San Francisco State University

Twitter User Demographic Data Prediction

- BUS 895 Research Project in Business

Lu-Ting Lee
Rex Cheung
Leyla Ozsen

Agenda

- Introduction (4 minutes)
- Background (2 minutes)
- Research Process (20 minutes)
- Analysis Finding (3 minutes)
- Conclusion & Future works (3 minutes)
- Q & A

Introduction

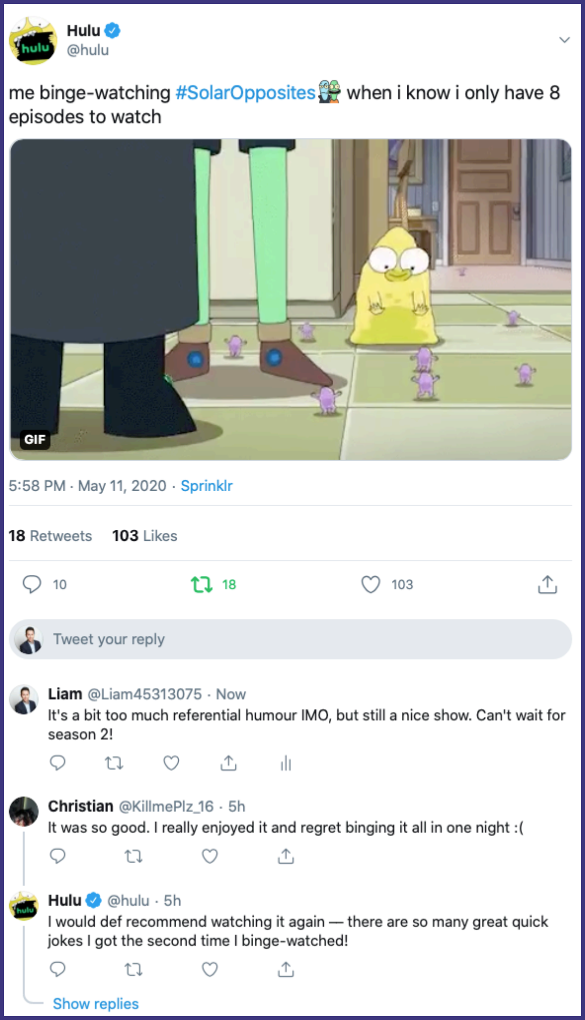
Introduction

What’s the goal?

“Predict Twitter users’ demographic data”

Why?

- **Better define the target audience**
- **Boost marketing campaign**
- **Demographic data: gender, age, location, education, income, ethnicity…**



Hulu @hulu

me binge-watching #SolarOpposites when i know i only have 8 episodes to watch

5:58 PM · May 11, 2020 · Sprinklr

18 Retweets 103 Likes

10 18 103

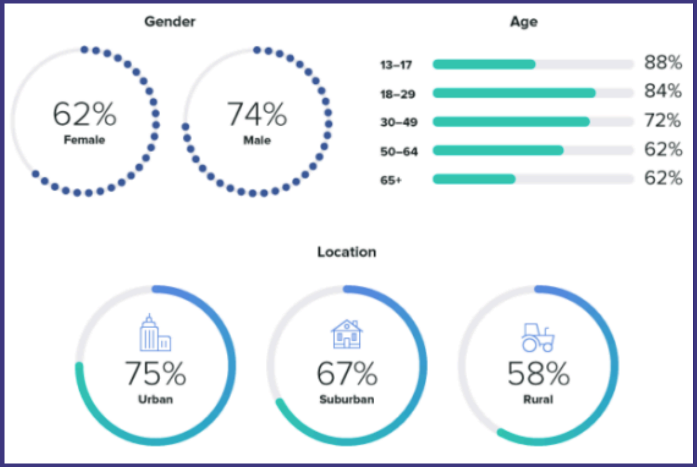
Tweet your reply

Liam @Liam45313075 · Now
It's a bit too much referential humour IMO, but still a nice show. Can't wait for season 2!

Christian @KillmePlz_16 · 5h
It was so good. I really enjoyed it and regret binging it all in one night :(

Hulu @hulu · 5h
I would def recommend watching it again — there are so many great quick jokes I got the second time I binge-watched!

