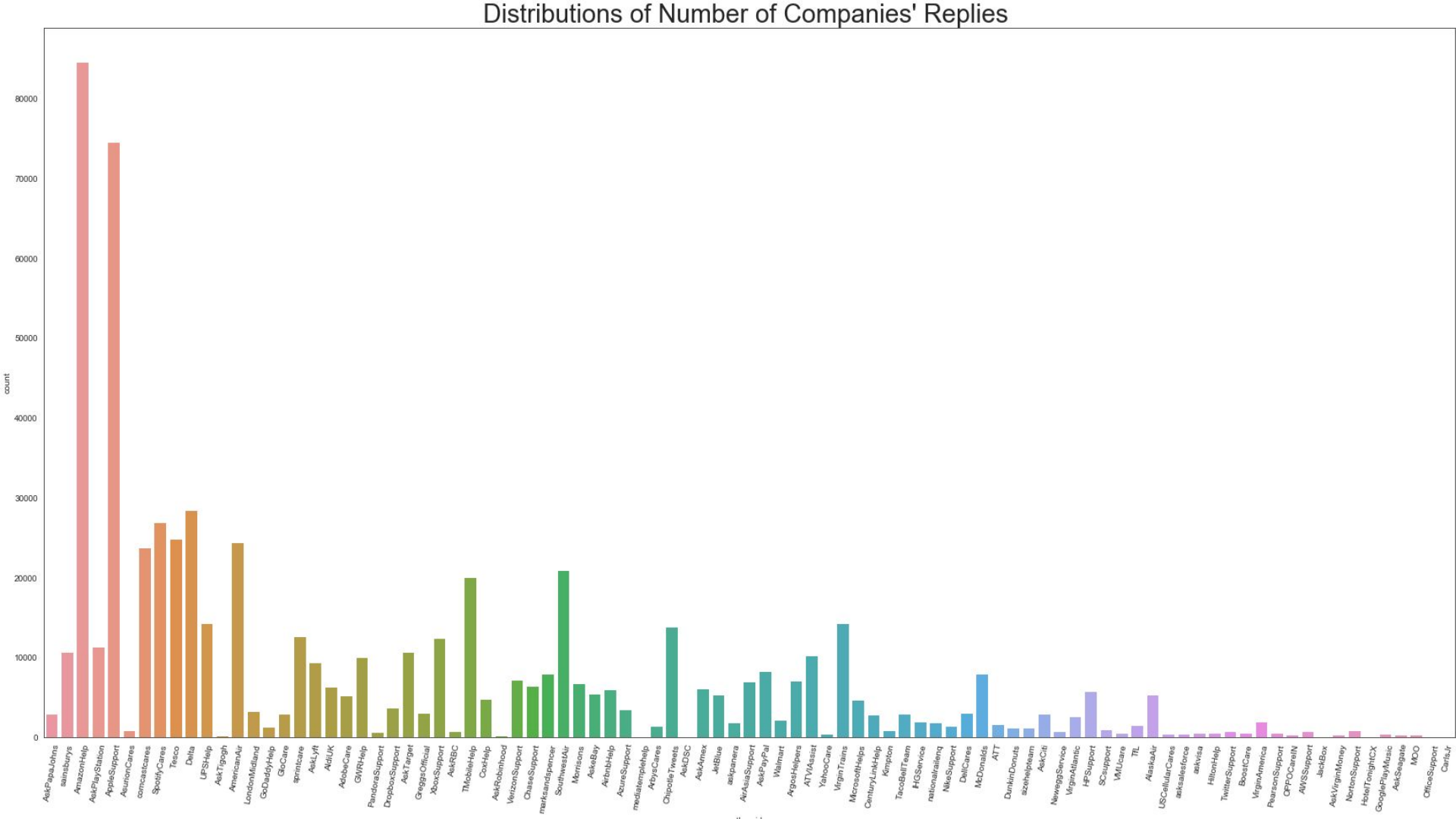
Stage 3

Explore the Data

Descriptive Analysis - # Reply by Company

Top Companies:

