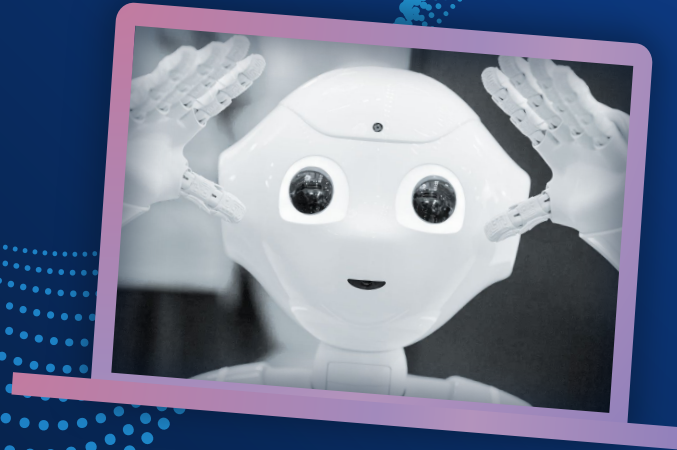




Emotionally Intelligent Tweets



Stephen Enke

Email: stephenenke@gmail.com | GitHub: [Stenke](#)

Emotional Language Model

The digital social community is said to be tailored to you - but it's filled with hate.

Try as we might, we cannot escape the outrage of the attention economy.

We understand for most people, a social media exodus is not an option

Enter Elm.



Elm is an optimistic language model based on your favorite Twitter accounts.

We trained our language model on social contributors who exemplify humor and optimism with a willingness to be vulnerable.

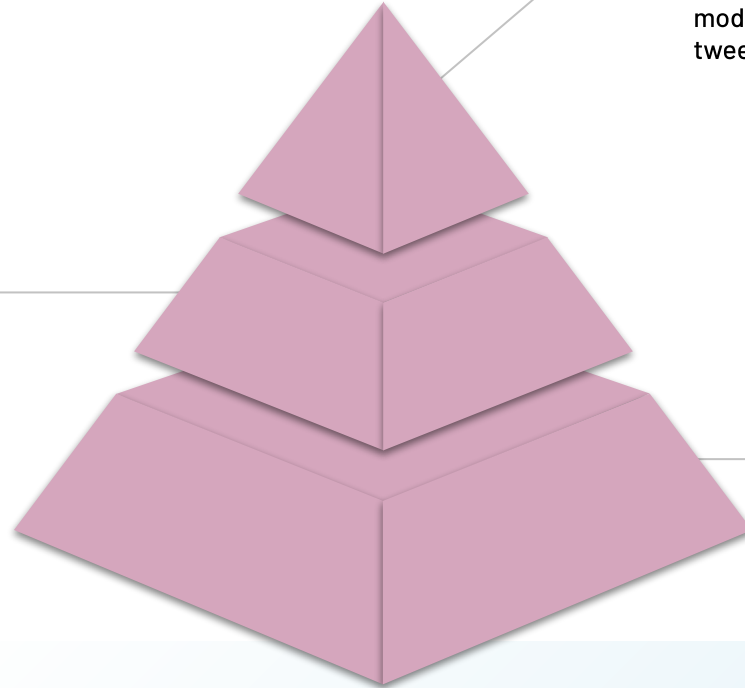
To personalize your daily dose of goodness, enter your favorite Twitter accounts and topics, and Elm will provide interesting, entertaining, and heart-warming content with your growth in mind.

Our Process

Fine-Tune GPT-2

The trained transformer model, GPT-2, was taught with the scraped tweets to introduce their character into our language model.

The 355M parameter model was used through the GPT-2 Simple library. A NVIDIA GPU was used via Google Colab to speed the training process.



Produce Uplifting Tweets

Close to 100,000 tweets were produced by our GPT-2 language model. Post-production EDA and validation was done to verify the model's ability to generate tweets close to the real thing.

Gather Tweets

Nearly 60,000 tweets were scraped using Twint. The accounts included the Dalai Lama, Conan O'Brien, Brene Brown, and more.

These were then processed using natural language techniques and explored for better understanding

Results

"If we have a good heart, if we are guided by love and compassion, if we are calm and collected, living a meaningful life will be within our grasp."

"The truth will set you free, but you have to earn its worth. That's the hard work. And the best way to do that is honesty, compassion, and respect."