Show replies



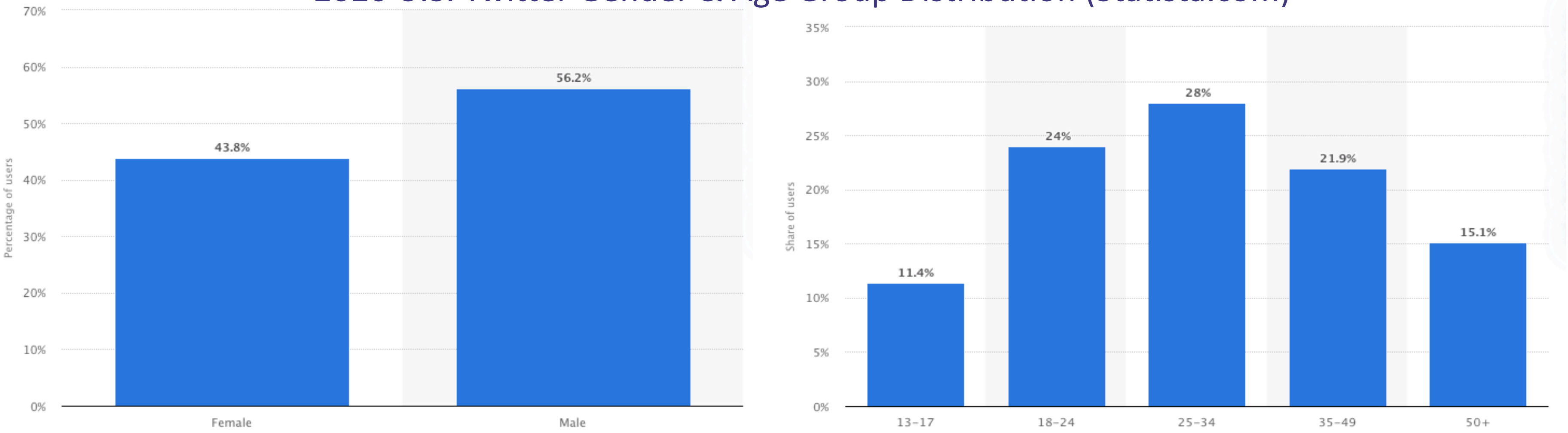
Credit by: https://www.pinterest.com/pin/437341813813402395

Introduction

○ **Scales:**

- Demographic data – Gender, Age Group, and Location
- USA domestic tweets *related to entertainment industry

2020 U.S. Twitter Gender & Age Group Distribution (Statista.com)



Introduction

○ **Exploration:**

- Data collection/ labeling
- Data analysis
- Text data feature extraction
- Prediction modeling

Background

Background

- **Intuition Intelligence**
  - Startup based in CA
  - Specialize on social media marketing / advertising
  - B2B software services
  - SaaS products

Research Process

1. Assumption
2. Data Collection/ Labeling
3. Modeling for Prediction

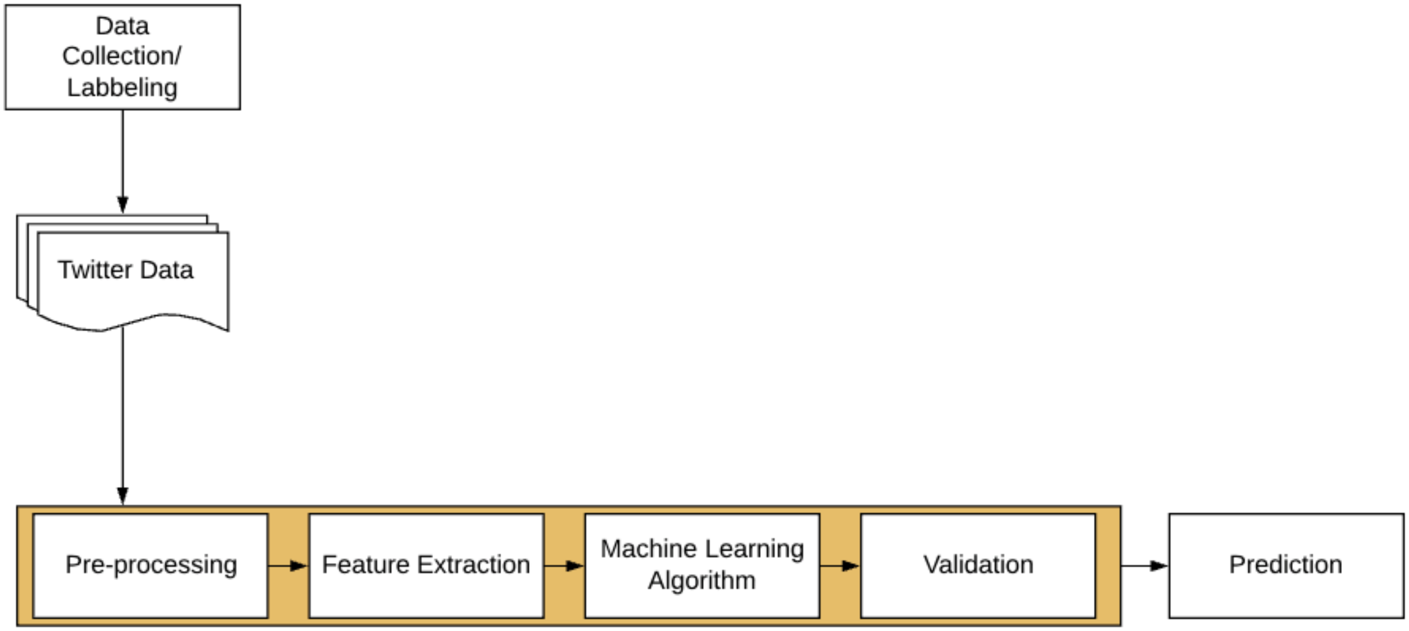
Assumption

- Twitter users’ writing style is different by gender or age group
- All Twitter users have an equal willingness to disclose demographic data in their Twitter account profile

Process Flow

2.

