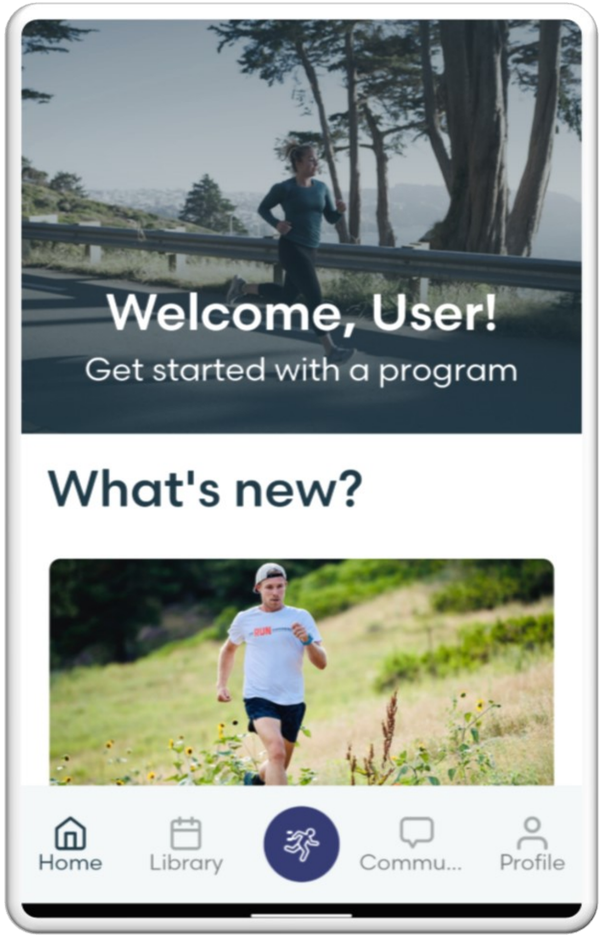
Beyond Words

Predicting user decision with text data

Eric Zhang

Insight 20C.DS.SV

An app for running exercise

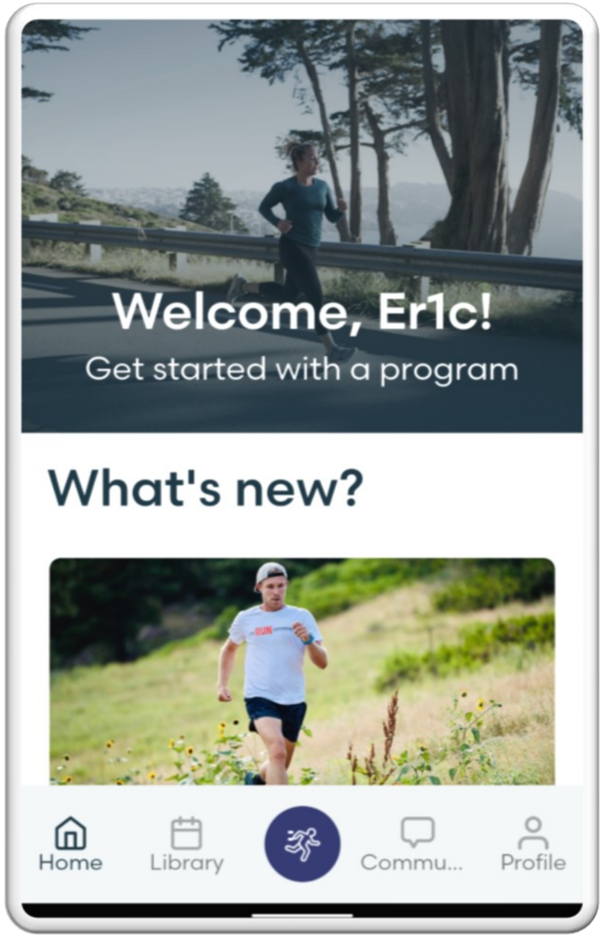


Welcome, User!

Get started with a program

What's new?

Imagine Eric uses this app for running exercise

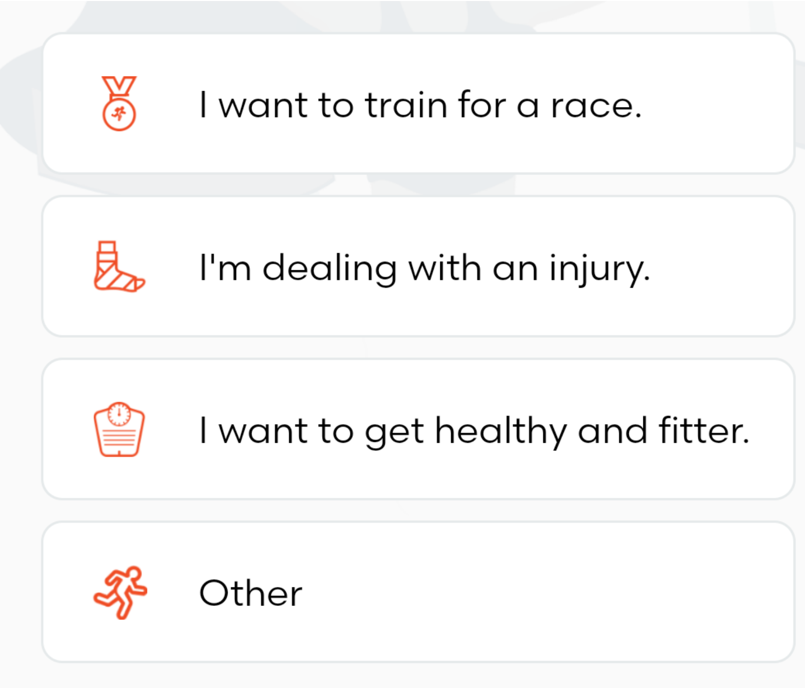


Welcome, Er1c!

Get started with a program

What's new?

Imagine Eric uses this app for running exercise

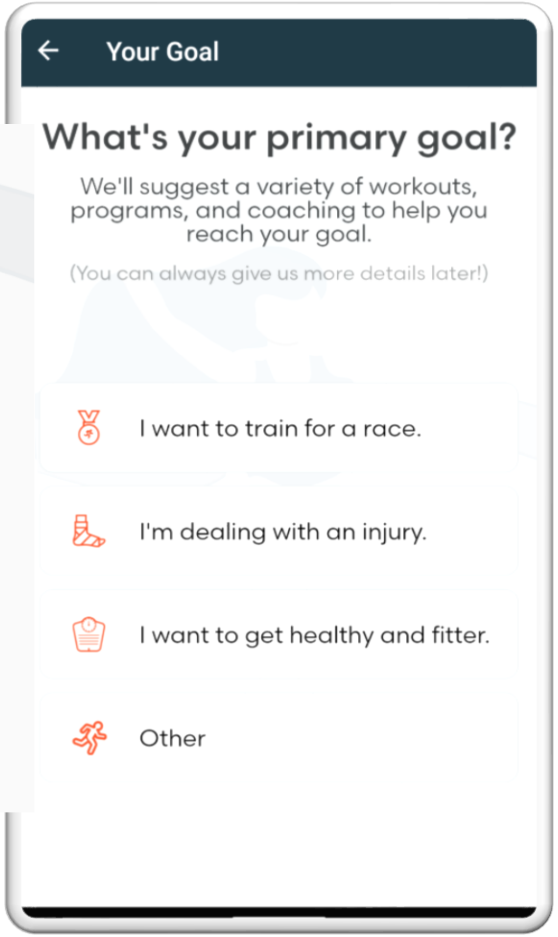


I want to train for a race.

I'm dealing with an injury.

I want to get healthy and fitter.

Other



What's your primary goal?

We'll suggest a variety of workouts, programs, and coaching to help you reach your goal.

(You can always give us more details later!)