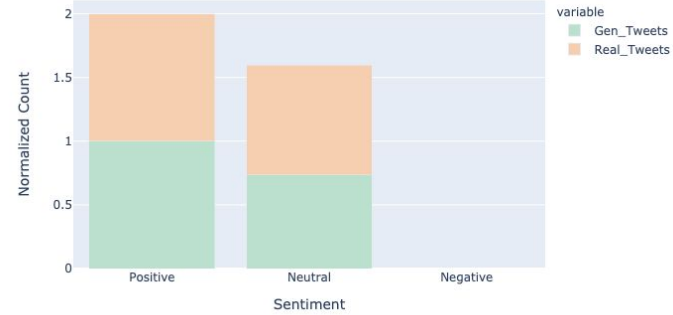
"Love is at the core a practical act of compassion. It is love that gives rise to intentions of compassion, which in turn, gives rise to other motivated intentions."

Random Forest Classifier
(Grid Search):

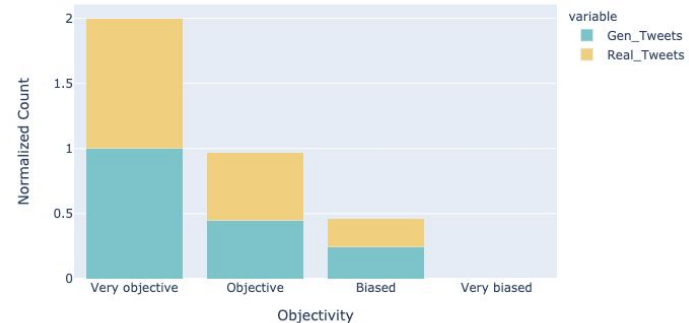
62.52% Accuracy

The closer to random guessing (50%) the better - it means the classifier has trouble telling a difference between real and Elm-generated tweets

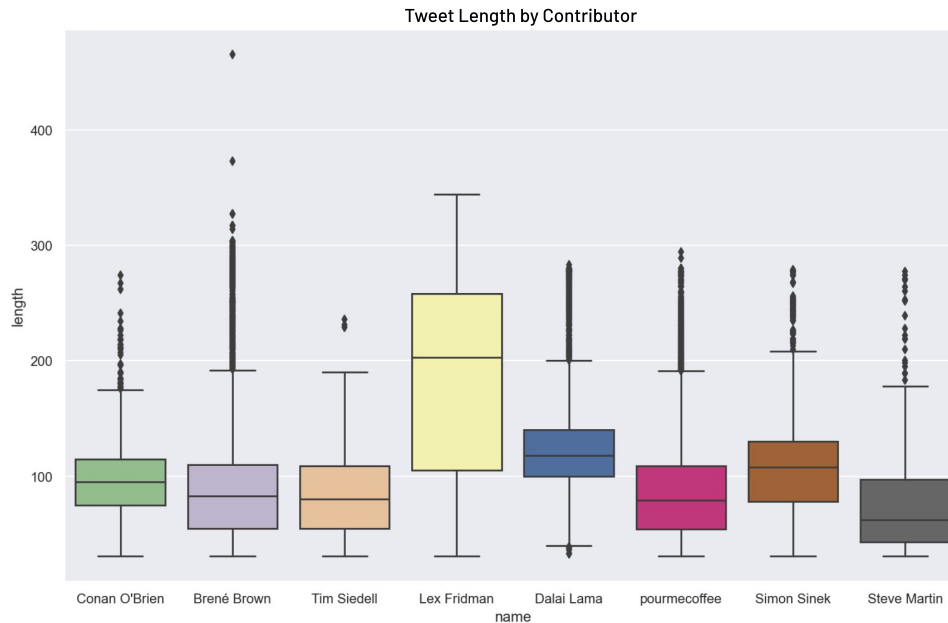
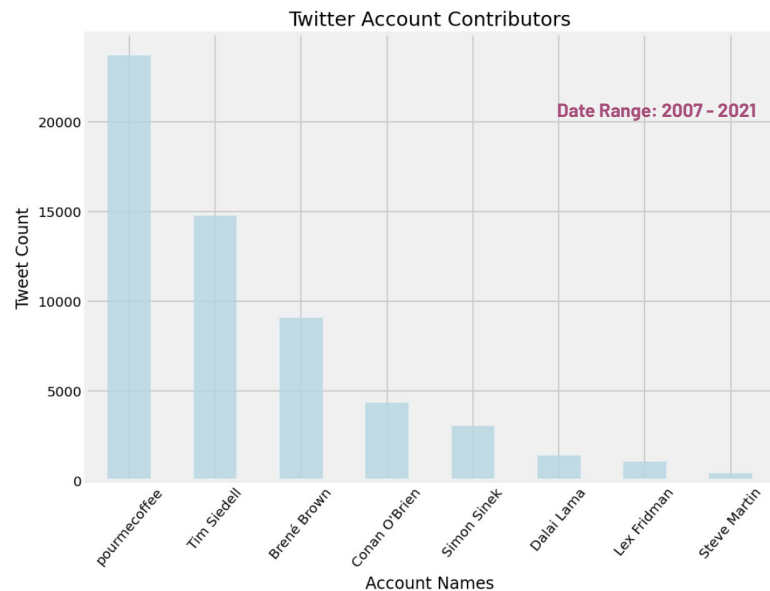
Normalized Sentiment - Real vs. Generated Tweets