3. Modeling



Data Collection/ Labbeling

Twitter Data

Pre-processing

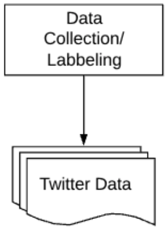
Feature Extraction

Machine Learning Algorithm

Validation

Prediction

Data Collection/ Labeling



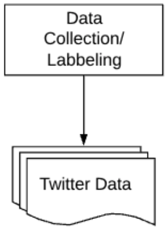
○ **Utilize Tweepy Library:**

- Third-party open source
- Extract data form Twitter Streaming API (application programing interface)

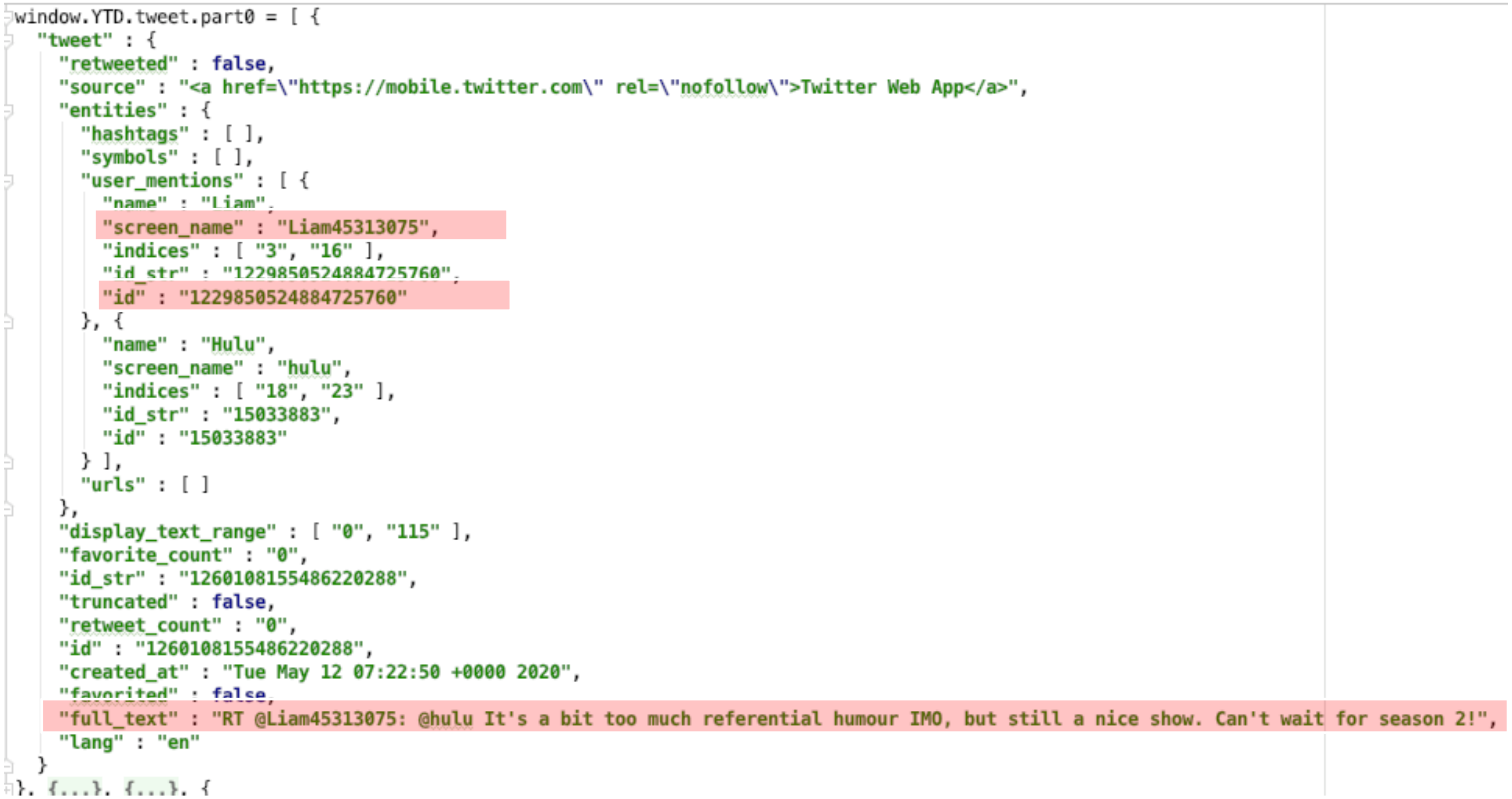
○ **Filter & Labeling:**

- U.S. only & English only
- Exclude Retweet
- Apply **labeling rules**

Data Collection

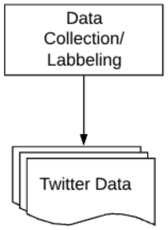


JS data from Twitter API



```
window.YTD.tweet.part0 = [ {
  "tweet" : {
    "retweeted" : false,
    "source" : "<a href=\"https://mobile.twitter.com\" rel=\"nofollow\">Twitter Web App</a>",
    "entities" : {
      "hashtags" : [ ],
      "symbols" : [ ],
      "user_mentions" : [ {
        "name" : "Liam",
        "screen_name" : "Liam45313075",
        "indices" : [ "3", "16" ],
        "id_str" : "1229850524884725760",
        "id" : "1229850524884725760"
      }, {
        "name" : "Hulu",
        "screen_name" : "hulu",
        "indices" : [ "18", "23" ],
        "id_str" : "15033883",