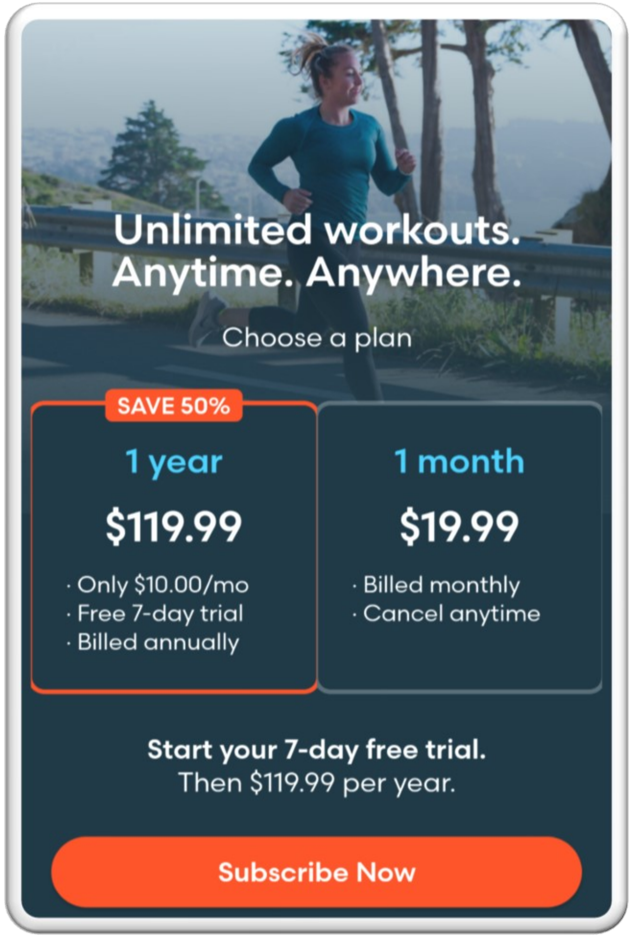
I want to train for a race.

I'm dealing with an injury.

I want to get healthy and fitter.

Other

Will Eric subscribe after the free trial



Unlimited workouts.
Anytime. Anywhere.

Choose a plan

1 year

$119.99

- Only $10.00/mo
- Free 7-day trial
- Billed annually

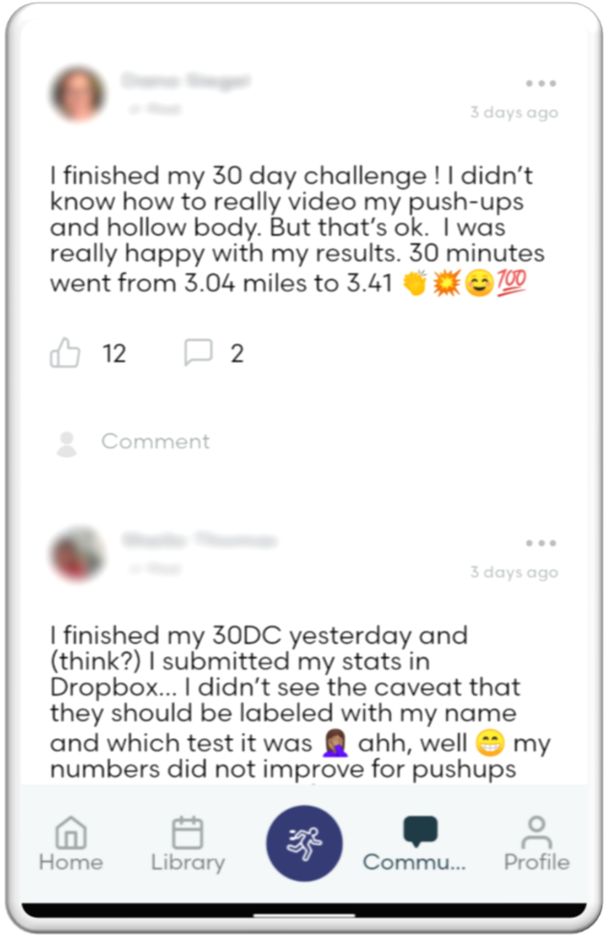
1 month

$19.99

- Billed monthly
- Cancel anytime

Start your 7-day free trial.
Then $119.99 per year.

Imagine if you have user text data during the free trial



[redacted]

3 days ago

I finished my 30 day challenge ! I didn't know how to really video my push-ups and hollow body. But that's ok. I was really happy with my results. 30 minutes went from 3.04 miles to 3.41 👏💥😊💯

👍 12 💬 2

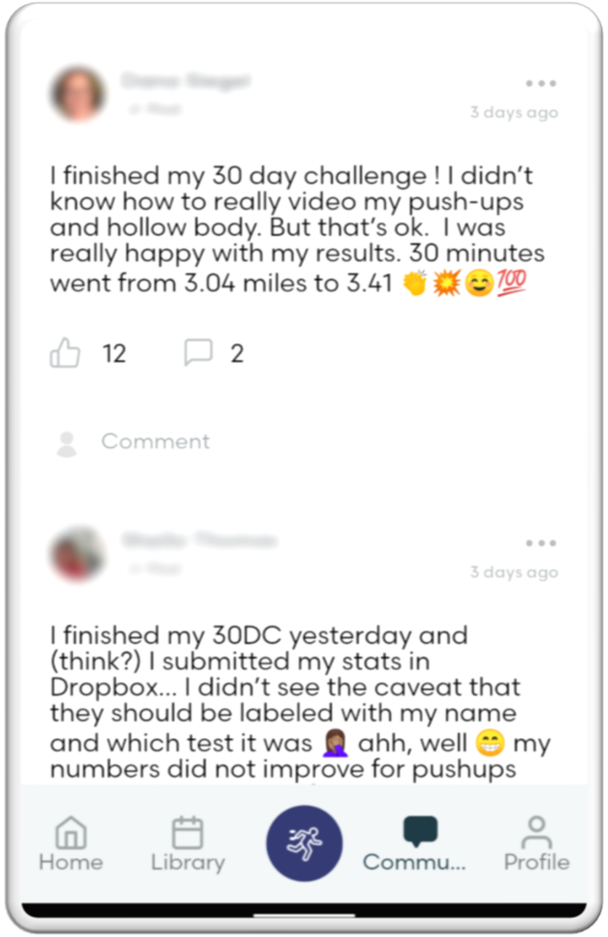
Comment

[redacted]

3 days ago

I finished my 30DC yesterday and (think?) I submitted my stats in Dropbox... I didn't see the caveat that they should be labeled with my name and which test it was 🤦‍♀️ ahh, well 😁 my numbers did not improve for pushups

What would you do as a data scientist



[redacted]

3 days ago

I finished my 30 day challenge ! I didn’t know how to really video my push-ups and hollow body. But that’s ok. I was really happy with my results. 30 minutes went from 3.04 miles to 3.41 👏💥😊💯

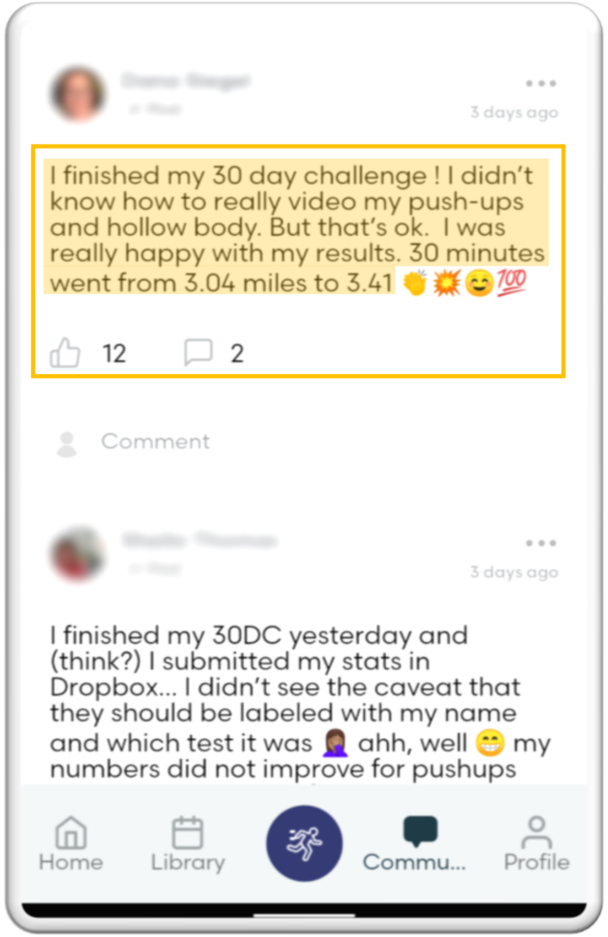
[redacted]

3 days ago

I finished my 30DC yesterday and (think?) I submitted my stats in Dropbox… I didn’t see the caveat that they should be labeled with my name and which test it was 🤦🏾‍♀️ ahh, well 😁 my numbers did not improve for pushups

What would you do as a data scientist

- Text:
  - Characters, words



I finished my 30 day challenge ! I didn't know how to really video my push-ups and hollow body. But that's ok. I was really happy with my results. 30 minutes went from 3.04 miles to 3.41 👏💥😊💯

