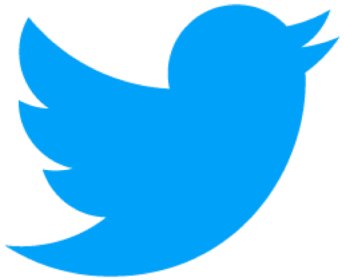


Twitter Big Data Project

By Rochita Sundar

1st May 2022

Objective

- Launch an AWS EC2 instance
- Gather 100K + tweets using Twitter Developer API. Scrape by keyword ‘Netflix’ or ‘netflix’
- Stream the tweets using kinesis firehose & store data in AWS S3 buckets
- Launch a databricks notebook, work in distributed environment & build a PySpark Sentiment Analysis ML model
- Write the sentiment predictions back in S3 buckets
- Run MySQL queries using AWS Athena & build an interactive visual dashboard using QuickSight to convey insights

Data

- 100,000+ tweets were collected over 10+ hours (on 27th & 28th April 2022). E.g.:

	tweet_id	user_name
1	1519451533603409921	kimi🍂
2	1519451539550744582	ella :) 🎗
3	1519451710497992708	Jaimie Buchanan #FBPA #FBPE 🇪🇺🇺🇦🇺🇦✊🏾✊🏾✊
4	1519450802884124674	Hank the Tank
5	1519451329508397058	FOX 42 KPTM
6	1519451395866316800	SofiaCarson

user_screen_name	user_followers_count	user_statuses_count	user_location
kimijaee	192	39337	☉♌ ☽♉ ↑♎ — cishet men dni
dvqella	116	2053	they/she/he
JJ_Enchanted	1579	11440	Edinburgh, Scotland
mirandabridgela	113	1696	Santa Claus, Arizona
FOX42KPTM	16655	69461	Omaha, NE
sofiacirsin	12	58	None

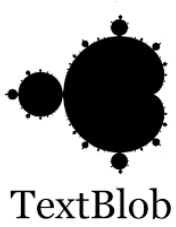
tweet_text
@joelocke03 @netflix which means they're welcome AGAIN which means season TWO
RT @joelocke03: @netflix I mean, you're 100% welcome...
RT @krishgm: Govt will publish plan to privatise Channel 4 tomorrow. Ministers say will 'unleash its potential' to compete with Netflix (wh...
RT @bingetowntv: Who else wishes @netflix would #savetheOA ? #TheOA https://t.co/9gw4Gl1qqH
It's nearly May and that means that it is time to take a look at what titles Netflix will be adding over the next f... https://t.co/PV6bSi4wXf
RT @ThePlaylistNews: First Look: 'PURPLE HEARTS.' Despite their differences and against all odds, an aspiring singer-songwriter (Sofia Cars...

tweet_hashtags
[]
[]
[]
[{'text': 'savetheOA', 'indices': [48, 58]}, {'text': 'TheOA', 'indices': [61, 67]}]
[]
[]

tweet_created_at
Wed Apr 27 23:00:50 +0000 2022
Wed Apr 27 23:00:51 +0000 2022
Wed Apr 27 23:01:32 +0000 2022
Wed Apr 27 22:57:56 +0000 2022
Wed Apr 27 23:00:01 +0000 2022
Wed Apr 27 23:00:17 +0000 2022

Data Labeling

- A lexicon based method (TextBlob) was used to label tweets as having a positive, neutral or negative sentiment