        "id" : "15033883"
      } ],
      "urls" : [ ]
    },
    "display_text_range" : [ "0", "115" ],
    "favorite_count" : "0",
    "id_str" : "1260108155486220288",
    "truncated" : false,
    "retweet_count" : "0",
    "id" : "1260108155486220288",
    "created_at" : "Tue May 12 07:22:50 +0000 2020",
    "favorited" : false,
    "full_text" : "RT @Liam45313075: @hulu It's a bit too much referential humour IMO, but still a nice show. Can't wait for season 2!",
    "lang" : "en"
  }
}, {...}, {...}, {
```

Data Labeling

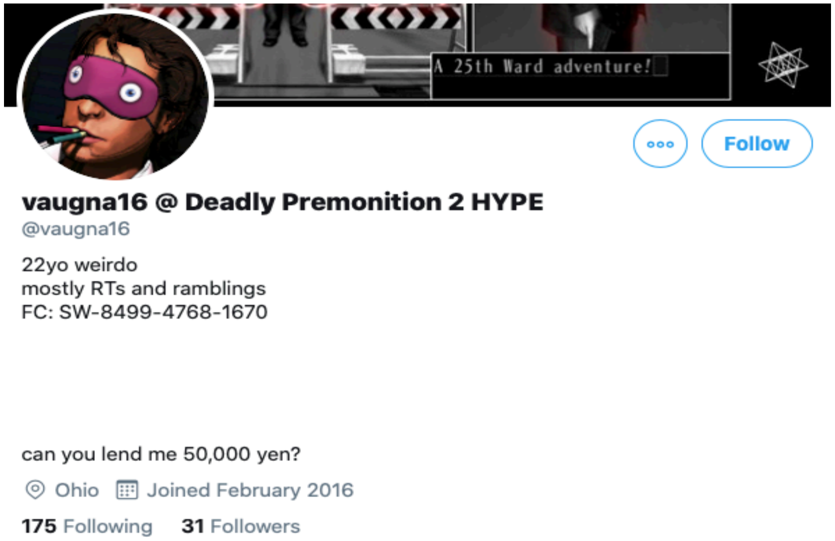


Data Collection/ Labbeling

Twitter Data

- **Labeling Rules – Examples**

  Description > Picture > Can’t define



vaugna16 @ Deadly Premonition 2 HYPE

@vaugna16

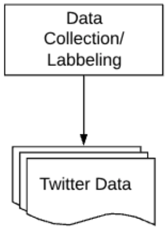
22yo weirdo
mostly RTs and ramblings
FC: SW-8499-4768-1670

can you lend me 50,000 yen?

Ohio   Joined February 2016

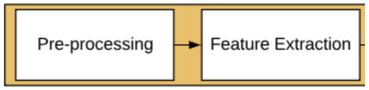
175 Following   **31** Followers

Data Collection/ Labeling



- **Result:**
  - Raw data from Twitter API ~ 2 million
  - 2,667 labeled observation x 14 features (~ 0.1% left)
    - 2,029 valid data for gender
    - 1,628 valid data for age group