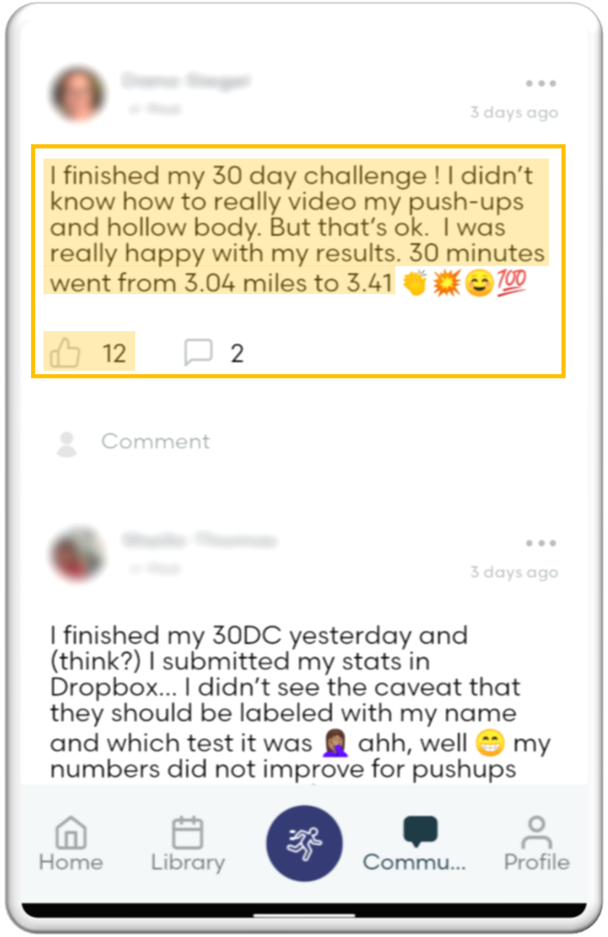
I finished my 30DC yesterday and (think?) I submitted my stats in Dropbox... I didn't see the caveat that they should be labeled with my name and which test it was 🤦‍♀️ ahh, well 😁 my numbers did not improve for pushups

Total characters: 150

Total words: 36

What would you do as a data scientist

- Text:
  - Characters, words
  - Likes



3 days ago

I finished my 30 day challenge ! I didn't know how to really video my push-ups and hollow body. But that's ok. I was really happy with my results. 30 minutes went from 3.04 miles to 3.41 👏💥😊💯

12 2

Comment

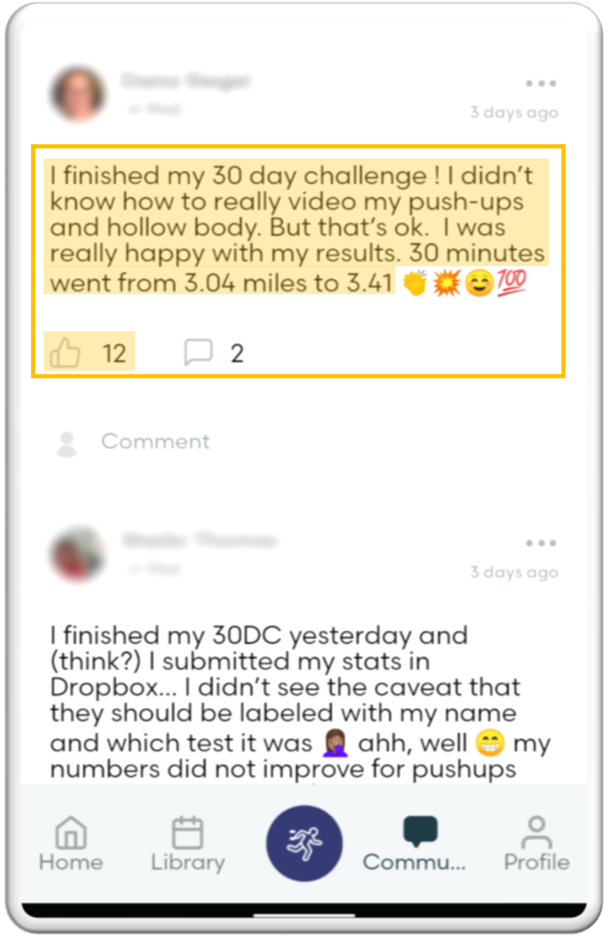
3 days ago

I finished my 30DC yesterday and (think?) I submitted my stats in Dropbox... I didn't see the caveat that they should be labeled with my name and which test it was 🤦‍♀️ ahh, well 😁 my numbers did not improve for pushups

Total characters: 150
Total words: 36
Total likes: 12

What would you do as a data scientist

- Text:
  - Characters, words
  - Likes
- Sentiment:



3 days ago

I finished my 30 day challenge ! I didn't know how to really video my push-ups and hollow body. But that's ok. I was really happy with my results. 30 minutes went from 3.04 miles to 3.41 👏💥😄💯

12   2

Comment

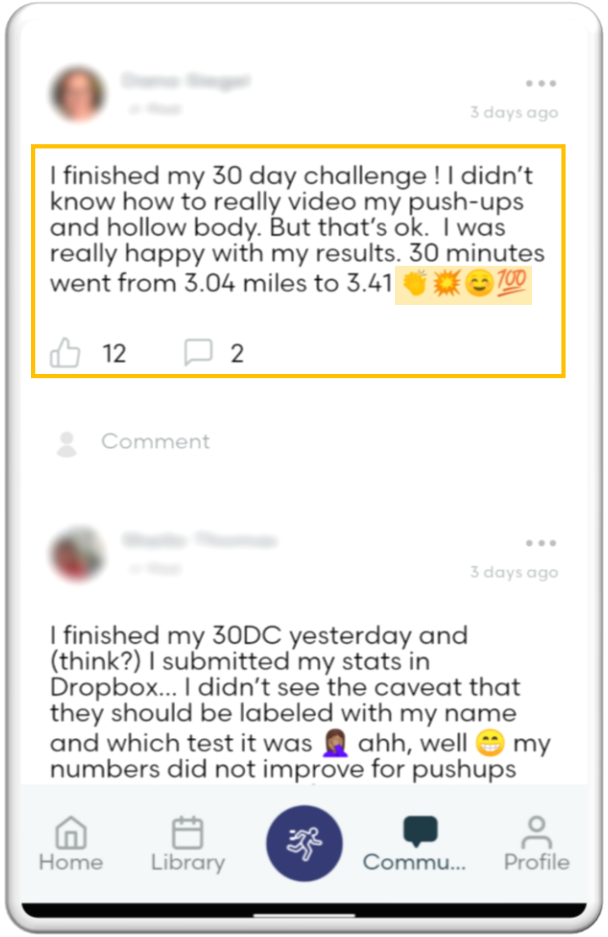
3 days ago

I finished my 30DC yesterday and (think?) I submitted my stats in Dropbox… I didn't see the caveat that they should be labeled with my name and which test it was 🤦🏽‍♀️ ahh, well 😁 my numbers did not improve for pushups

Total characters: 150
Total words: 36
Total likes: 12

What would you do as a data scientist

- Text:
  - Characters, words
  - Likes
- Sentiment:
  - Text and Emoji



3 days ago

I finished my 30 day challenge ! I didn't know how to really video my push-ups and hollow body. But that's ok. I was really happy with my results. 30 minutes went from 3.04 miles to 3.41 👏💥😊💯

👍 12 💬 2

Comment

3 days ago

I finished my 30DC yesterday and (think?) I submitted my stats in Dropbox… I didn't see the caveat that they should be labeled with my name and which test it was 🤦‍♀️ ahh, well 😁 my numbers did not improve for pushups