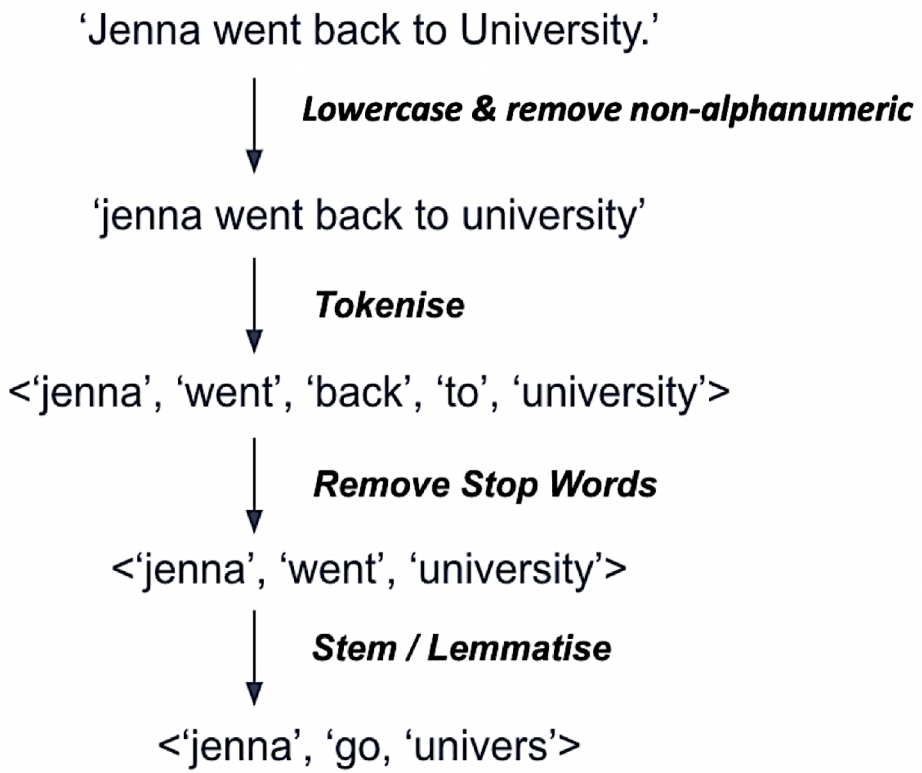


tweet_text	tweet_sentiment_label
@joelocke03 @netflix which means they're welcome AGAIN which means season TWO	positive
RT @joelocke03: @netflix I mean, you’re 100% welcome...	negative
RT @krishgm: Govt will publish plan to privatise Channel 4 tomorrow. Ministers say will 'unleash its potential' to compete with Netflix (wh...	neutral
RT @bingetowntv: Who else wishes @netflix would #savetheOA ? #TheOA https://t.co/9gw4Gl1qqH	neutral
It's nearly May and that means that it is time to take a look at what titles Netflix will be adding over the next f... https://t.co/PV6bSi4wXf	positive
RT @ThePlaylistNews: First Look: 'PURPLE HEARTS.' Despite their differences and against all odds, an aspiring singer-songwriter (Sofia Cars...	positive

Data Preprocessing

- Remove duplicates/ missing values
- Remove URLs
- Remove special characters
- Substituting multiple spaces with single space
- Lowercase all text
- Trim the leading/trailing whitespaces

- Tokenisation
- Stop word removal
- Count vectorizer & TF-IDF Vectorizer

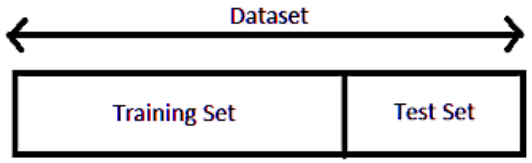


https://medium.com/predict/how-does-nlp-pre-processing-actually-work-8d097c179af1
https://realpython.com/sentiment-analysis-python/

Machine Learning – Logistic Regression

- Build a ML pipeline of steps

```
# Create transformers for the ML pipeline
tokenizer = Tokenizer(inputCol="tweet_text", outputCol="tokens")
stopword_remover = StopWordsRemover(inputCol="tokens", outputCol="filtered")
cv = CountVectorizer(vocabSize=2**16, inputCol="filtered", outputCol='cv')
idf = IDF(inputCol='cv', outputCol="1gram_idf", minDocFreq=5) #minDocFreq: remove sparse terms
assembler = VectorAssembler(inputCols=["1gram_idf"], outputCol="features")
label_encoder= StringIndexer(inputCol = "tweet_sentiment_label", outputCol = "label")
lr = LogisticRegression(maxIter=100)
pipeline = Pipeline(stages=[tokenizer, stopword_remover, cv, idf, assembler, label_encoder, lr])
```



- Split dataset into training & testing set

Fit pipeline on training set & evaluate performance on testing set

Tune model parameters such as 'N'-gram for TF-IDF vectorizers or maxiter for Logistic Regression