Modeling- Pre-processing

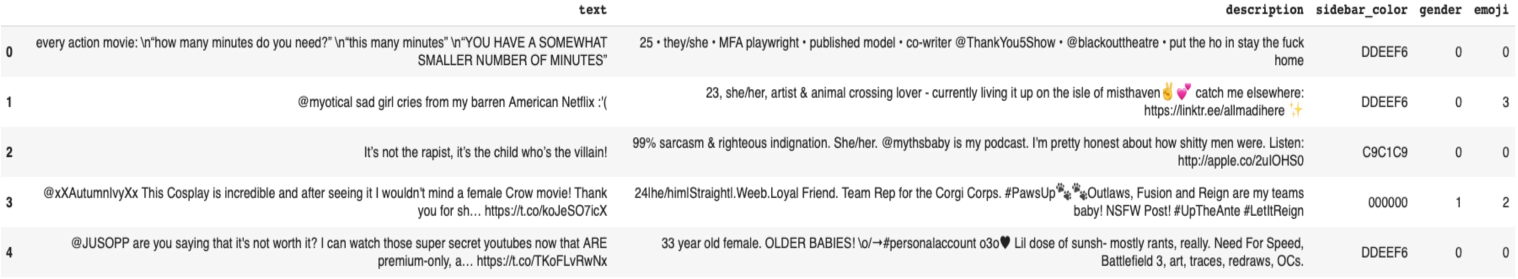


○ **Clean**

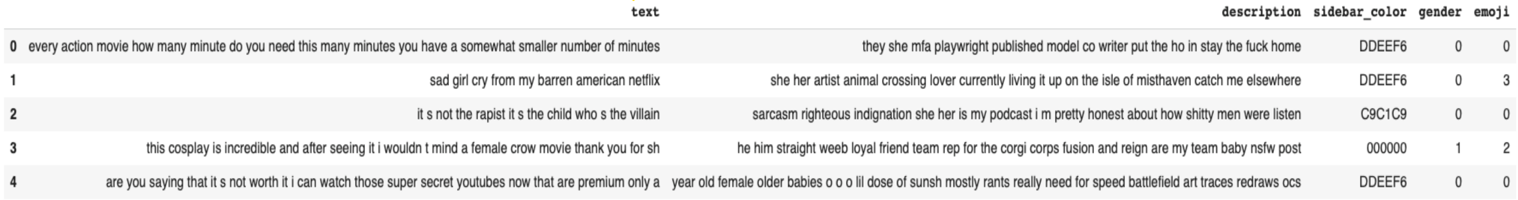
- Hashtag
- URL
- Emoji
- Non-alphabet
- Lower case
- ~~Stop-words~~

○ **Apply**

- Lemmatize

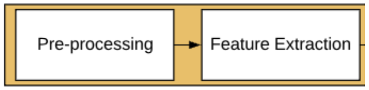


	text	description	sidebar_color	gender	emoji
0	every action movie: \n“how many minutes do you need?” \n“this many minutes” \n“YOU HAVE A SOMEWHAT SMALLER NUMBER OF MINUTES”	25 • they/she • MFA playwright • published model • co-writer @ThankYou5Show • @blackouttheatre • put the ho in stay the fuck home	DDEEF6	0	0
1	@myotical sad girl cries from my barren American Netflix :'(	23, she/her, artist & animal crossing lover - currently living it up on the isle of misthaven✌️💕 catch me elsewhere: https://linktr.ee/allmadihere ✨	DDEEF6	0	3
2	It’s not the rapist, it’s the child who’s the villain!	99% sarcasm & righteous indignation. She/her. @mythsbaby is my podcast. I'm pretty honest about how shitty men were. Listen: http://apple.co/2uIOHS0	C9C1C9	0	0
3	@xXAutumnIvyXx This Cosplay is incredible and after seeing it I wouldn't mind a female Crow movie! Thank you for sh… https://t.co/koJeSO7icX	24lhe/himlStraightl.Weeb.Loyal Friend. Team Rep for the Corgi Corps. #PawsUp🐾🐾Outlaws, Fusion and Reign are my teams baby! NSFW Post! #UpTheAnte #LetItReign	000000	1	2
4	@JUSOPP are you saying that it's not worth it? I can watch those super secret youtubes now that ARE premium-only, a… https://t.co/TKoFLvRwNx	33 year old female. OLDER BABIES! \o/→#personalaccount o3o🖤 Lil dose of sunsh- mostly rants, really. Need For Speed, Battlefield 3, art, traces, redraws, OCs.	DDEEF6	0	0



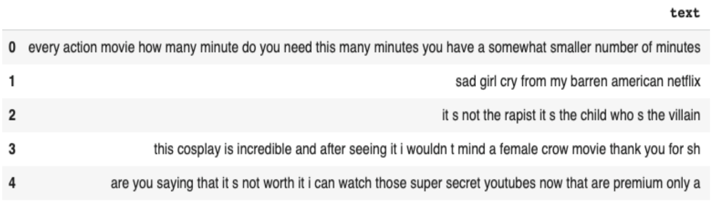
	text	description	sidebar_color	gender	emoji
0	every action movie how many minute do you need this many minutes you have a somewhat smaller number of minutes	they she mfa playwright published model co writer put the ho in stay the fuck home	DDEEF6	0	0
1	sad girl cry from my barren american netflix	she her artist animal crossing lover currently living it up on the isle of misthaven catch me elsewhere	DDEEF6	0	3
2	it s not the rapist it s the child who s the villain	sarcasm righteous indignation she her is my podcast i m pretty honest about how shitty men were listen	C9C1C9	0	0
3	this cosplay is incredible and after seeing it i wouldn t mind a female crow movie thank you for sh	he him straight weeb loyal friend team rep for the corgi corps fusion and reign are my team baby nsfw post	000000	1	2
4	are you saying that it s not worth it i can watch those super secret youtubes now that are premium only a	year old female older babies o o o lil dose of sunsh mostly rants really need for speed battlefield art traces redraws ocs	DDEEF6	0	0