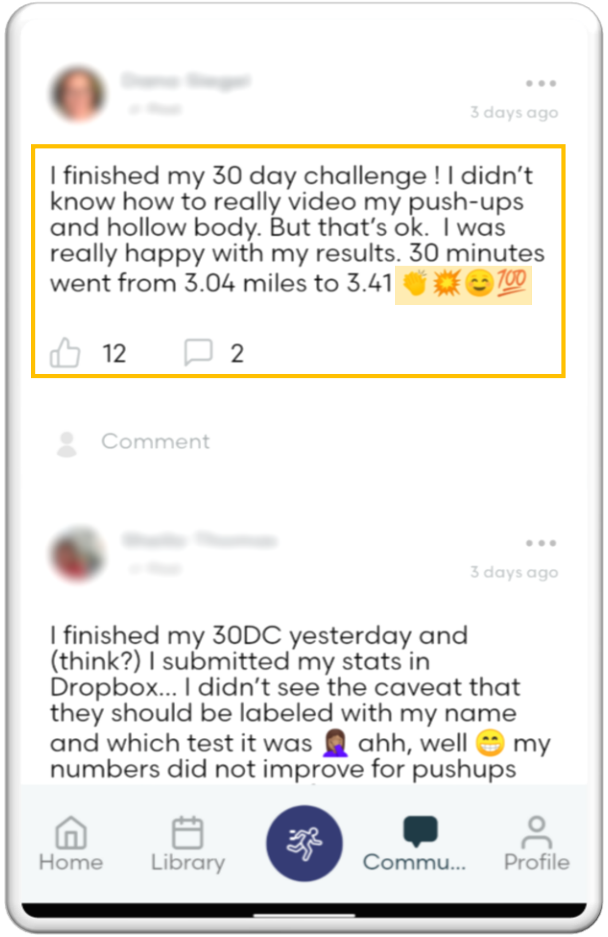
Total characters: 150
Total words: 36
Total likes: 12

Emojis:

👏
💥
😊
💯

What would you do as a data scientist

- Text:
  - Characters, words
  - Likes
- Sentiment:
  - Text and Emoji



3 days ago

I finished my 30 day challenge ! I didn't know how to really video my push-ups and hollow body. But that's ok. I was really happy with my results. 30 minutes went from 3.04 miles to 3.41 👏💥😊💯

👍 12 💬 2

Comment

3 days ago

I finished my 30DC yesterday and (think?) I submitted my stats in Dropbox... I didn't see the caveat that they should be labeled with my name and which test it was 🤦 ahh, well 😁 my numbers did not improve for pushups

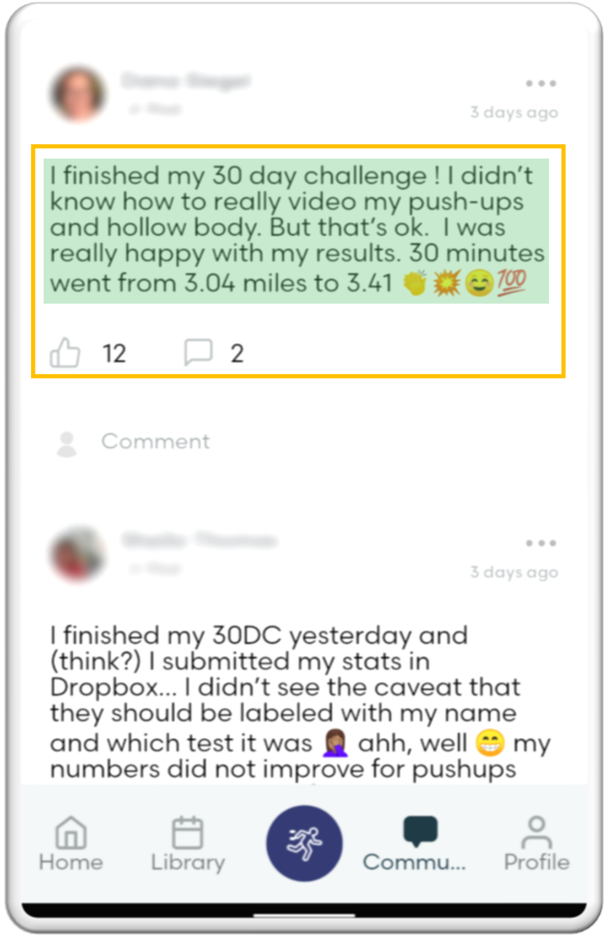
Total characters: 150
Total words: 36
Total likes: 12

Emojis:

👏 >>> Clapping Hands
💥 >>> Collision
😊 >>> Smiley Face
💯 >>> Hundred Points

What would you do as a data scientist

- Text:
  - Characters, words
  - Likes
- Sentiment:
  - Text and Emoji
  - Tone
  - Factual



3 days ago

I finished my 30 day challenge ! I didn't know how to really video my push-ups and hollow body. But that's ok. I was really happy with my results. 30 minutes went from 3.04 miles to 3.41 👏💥😊💯

👍 12 💬 2

Comment

3 days ago

I finished my 30DC yesterday and (think?) I submitted my stats in Dropbox... I didn't see the caveat that they should be labeled with my name and which test it was 🤦🏾‍♀️ ahh, well 😁 my numbers did not improve for pushups

Total characters: 150
Total words: 36
Total likes: 12

Emojis:

👏 >>> Clapping Hands
💥 >>> Collision
😊 >>> Smiley Face
💯 >>> Hundred Points

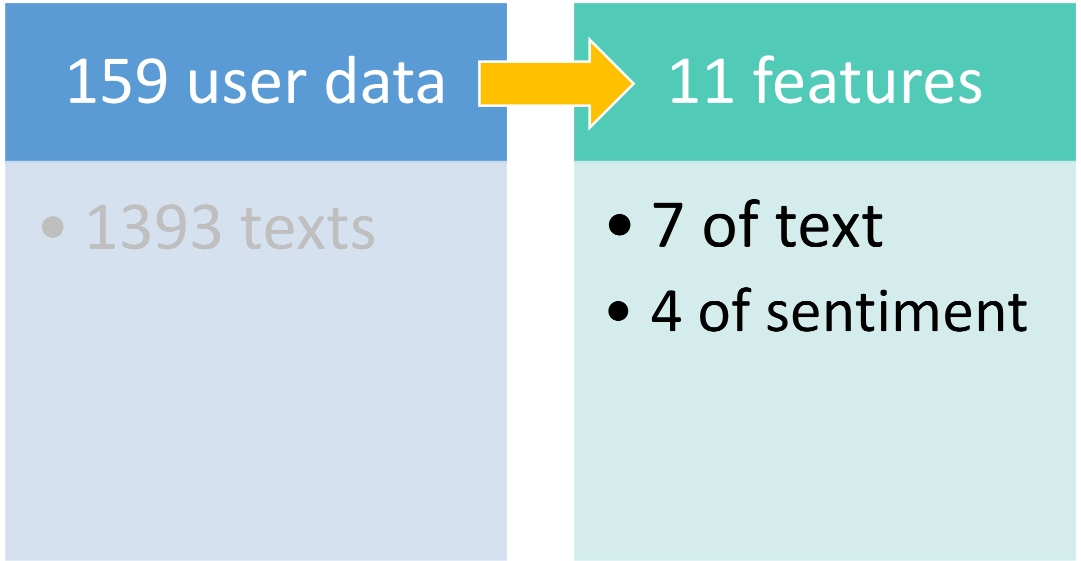
Sentiment:
Tone: Positive (1)
Factual: Rich (1)

Processing data with machine learning models

159 user data

- 1393 texts

Processing data with machine learning models



159 user data

- 1393 texts

11 features

- 7 of text
- 4 of sentiment

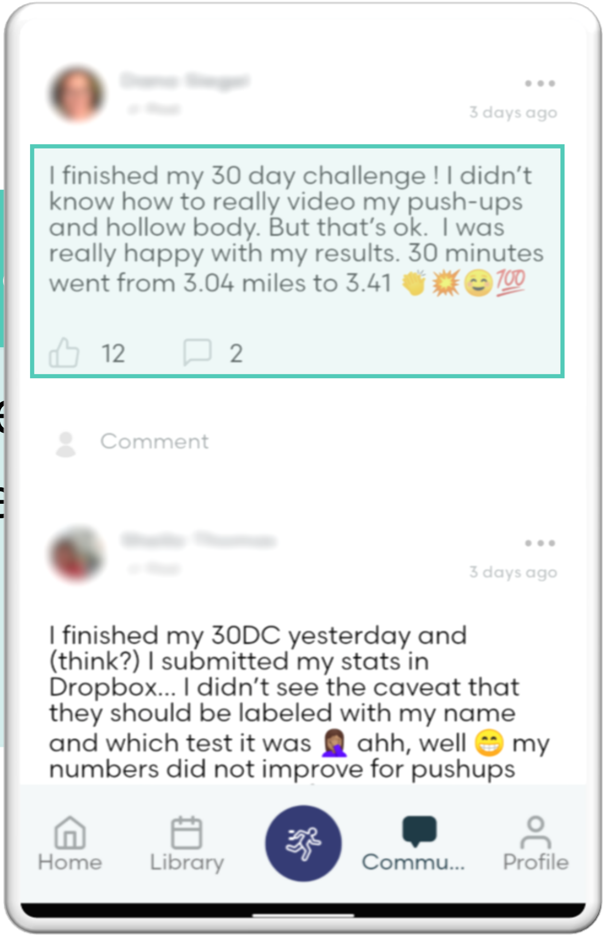
Processing data with machine learning models