Build predictions using final chosen model

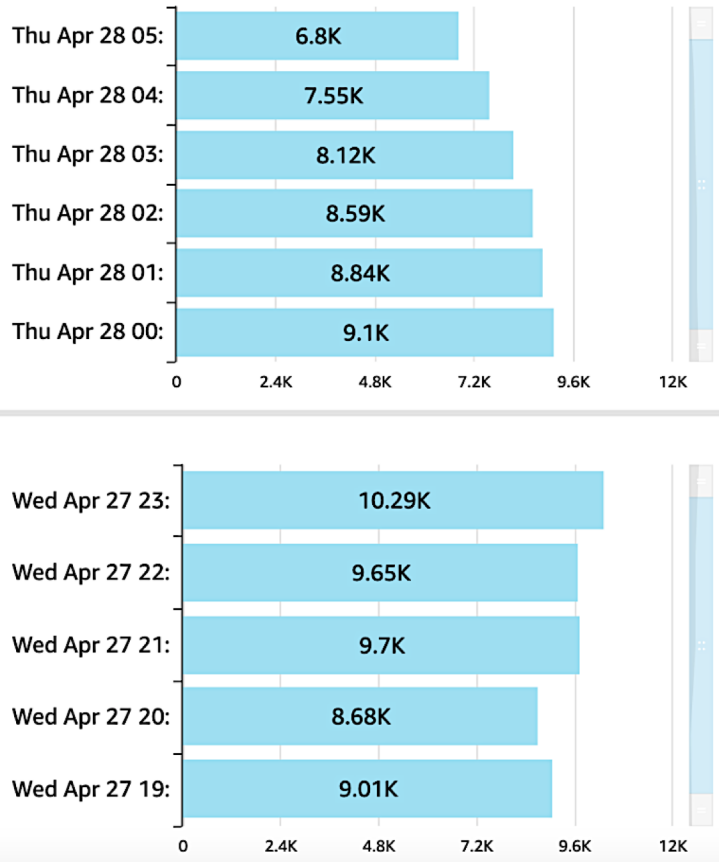
- Accuracy score of 0.9357 & ROC-AUC score of 0.9359 were obtained for the final chosen model

QuickSight Dashboard

- Interactive dashboard to filter by predicted sentiment as well as by right/wrong predictions

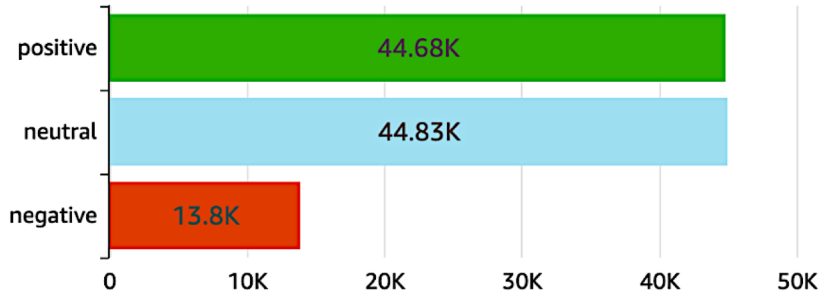
Collected 100,000+ tweets between Wednesday April 27th 7pm to Thursday April 28th 5am. Scrapped tweets & user information specifically for keyword 'Netflix' or 'netflix' in tweets.

Tweet count by day & hour



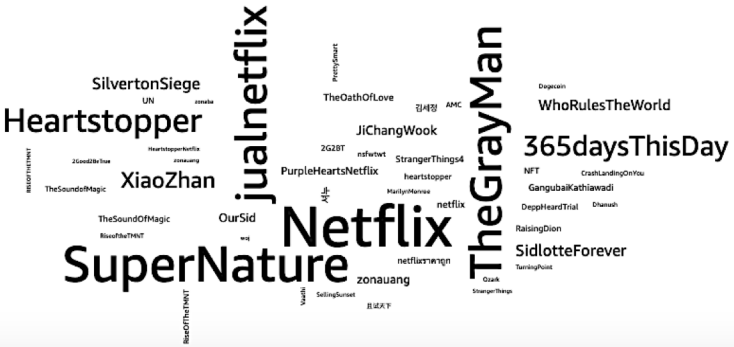
Count of Tweet Sentiments predicted by ML model

Select all
Right prediction
Wrong prediction



Word Cloud of # Firsthashtag in tweets by count

Showing top 50 (Size indicates count)