Modeling- Feature Extraction



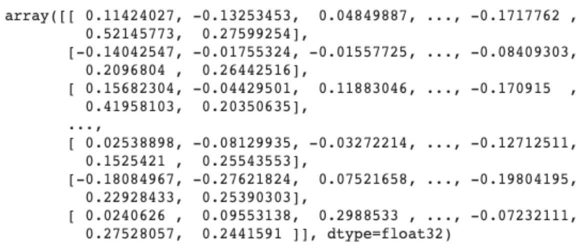
○ Utilize Distil-BERT for text data transforming:

- Tokenize
- Fit pre-trained deep neural network model
- Output equal dimension matrix



	text
0	every action movie how many minute do you need this many minutes you have a somewhat smaller number of minutes
1	sad girl cry from my barren american netflix
2	it s not the rapist it s the child who s the villain
3	this cosplay is incredible and after seeing it i wouldn t mind a female crow movie thank you for sh
4	are you saying that it s not worth it i can watch those super secret youtubes now that are premium only a

(908,)

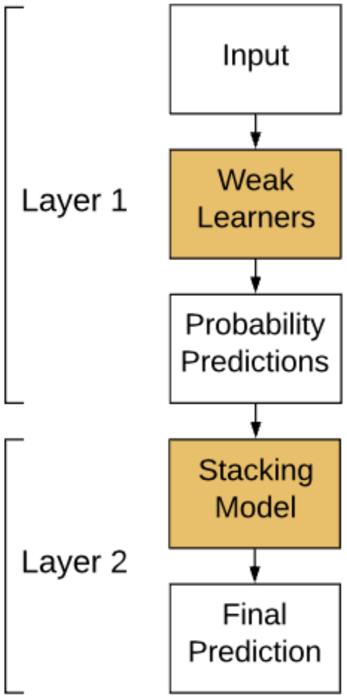


```
array([[ 0.11424027, -0.13253453,  0.04849887, ..., -0.1717762 ,
         0.52145773,  0.27599254],
       [-0.14042547, -0.01755324, -0.01557725, ..., -0.08409303,
         0.2096804 ,  0.26442516],
       [ 0.15682304, -0.04429501,  0.11883046, ..., -0.170915  ,
         0.41958103,  0.20350635],
       ...,
       [ 0.02538898, -0.08129935, -0.03272214, ..., -0.12712511,
         0.1525421 ,  0.25543553],
       [-0.18084967, -0.27621824,  0.07521658, ..., -0.19804195,
         0.22928433,  0.25390303],
       [ 0.0240626 ,  0.09553138,  0.2988533 , ..., -0.07232111,
         0.27528057,  0.2441591 ]], dtype=float32)
```

(908, 768)

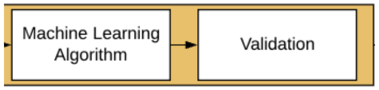
Modeling - ML Structure

- A 2-layers ensemble model with 8 weak learners

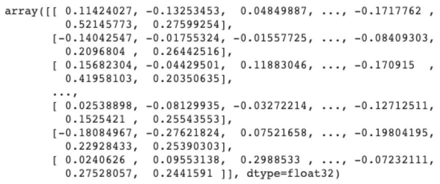


We fit the output from layer 1 to layer 2

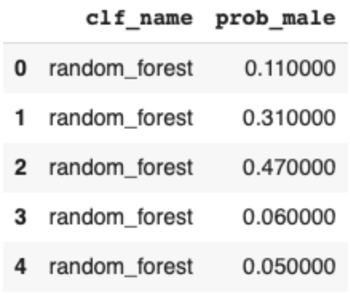
Modeling - ML Structure



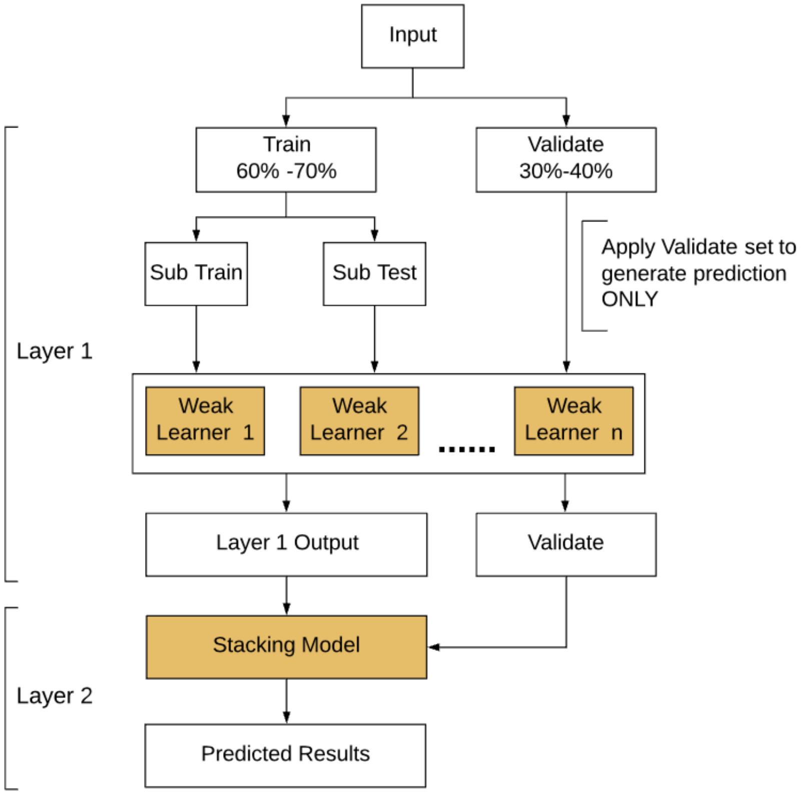
- Transform output matrix to probability after layer 1