1. AppleSupport - 50K
2. AmazonHelp - 50K
3. Uber_Support - 24K
4. Delta - 19K
5. SpotifyCares - 16K


Distributions of Number of Companies' Replies
count
0
10000
20000
30000
40000
50000
60000
70000
80000
AskPapaJohns
sainsburys
AmazonHelp
AskPlayStation
AppleSupport
AsurionCares
comcastcares
SpotifyCares
Tesco
Delta
UPSHelp
AskTigogh
AmericanAir
LondonMidland
GoDaddyHelp
GloCare
tmobilecare
AskLyft
AldiUK
AdobeCare
GWRHelp
PandoraSupport
DropboxSupport
AskTarget
GreggsOfficial
XboxSupport
AskRBC
TMobileHelp
CoxHelp
AskRobinhood
VerizonSupport
ChaseSupport
marksandspencer
SouthwestAir
Morrisons
AskeBay
AirbnbHelp
AzureSupport
mediatempleHelp
AtoysCares
ChipotleTweets
AskDSC
AskAmex
JetBlue
asksainsburys
AirAsiaSupport
AskPayPal
Walmart
ArgosHelpers
ATVIAssist
YahooCare
VirginTrains
MorrisonsHelps
CenturyLinkHelp
Kimpton
TacoBellTeam
IHGService
nationalrailenq
NikeSupport
DellCares
McDonalds
ATT
DunkinDonuts
sizehelpteam
AskCiti
NeweggService
VirginAtlantic
HPSupport
SCsupport
VMUcare
TfL
AlaskaAir
USCellularCares
asksalesforce
askvisa
HiltonHelp
TwitterSupport
BoostCare
VirginAmerica
PearsonSupport
OPPOCareIN
AWSSupport
JackBox
AskVirginMoney
NortonSupport
HotelTonightCX
GooglePlayMusic
AskSeagate
MOO
OfficeSupport
CarisJr

Descriptive Analysis - Top Words by Company

Found top 10 words for every company just to see if my ideas that words would differ was correct, here is what I found from just a few companies that shows correct:

- ATT: internet, pay, att, atampt, bill
- AdobeCare: cc, adobe, lightroom, premier, photoshop
- AirbnbHelp: airbnb, host, booking, reservation, book
- AlaskaAir: alaska, iflyalaska, delayed, first, flights
- Aldi: aldi, christmas, kevin, bought, stores
- AmazonHelp: prime, delivered, de, ordered, package, days
- AmericanAir: gate, plane, delayed, aa, flights, fly, flying

Get to Know Data - Hypotheses & Patterns

- Effective if good pre processing and good choice of classification algorithm
- The top words for each of the companies is quite different, so the model should be able to separate them out and make good predictions
- Few overlapping words in similar services (flight/gate/delayed for Airlines, etc.)

Problems with Dataset, Improvements?

- Top words could be fined down even more
  - Where do we stop?
  - Don’t want to lose too much data
- Smarter preprocessing
  - Difficult to know what to do and what not to do

Stage 4

Model The Data

Selection of ML model

- Obvious classification problem
- Here is a tweet, which one of these ~100 companies was it meant for (after removing the mention, that would be too obvious)
- First thought was to use Knn, but that would prove to be one of the worse options

Accuracy measurement, train/test split

- Accuracy Measurement with Decision Tree: 53%
- So, 53% of the time when given a tweet with the mention removed and grammar stripped down, it will return the correct company it was meant for.

- Train Split: 600,000 tweets
- Test Split: Remaining ~ 43,000 tweets

Improving Features

- Text is tricky
- Smarter preprocessing needed
  - What other ways should it be processed?
  - In what different ways should it be processed?
  - Was anything over processed?
- Stemming?

Model Evaluations

Decision Tree: 53.15%

Naive Bayes: 41.4%

Knn, k = 1: 33.8%

Knn, k = 3: 29.6%

Knn, k = 5: 27.25%

Knn, k = 7: 25.24%

Alternative Approaches

Another thing I may have done if time permitted was work out how to return the more probable results. Like: Apple - 70%, Comcast - 25%, Samsung - 3%, ATVI - 2%

I also preprocessed the results from the company but that was out of the scope out of my question, but perhaps that could be used to support what kind of words are used when talking about that companies products or services

Stage 5

Communicate the Data

What was already there, what was added?

The csv file was the only thing that was there, but I took parts of my project from three different places, altered them to fit my use case, and added other parts as needed:

Text Preprocessing:

https://www.kaggle.com/sudalairajkumar/getting-started-with-text-preprocessing

https://www.kaggle.com/snocco/textpreprocessing-on-twitter

Using Machine Learning with Text:

https://towardsdatascience.com/machine-learning-text-processing-1d5a2d638958

Prescriptive Analysis - Actionable Results, Applications, Benefit to Organization

- Create a better, more accurate program with a better combination of data selection, text preprocessing, and ML model selection
- Change results so that it outputs most likely 5 companies with %
- Create script/UI that can be run by a company looking for tweets that are likely mentioning their company specifically
- Find tweets talking about their company when not mentioned with @, then can respond to help customers and improve reputation