Word Cloud of Twitter user_names by follower_count

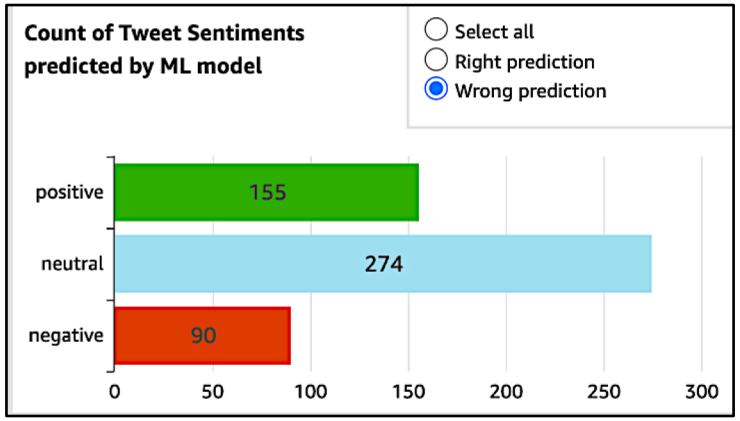
Showing top 50 user_names (Size indicates number of followers)

Select all
negative
neutral
positive

Word Cloud of Twitter user_names by statuses_count

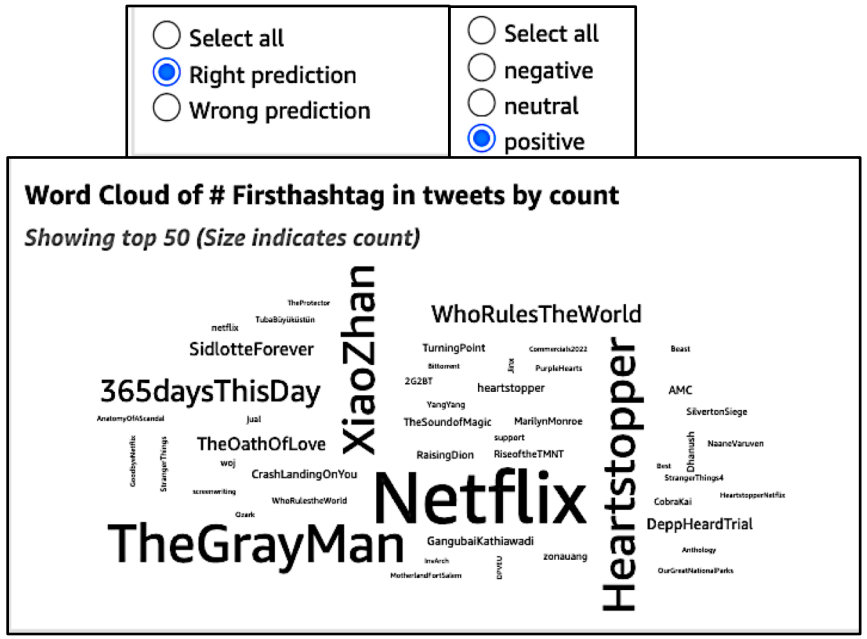
Showing top 30 user_names (Size indicates number of statuses)