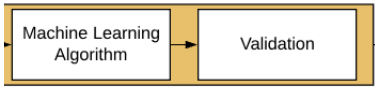
```
array([[ 0.11424027, -0.13253453,  0.04849887, ..., -0.1717762 ,
         0.52145773,  0.27599254],
       [-0.14042547, -0.01755324, -0.01557725, ..., -0.08409303,
         0.2096804 ,  0.26442516],
       [ 0.15682304, -0.04429501,  0.11883046, ..., -0.170915  ,
         0.41958103,  0.20350635],
       ...,
       [ 0.02538898, -0.08129935, -0.03272214, ..., -0.12712511,
         0.1525421 ,  0.25543553],
       [-0.18084967, -0.27621824,  0.07521658, ..., -0.19804195,
         0.22928433,  0.25390303],
       [ 0.0240626 ,  0.09553138,  0.2988533 , ..., -0.07232111,
         0.27528057,  0.2441591 ]], dtype=float32)
```



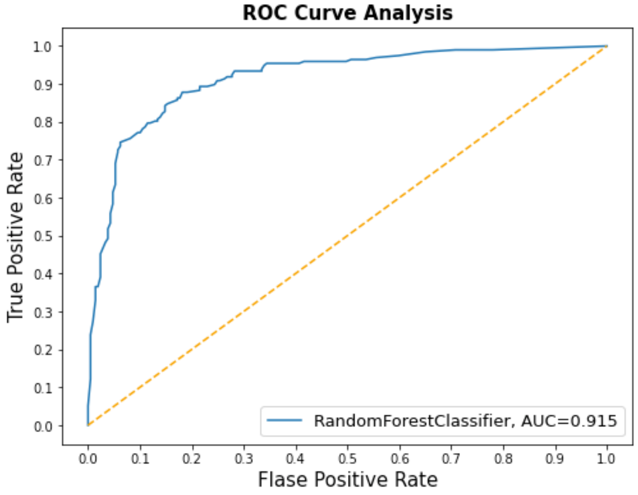
	clf_name	prob_male
0	random_forest	0.110000
1	random_forest	0.310000
2	random_forest	0.470000
3	random_forest	0.060000
4	random_forest	0.050000

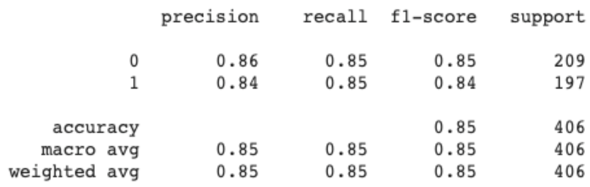


Modeling - Validation


Machine Learning Algorithm
Validation

- GridSearch + Cross-validation with validate set
- Get 0.84 Acc with RandomForestClassifier
- Define layer 2 ensemble classifier

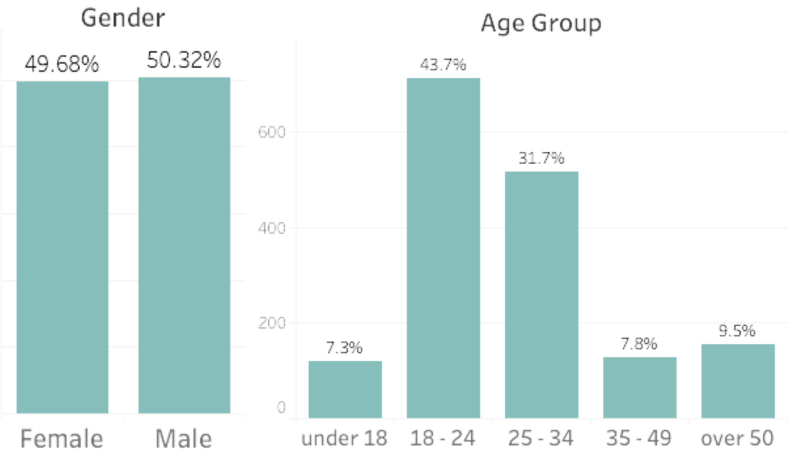



```
              precision    recall  f1-score   support

           0       0.86      0.85      0.85       209
           1       0.84      0.85      0.84       197

    accuracy                           0.85       406
   macro avg       0.85      0.85      0.85       406
weighted avg       0.85      0.85      0.85       406
```