159 user data

- 1393 texts

11 fe

- 7 of te
- 4 of se



[redacted]

[redacted]

3 days ago

I finished my 30 day challenge ! I didn’t know how to really video my push-ups and hollow body. But that’s ok. I was really happy with my results. 30 minutes went from 3.04 miles to 3.41 👏💥😊💯

12 2

Comment

[redacted]

[redacted]

3 days ago

I finished my 30DC yesterday and (think?) I submitted my stats in Dropbox... I didn’t see the caveat that they should be labeled with my name and which test it was 🤦 ahh, well 😁 my numbers did not improve for pushups

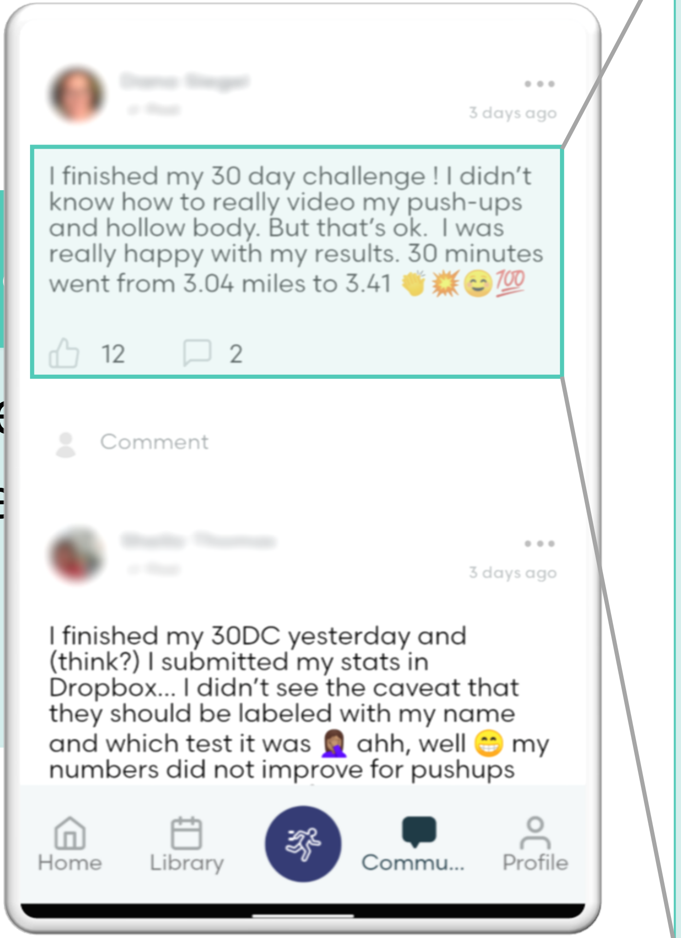
Processing data with machine learning models

159 user data

- 1393 texts

11 fe

- 7 of te
- 4 of se



3 days ago

I finished my 30 day challenge ! I didn't know how to really video my push-ups and hollow body. But that's ok. I was really happy with my results. 30 minutes went from 3.04 miles to 3.41 👏💥😄💯

12 2

Comment

3 days ago

I finished my 30DC yesterday and (think?) I submitted my stats in Dropbox... I didn't see the caveat that they should be labeled with my name and which test it was 🤦‍♀️ ahh, well 😁 my numbers did not improve for pushups

[7 Text Features]

- Total characters
- Total words
- Average character
- Average words
- Number of posts
- Number of likes
- Average likes

[4 Sentiment features]

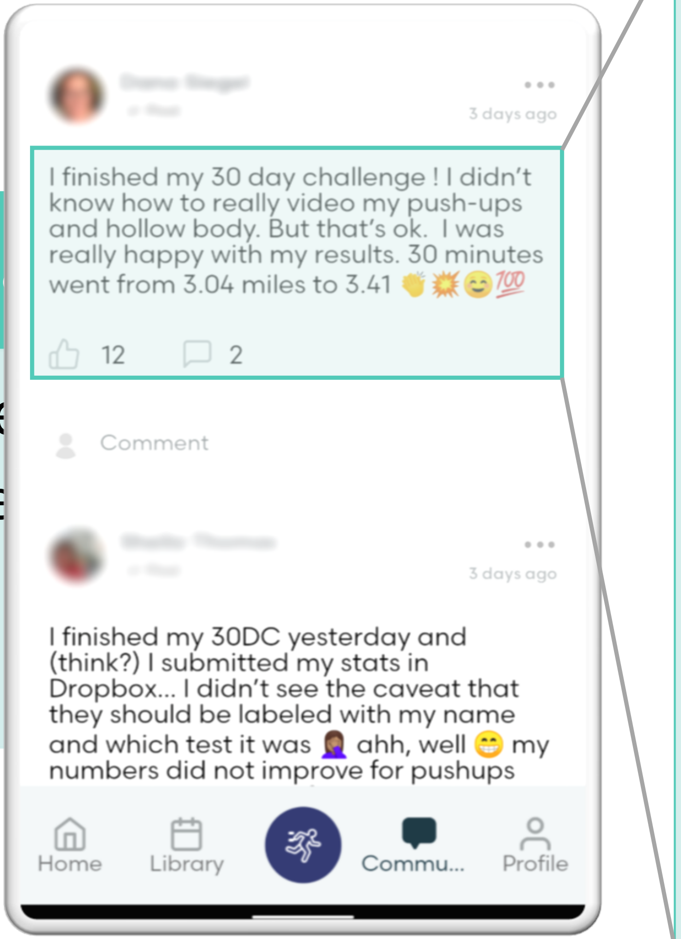
Processing data with machine learning models

159 user data

- 1393 texts

11 fe

- 7 of te
- 4 of se



3 days ago

I finished my 30 day challenge ! I didn't know how to really video my push-ups and hollow body. But that's ok. I was really happy with my results. 30 minutes went from 3.04 miles to 3.41 👏💥😄💯

12 2

Comment

3 days ago

I finished my 30DC yesterday and (think?) I submitted my stats in Dropbox... I didn't see the caveat that they should be labeled with my name and which test it was 🤦🏾‍♀️ ahh, well 😁 my numbers did not improve for pushups

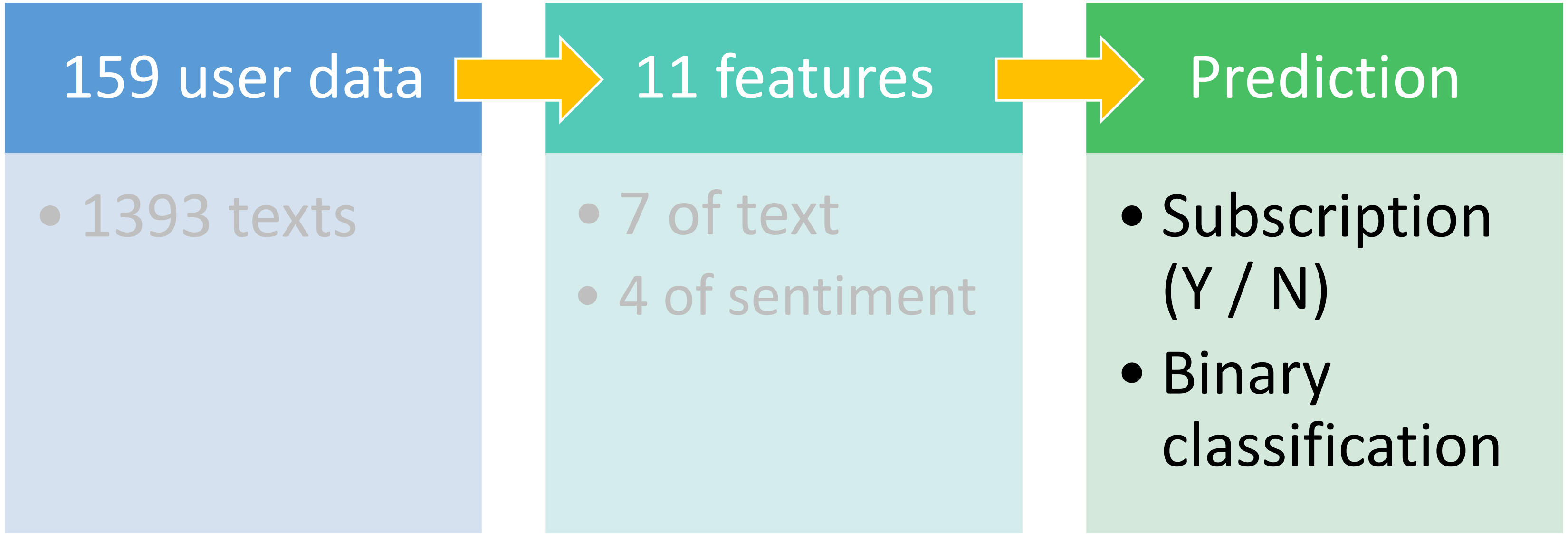
[7 Text Features]

