Predicting Age Group of Social Media Users

By Libby Merchant and Sydney Huxtable

Group 10

Business Problem

Implement a **supervised ML algorithm** to predict age groups of social media users based on individual language use.

The age segments available for global targeting are:

- 13-24
- 13-34
- 13-49
- 13-54
- 13+
- 18-24
- 18-34
- 18-49
- 18-54
- 18+
- 21-34
- 21-49
- 21-54
- 21+
- 25-49
- 25-54
- 25+
- 35-49
- 35-54
- 35+
- 50+

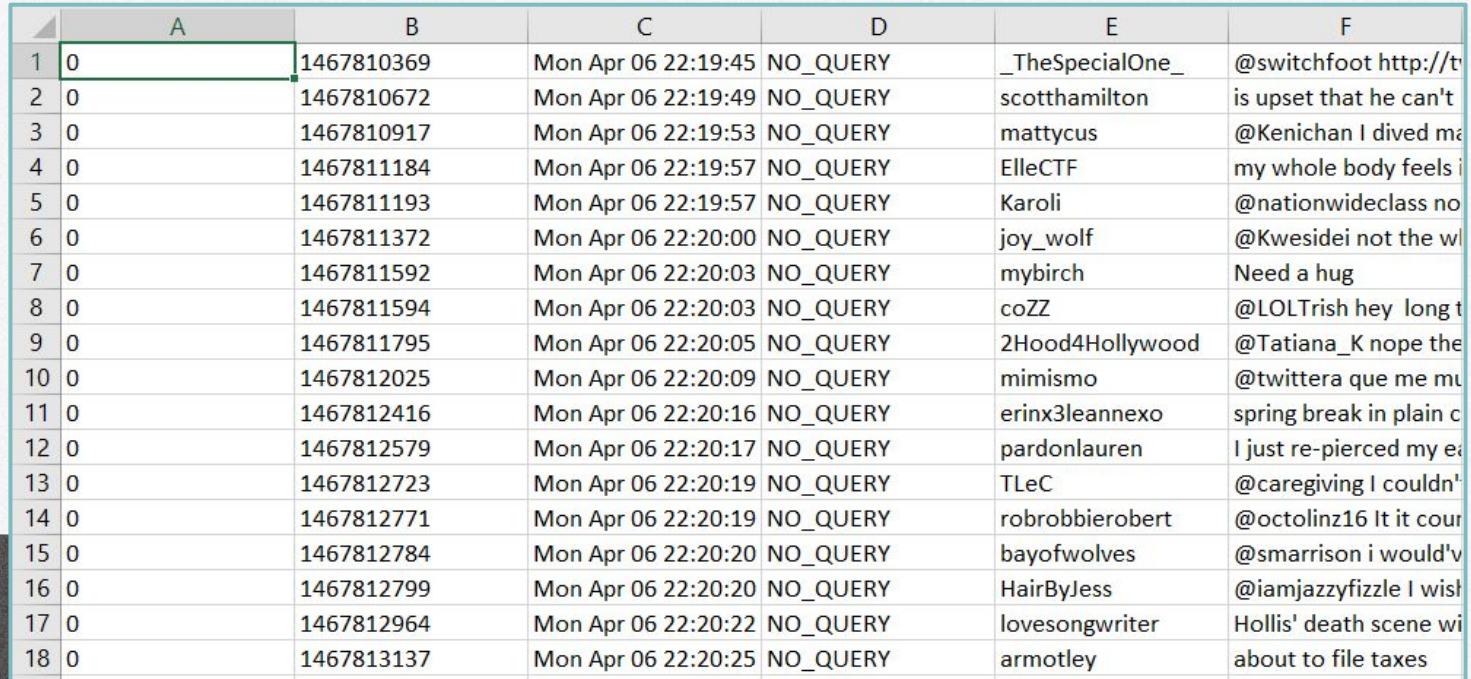
Celebrity Tweets Dataset

- GitHub repository of tweets from public figures such as Oprah, Kim K, and Bill Gates, used to analyze sentiment
- Who is the most positive figure based on Twitter activity?