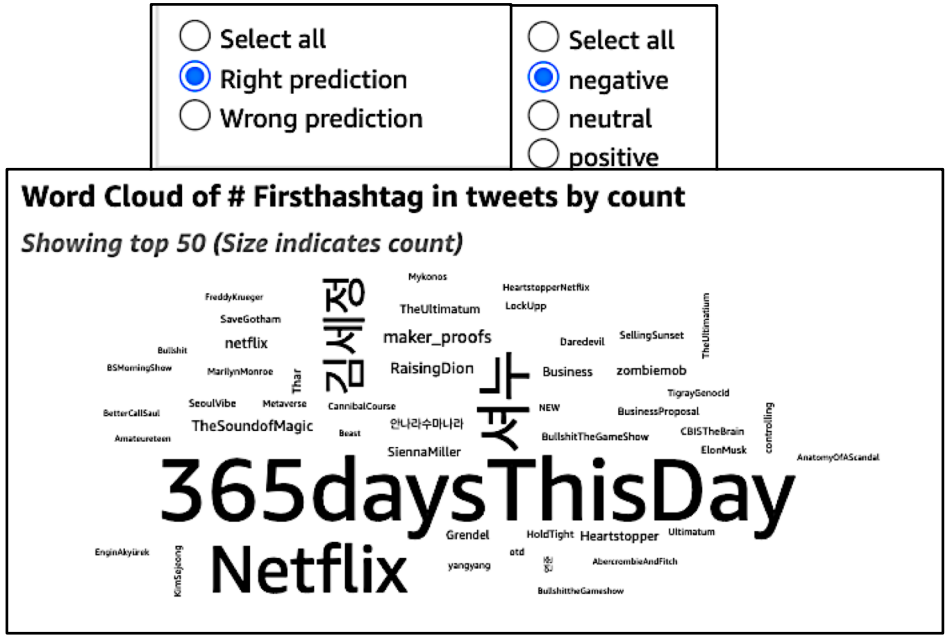
Dashboard Findings *(Some e.g.)*



- The model incorrectly predicts 155, 274 & 90 positive, neutral & negative sentiments

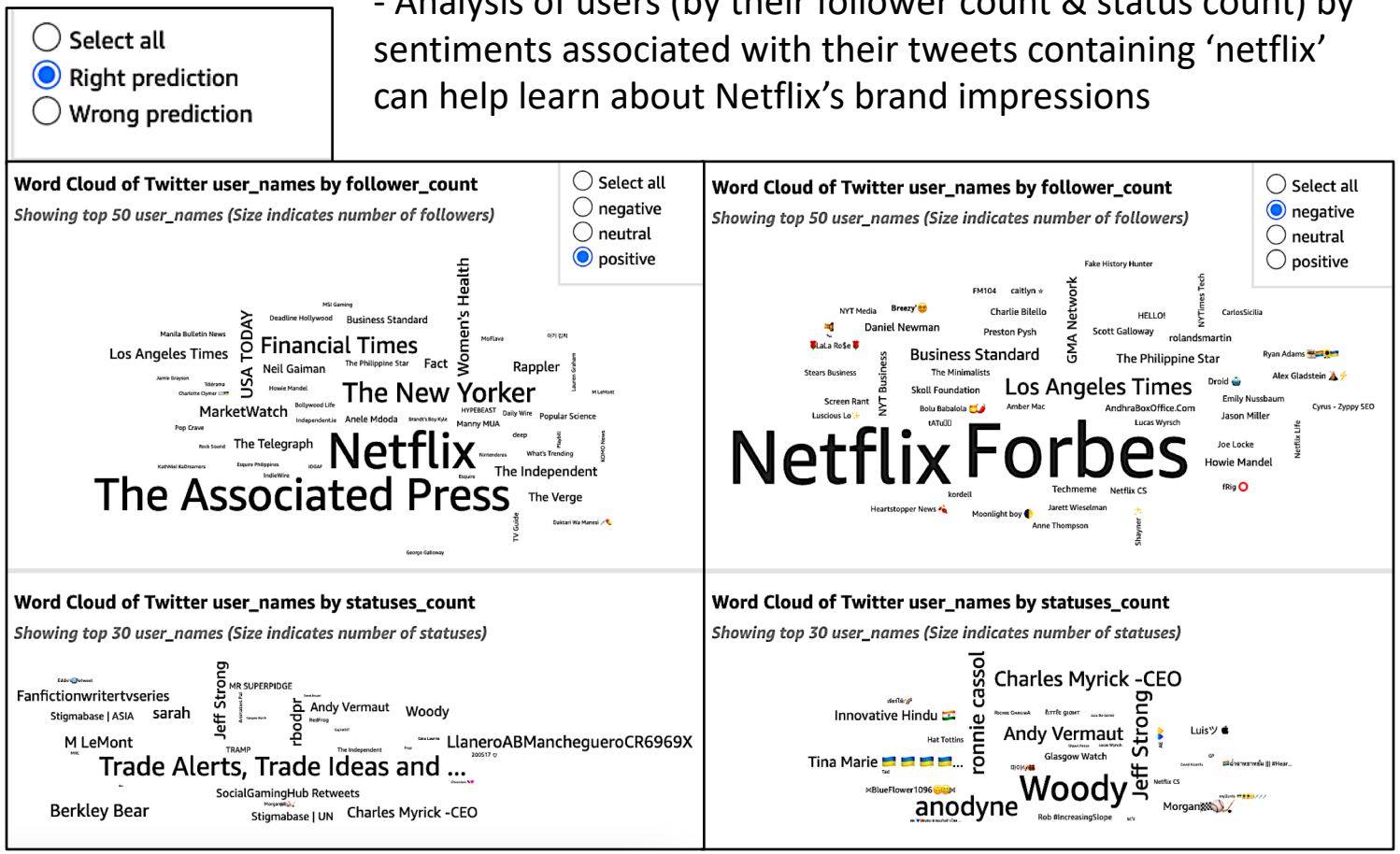
- Movie 365daysThisDay is associated with negative sentiments (perhaps bad reviews) while TheGrayMan & Heartstopper series are associated with positive sentiments (& good reviews). Chinese actor XiaoZhan is also associated with positive sentiments