- Total characters
- Total words
- Average character
- Average words
- Number of posts
- Number of likes
- Average likes

[4 Sentiment features]

- Total tone score
- Total factual score
- Average tone score
- Average factual score

Processing data with machine learning models



159 user data

- 1393 texts

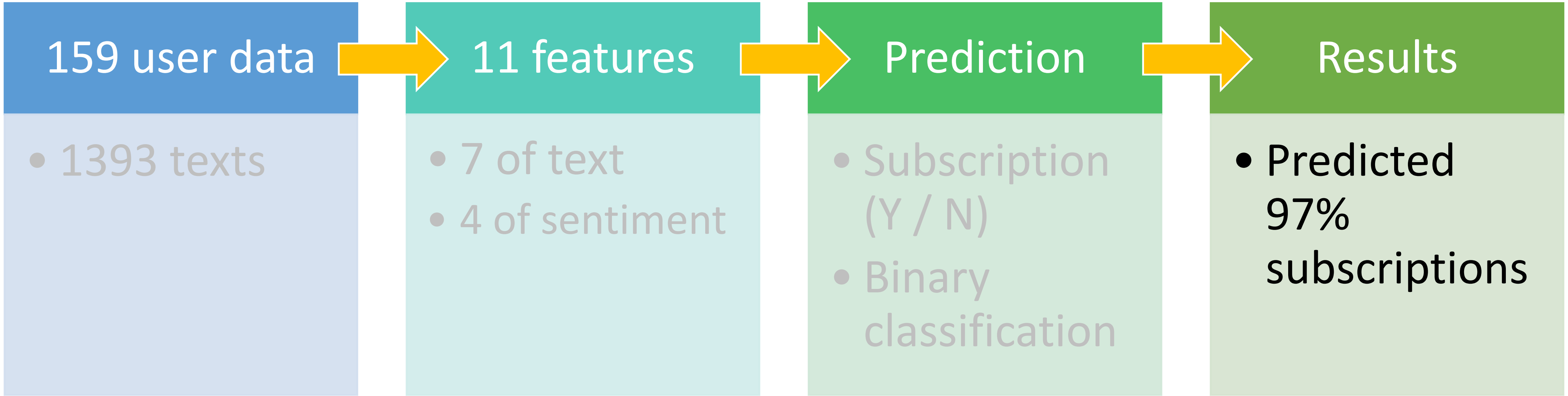
11 features

- 7 of text
- 4 of sentiment

Prediction

- Subscription (Y / N)
- Binary classification

Processing data with machine learning models



159 user data

- 1393 texts

11 features

- 7 of text
- 4 of sentiment

Prediction

- Subscription (Y / N)
- Binary classification

Results

- Predicted 97% subscriptions

How do these models work

Sentiment analysis

- Natural Language Processing (NLP)

How do these models work

Sentiment analysis

- Natural Language Processing (NLP)
- Pre-trained BERT
- Hand labelling 60% texts

How do these models work

Sentiment analysis

- Natural Language Processing (NLP)
- Pre-trained BERT
- Hand labelling 60% texts

Classification

- Ridge, Logistic Regression, XGBoost, Random Forest

scikit learn

dmlc XGBoost

How do these models work

Sentiment analysis

- Natural Language Processing (NLP)
- Pre-trained BERT
- Hand labelling 60% texts

Classification

- Ridge, Logistic Regression, XGBoost, Random Forest
- Stacking

Validation metrics

Sentiment	Accuracy
Tone	0.851
Factual	0.776

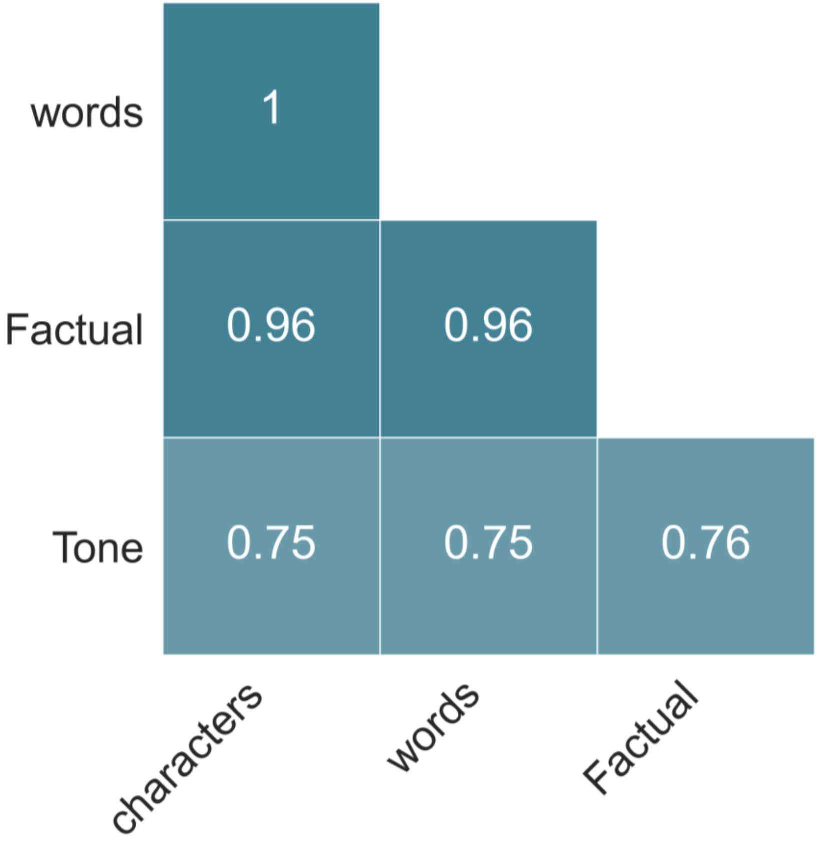
Classification	Precision	Recall
Regular classifiers	0.56 ± 0.03	0.89 ± 0.04
Stacked	0.58 ± 0.03	0.97 ± 0.03

Yes, the text data can predict user decision

Yes, the text data can predict user decision and …

- Strong correlation
  - Text features
  - Sentiment features

Strong Correlation!



Yes, the text data can predict user decision
and it gets even simpler!

- Strong correlation
    - Text features
    - Sentiment features

Text features

Yes, the text data can predict user decision
and it gets even simpler!

- Strong correlation
    - Text features
    - Sentiment features

Text features

- Good enough?

Yes, the text data can predict user decision and it gets even simpler!

- Strong correlation
  - Text features
  - Sentiment features

Text features

- Good enough?

Classification	Precision	Recall
Text features	0.57 ± 0.04	0.90 ± 0.02
Sentiment features	0.56 ± 0.04	0.92 ± 0.04

Yes, the text data can predict user decision
and it gets even simpler!

- Strong correlation
    - Text features
    - Sentiment features

Text features

- Good enough!

Classification	Precision	Recall
Text features	0.57 ± 0.04	0.90 ± 0.02
Sentiment features	0.56 ± 0.04	0.92 ± 0.04
All features	0.56 ± 0.03	0.89 ± 0.04

Yes, the text data can predict user decision
and it gets even simpler!

- Strong correlation
  - Text features
  - Sentiment features

Text features

- Good enough!
- Easy to scale up

Classification	Precision	Recall
Text features	0.57 ± 0.04	0.90 ± 0.02
Sentiment features	0.56 ± 0.04	0.92 ± 0.04
All features	0.56 ± 0.03	0.89 ± 0.04

Yes, the text data can predict user decision and it gets even simpler!

- Strong correlation
  - Text features
  - Sentiment features

Text features

  - Good enough!
  - Easy to scale up

Takeaways

Yes, the text data can predict user decision and it gets even simpler!

- Strong correlation
  - Text features
  - Sentiment features

Text features

- Good enough!
- Easy to scale up

Takeaways

- User communication: strong indicator

Yes, the text data can predict user decision and it gets even simpler!