348	BarackObama	Today is the one-year anniversary of the Clean Power Plan—show your support in the fight to #ActOnClimate.
349	BarackObama	The Clean Power Plan will cut carbon pollution from power plants by 32 percent by 2030. #ActOnClimate
350	BarackObama	Show President Obama some love on his 55th birthday—sign the birthday card.
351	BarackObama	Step 1: Sign President Obama's birthday card. Step 2: Celebrate.
352	BarackObama	Senate leaders are weakening our highest court by refusing to vote on Judge Garland. #DoYourJob
353	BarackObama	It has been a record-breaking 139 days since Judge Garland was nominated and still no hearing or vote. #DoYou
354	realDonaldTrump	Our relationship with Mexico is getting closer by the hour. Some really good people within both the new and old
355	realDonaldTrump	"The FBI looked at less than 1%" of Crooked's Emails!
356	realDonaldTrump	"The FBI only looked at 3000 of 675,000 Crooked Hillary Clinton Emails." They purposely didn't look at the disaste
357	realDonaldTrump	Big story out that the FBI ignored tens of thousands of Crooked Hillary Emails, many of which are REALLY BAD. Al
358	realDonaldTrump	.@LindseyGrahamSC "Every President deserves an Attorney General they have confidence in. I believe every Presi

Sentiment140 Twitter Dataset

- Sentiment140 dataset with 1.6 million tweets extracted using Twitter API
- Tweets have been annotated (0 = negative, 4 = positive) to analyze sentiment
- Columns: target, IDS, date, flag, user, text (Kaggle)



	A	B	C	D	E	F
1	0	1467810369	Mon Apr 06 22:19:45	NO_QUERY	_TheSpecialOne_	@switchfoot http://t
2	0	1467810672	Mon Apr 06 22:19:49	NO_QUERY	scotthamilton	is upset that he can't
3	0	1467810917	Mon Apr 06 22:19:53	NO_QUERY	mattycus	@Kenichan I dived ma
4	0	1467811184	Mon Apr 06 22:19:57	NO_QUERY	ElleCTF	my whole body feels i
5	0	1467811193	Mon Apr 06 22:19:57	NO_QUERY	Karoli	@nationwideclass no
6	0	1467811372	Mon Apr 06 22:20:00	NO_QUERY	joy_wolf	@Kwesidei not the wl
7	0	1467811592	Mon Apr 06 22:20:03	NO_QUERY	mybirch	Need a hug
8	0	1467811594	Mon Apr 06 22:20:03	NO_QUERY	coZZ	@LOLTrish hey long t
9	0	1467811795	Mon Apr 06 22:20:05	NO_QUERY	2Hood4Hollywood	@Tatiana_K nope the
10	0	1467812025	Mon Apr 06 22:20:09	NO_QUERY	mimismo	@twittera que me mu
11	0	1467812416	Mon Apr 06 22:20:16	NO_QUERY	erinx3leannexo	spring break in plain c
12	0	1467812579	Mon Apr 06 22:20:17	NO_QUERY	pardonlauren	I just re-pierced my ea
13	0	1467812723	Mon Apr 06 22:20:19	NO_QUERY	TLeC	@caregiving I couldn'
14	0	1467812771	Mon Apr 06 22:20:19	NO_QUERY	robrobbierobert	@octolinz16 It it cour
15	0	1467812784	Mon Apr 06 22:20:20	NO_QUERY	bayofwolves	@smarrison i would've
16	0	1467812799	Mon Apr 06 22:20:20	NO_QUERY	HairByJess	@iamjazzyfizzle I wish
17	0	1467812964	Mon Apr 06 22:20:22	NO_QUERY	lovesongwriter	Hollis' death scene wi
18	0	1467813137	Mon Apr 06 22:20:25	NO_QUERY	armotley	about to file taxes

Data Preprocessing

```
# merge all tweets for one user into a single string
tweet_str = ' '.join(user_twts)
print('Original tweets:', tweet_str)


# remove emoji
tweet_str = clean(tweet_str, no_emoji=True)
print(tweet_str)


# change uppercase letters to lowercase, ignoring emoticons
emoticons = [':D', 'XD', ':P', ';D', 'XP', ':D D', '=D', '=P']

save_emot = []
for i in emoticons:
    if i in tweet_str:
        save_emot.append(i)
        tweet_str = re.sub(i, '', tweet_str)  # remove emoticons with capital letters from tweets
```

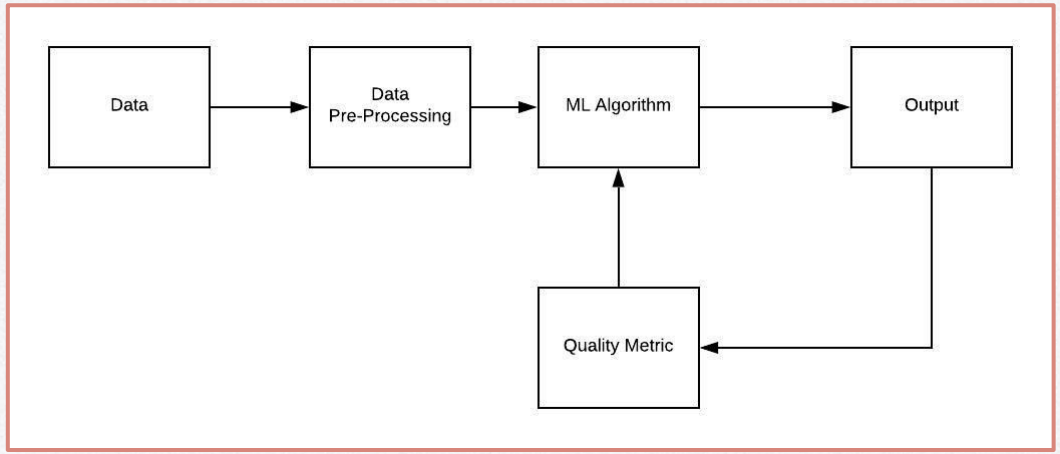
- Takes a separate csv file of tweets for each user
- Converts all characters to lowercase * keeping the case of emoticons
- Compiles all tweets by a user into a single string for analysis

```
# write tweets to csv file - CHANGE FILE
filename = 'KimKardashian_cleaned.txt'
with open(filename, 'w') as new_txt:
    new_txt.write(tweet_str)
```