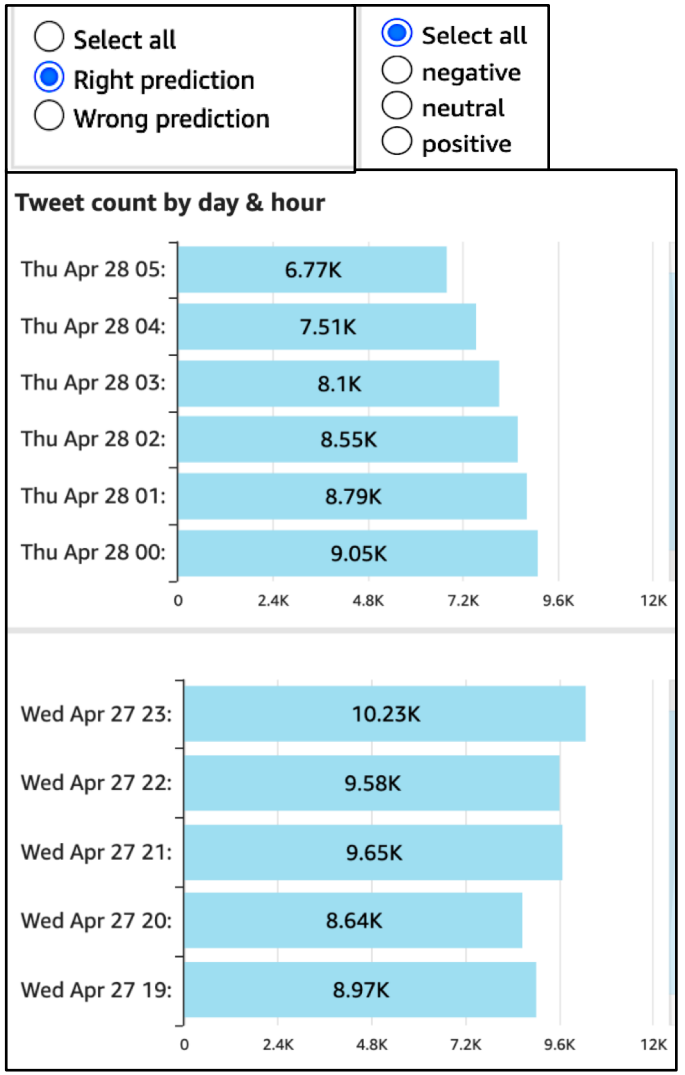




Dashboard Findings *(Some e.g.)*

- Analysis of users (by their follower count & status count) by sentiments associated with their tweets containing ‘netflix’ can help learn about Netflix’s brand impressions





- Maximum number of tweets containing ‘Netflix’ are closer to midnight than late evenings or early mornings

Challenges & Next Steps

- Some of the attributes collected were not useable as either data was not available like ‘geo’ or the data seemed incorrect like ‘location’

- Since only recent tweets were scrapped, reply_count, retweet_count & favorite_count were not useable

user_location
⊙♌ ☽♉ ↑♎ — cishet men dni
they/she/he
Edinburgh, Scotland
Santa Claus, Arizona
Omaha, NE
None

tweet_geo
None
None
None
None
None
None

tweet_reply_count	tweet_retweet_count	tweet_favorite_count
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0

- TextBlob lexicon based approach may not be the best way to assign initial sentiments to tweets as it negates the emotion/ logic behind the text & focuses only on count of positive or negative words
E.g., ‘bad’ → assigned negative sentiment

	I don't care how bad it is ryan hasn't been in a movie in four years so you KNOW I'm watching it	negative

- For next steps: project can be extended to build a real time analytical dashboard instead of static dashboard as well as scrape historical tweets rather than only new ones.
Other data labeling methods can be investigated to label the tweets more accurately