- Strong correlation
  - Text features
  - Sentiment features

Text features

- Good enough!
- Easy to scale up

Takeaways

- User communication: strong indicator

Yes, the text data can predict user decision and it gets even simpler!

- Strong correlation
  - Text features
  - Sentiment features

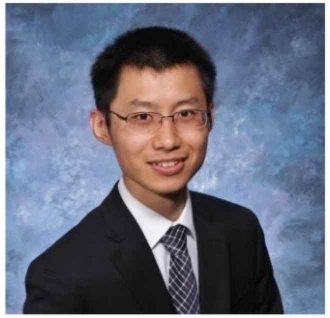
Text features

- Good enough!
- Easy to scale up

Takeaways

- User communication: strong indicator
- Recommend A/B testing on app features (causation)

Zelong (Eric) Zhang

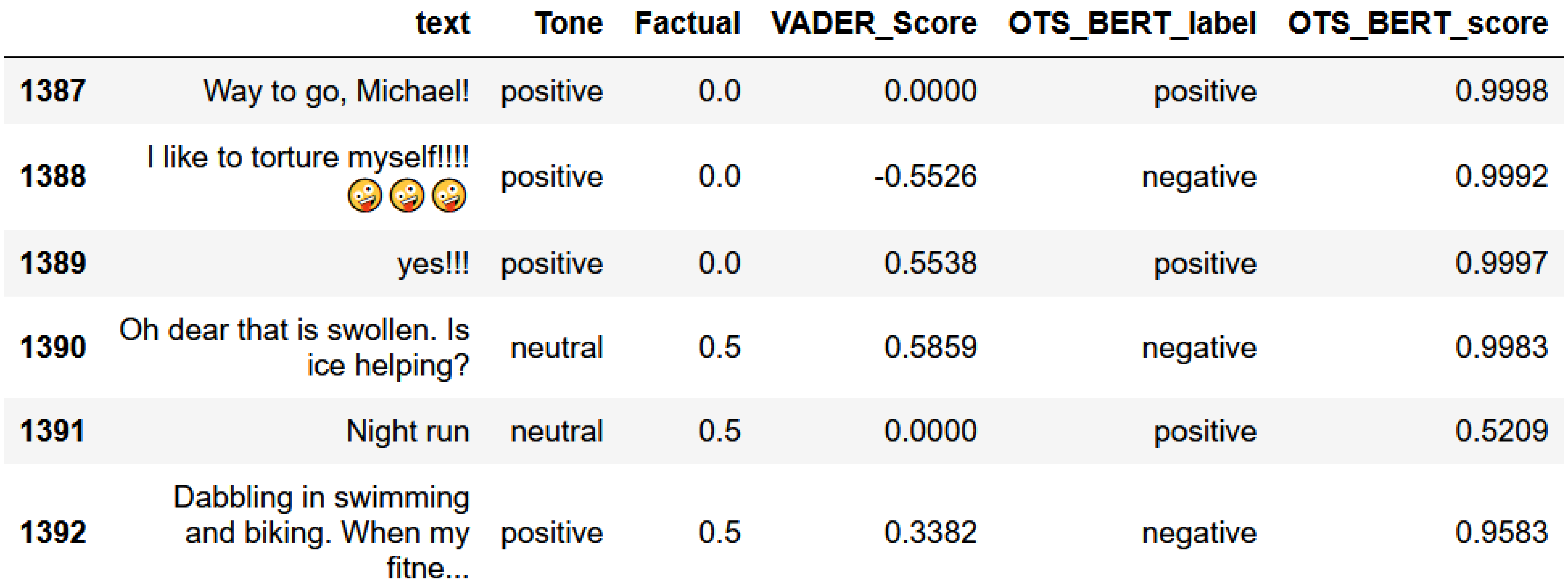


- PhD in Computational Chemistry
- Award-winning film (US DOE), photography
- User Experience and Decision-Making

Stony Brook University

LSU
LOUISIANA STATE UNIVERSITY

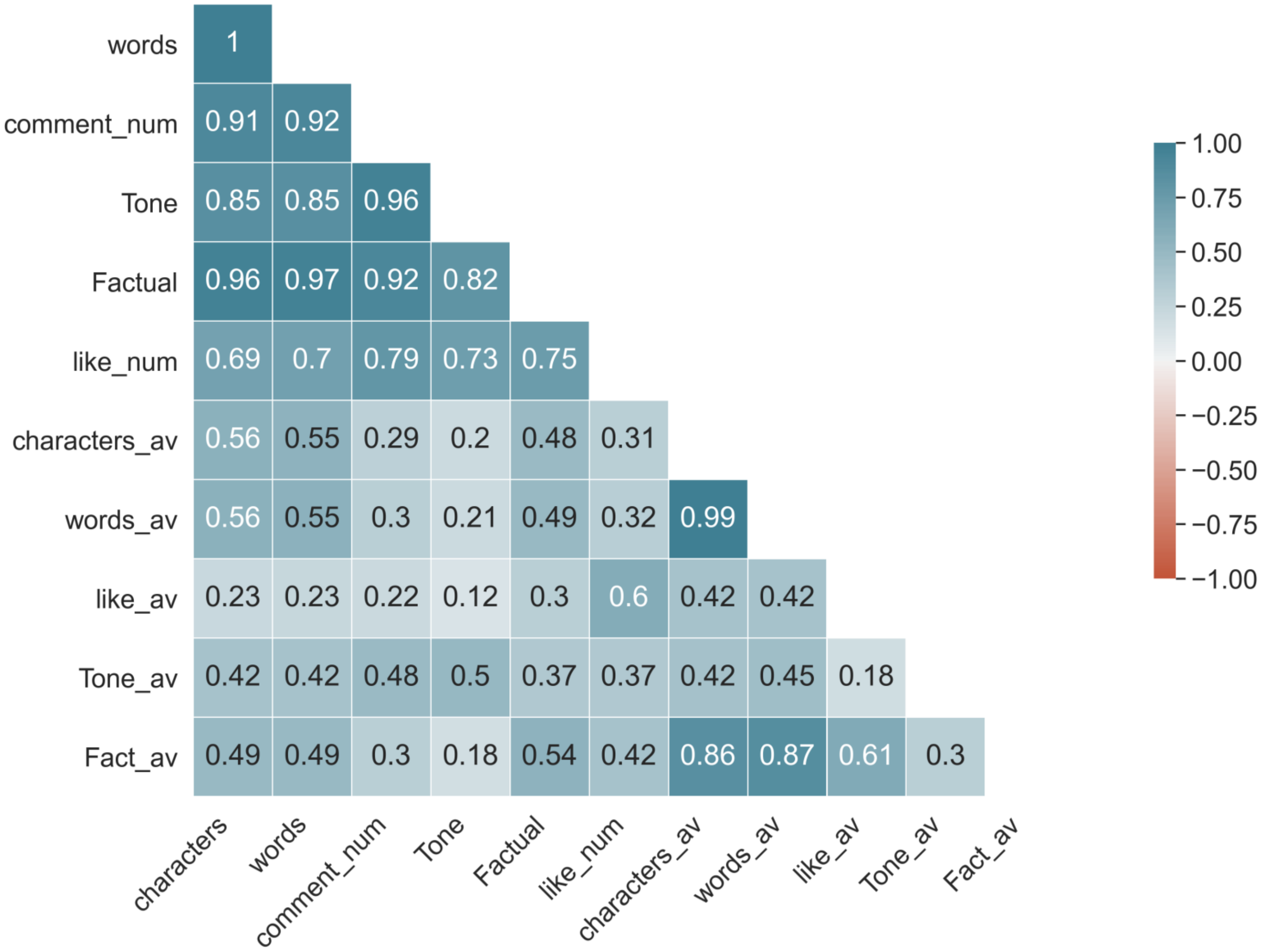
Sanity check of sentiment analysis



	text	Tone	Factual	VADER_Score	OTS_BERT_label	OTS_BERT_score
1387	Way to go, Michael!	positive	0.0	0.0000	positive	0.9998
1388	I like to torture myself!!!! 🤪🤪🤪	positive	0.0	-0.5526	negative	0.9992
1389	yes!!!	positive	0.0	0.5538	positive	0.9997
1390	Oh dear that is swollen. Is ice helping?	neutral	0.5	0.5859	negative	0.9983
1391	Night run	neutral	0.5	0.0000	positive	0.5209
1392	Dabbling in swimming and biking. When my fitne...	positive	0.5	0.3382	negative	0.9583