- Text file produced for each user is now our input in the linear regression!

Linear Regression Model

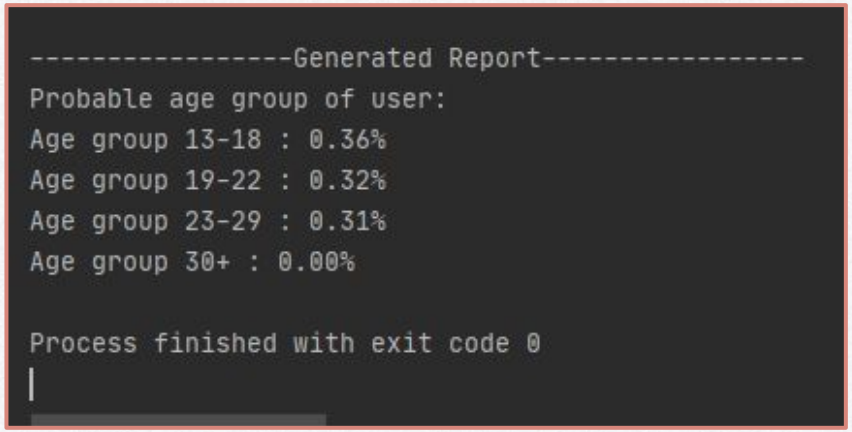
- Looks for words and phrases from the lexica
- Counts the number of occurrences for each and uses counts as the input for our regression
- *Developing Age and Gender Predictive Lexica over Social Media*



Linear Regression Model Cont.

Features:

- 4 different 200-dimensional matrices for 4 age brackets
- Make matrix with count of each word
- Linear regression: Matrix multiplication, sum resulting matrix
- Repeat for each age bracket
- Normalize outputs
- Calculate % match for each model



Key Results

- **Accuracy**
  - Tested model accuracy on celebrities
  - Information on age is readily available
- Produce % match for four age brackets
  - Correct guesses: Bill Gates (30+), Donald Trump (30+), Oprah (30+), Barack Obama (30+)
  - Incorrect guesses: Kim Kardashian (23-29)
  - Partially correct: Justin Bieber (19-22)
- Demonstrates one of the insidious ways social media giants use data

Learning Outcomes

Insights:

- Model tended to underestimate the age of users
  - Only 1 year for younger groups, average of ~20 years for older
  - ~4 years between publication of lexicon and data

Possible issues:

- Cross-platform applications
- Linguistic shifts (inverse correlation between emoticons & age)
- Differences in verified accounts vs private social media users
- Same words, different environments (ex: class vs. class vs. class)

Future opportunities:

- Replace current % match with Softmax to calculate probability
- Test on more users < age 30
- Compare with gender lexicon

Thank you!

Questions or Comments?

References

- World Well Being Project paper
- Predicting age groups of Twitter users based on language and metadata features (LDA algorithm)
- Predicting age and gender in online social networks (linear regression, emoticons!)
- Predicting age and gender in online social networks (linear regression, emoticons!)
- Age Prediction in Blogs: A Study of Style, Content, and Online Behavior in Pre- and Post-Social Media Generations (logistic regression)
- Prediction of Age, Sentiment, and Connectivity from Social Media Text | SpringerLink (EMD algorithm)
- "How Old Do You Think I Am?" A Study of Language and Age in Twitter | Proceedings of the International AAAI Conference on Web and Social Media (logistic and linear regression)
- A comparative evaluation of personality estimation algorithms for the twin recommender system | Proceedings of the 3rd international workshop on Search and mining user-generated contents (linear regression, M5' model tree, M5' regression tree and support vector machines for regression; TWIN rec system)
- SentiReview: Sentiment analysis based on text and emoticons (SVM, naive bayes, maximum entropy, neural networks)
- Determining the Age of the Author of the Text Based on Deep Neural Network Models