Analysis Finding

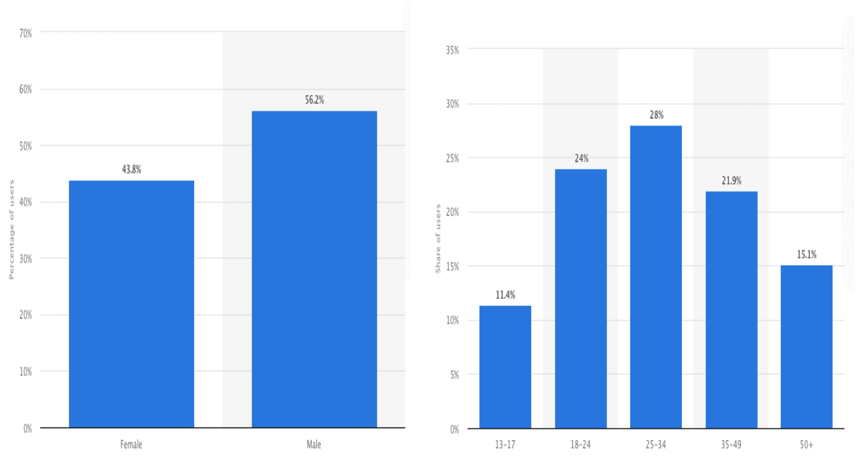
Analysis Finding

Our data distribution



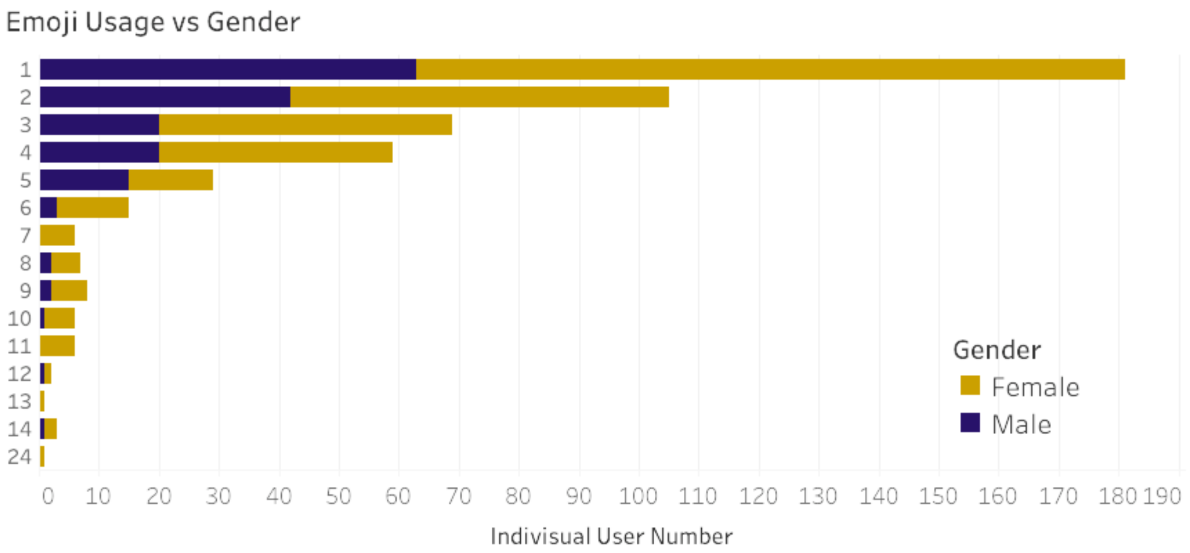
Total: 2029   Total: 1628

2020 U.S. Twitter Gender & Age Group Distribution (Statista.com)



Analysis Finding

- 36% data include at least one emoji
- Female use emoji more often



Conclusion & Future Works

Conclusion & Future Works

1. **Define Data Collection Methods**
2. **Define Labeling Rules**
3. **Establish a baseline model with 0.84 Acc & 0.92 AUC**

Future Works

1. **Improve labeling methods**
2. **Feature extraction**
    - Emoji
    - Sidebar color
3. **Utilize or mix with other approach**
    - Facial recognition
4. **Age prediction modeling**

Q & A

Reference

- Tweepy https://www.tweepy.org
- DistilBERT https://huggingface.co/transformers/model_doc/distilbert.html#Analys
- Lucidchart https://www.lucidchart.com
- Statista https://www.statista.com
- https://www.semanticscholar.org/paper/Why-Gender-and-Age-Prediction-from-Tweets-is-Hard%3A-Nguyen-Trieschnigg/ee06f058ca34a664d168f9a8f179d2db535c4e18
- https://www.sciencedirect.com/science/article/abs/pii/S0957417416301464
- https://www.researchgate.net/publication/221615645_Predicting_age_and_gender_in_online_social_networks
- https://www.researchgate.net/publication/221615645_Predicting_age_and_gender_in_online_social_networks

Thank you

liamlee9798@gmail.com