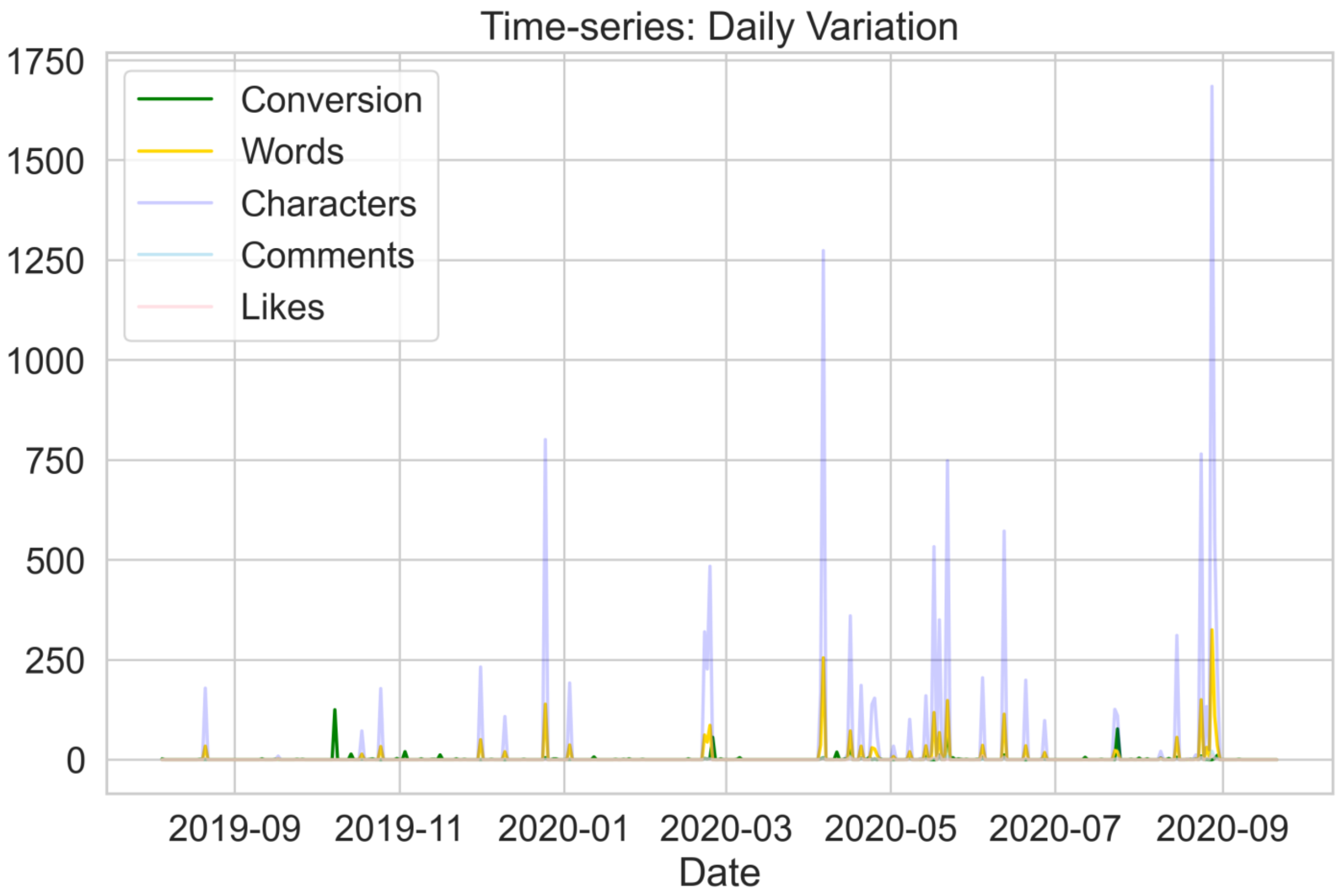
VADER: a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media.
OTS BERT: “distilbert-base-uncased-finetuned-sst-2-english"

Strong Correlations between Features

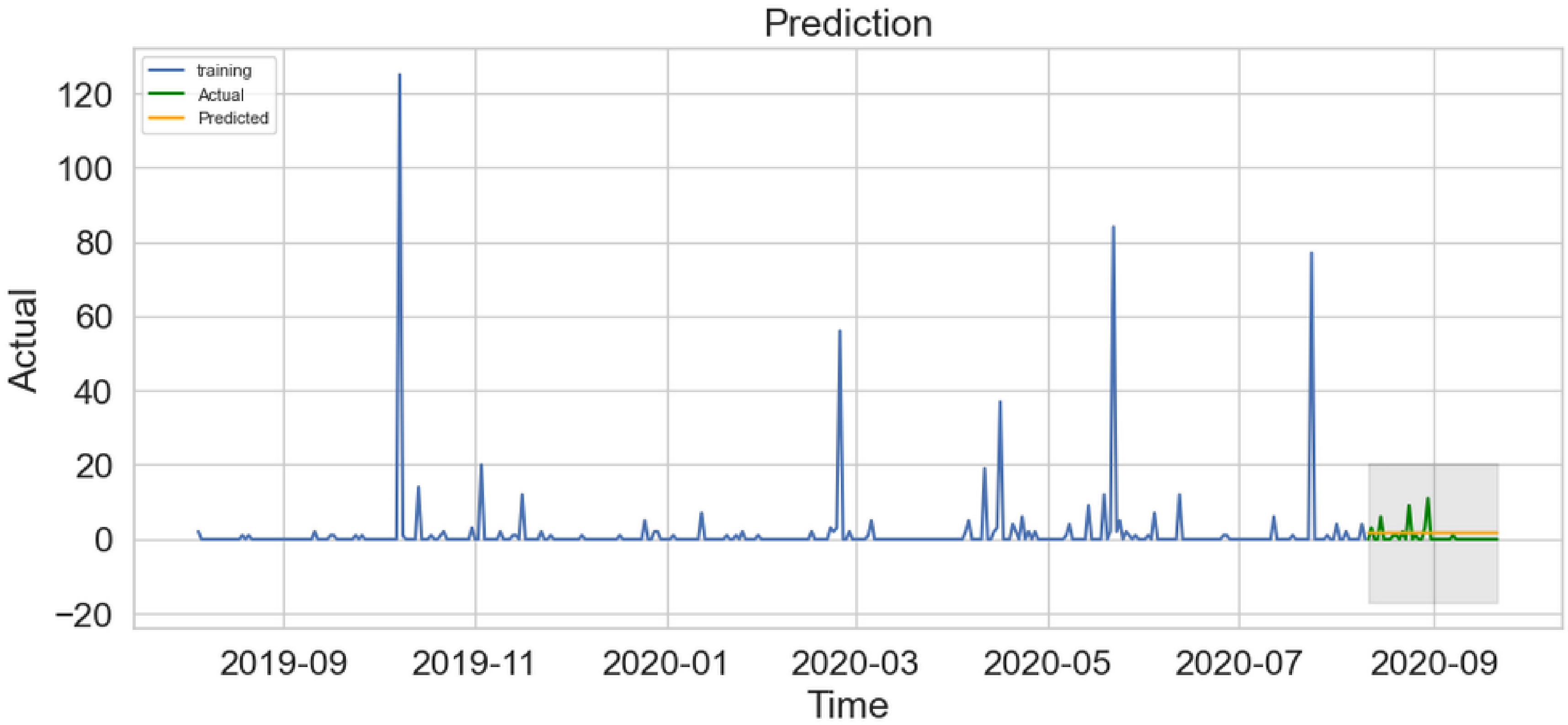


	characters	words	comment_num	Tone	Factual	like_num	characters_av	words_av	like_av	Tone_av	Fact_av
words	1										
comment_num	0.91	0.92									
Tone	0.85	0.85	0.96								
Factual	0.96	0.97	0.92	0.82							
like_num	0.69	0.7	0.79	0.73	0.75						
characters_av	0.56	0.55	0.29	0.2	0.48	0.31					
words_av	0.56	0.55	0.3	0.21	0.49	0.32	0.99				
like_av	0.23	0.23	0.22	0.12	0.3	0.6	0.42	0.42			
Tone_av	0.42	0.42	0.48	0.5	0.37	0.37	0.42	0.45	0.18		
Fact_av	0.49	0.49	0.3	0.18	0.54	0.42	0.86	0.87	0.61	0.3	

Time-series


Time-series: Daily Variation
Conversion
Words
Characters
Comments
Likes
1750
1500
1250
1000
750
500
250
0
2019-09
2019-11
2020-01
2020-03
2020-05
2020-07
2020-09
Date

Time-series Analysis (ARIMAX)



statsmodels