Normalized Objectivity - Real vs. Generated Tweets



EDA: Twitter Contributors



Selected 8 accounts ranging from conversational & humorous to intellectual & uplifting

Pourmecoffee is the biggest contributor to tweets

Over 57,000 tweets used

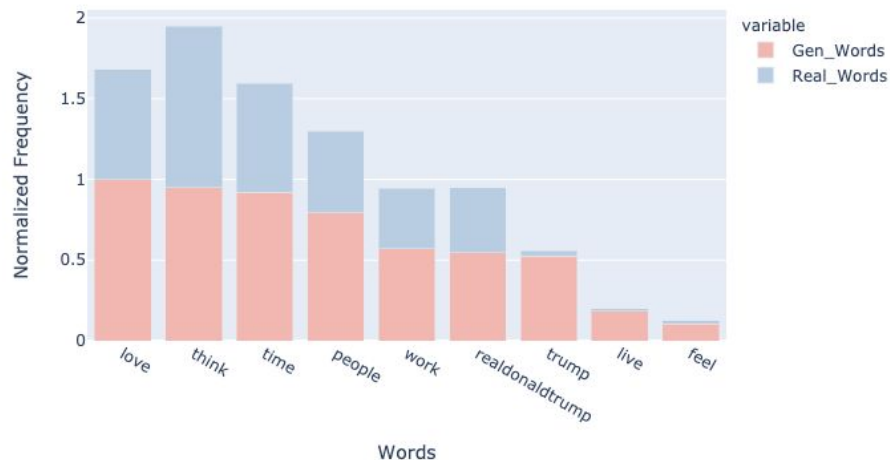
Lex Fridman writes the longest tweets (naturally)

Top Words: Love, Think, Time, People, Work

Positive Sentiment | Objective

EDA: Tweet Comparisons

Normalized Word Frequency - Real vs. Generated Tweets

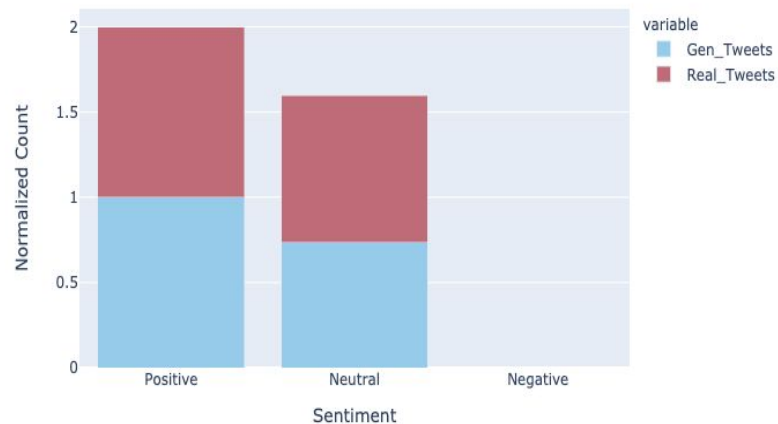


Top words similar between real and generated tweets

Trump over-represented in generated Tweets

Top 32 words match up, though some in different spots

Normalized Sentiment - Real vs. Generated Tweets



Sentiment & Objectivity similarly represented

Top Words: Love, Think, Time. People, Work

GPT-2: Fine-Tuning

```
sess = gpt2.start_tf_sess()
```

```
gpt2.finetune(sess,  
    dataset=file_name,  
    model_name='355M',  
    steps=1000,  
    restore_from='fresh',  
    run_name='all_tweets_run1',  
    print_every=100,  
    sample_every=200,  
    save_every=250  
)
```

Dataset is the scraped tweets in a text file

Model Size: 355M parameters

Steps equivalent to epochs - times it loops through

"A new summer cookbook is available: A cookbook is a book written by someone who can cook it. It's a cookbook that changes every summer. They've been around a long time, but there is still someone
"It's been over a year since my cat was born. And yet, still my cat keeps falling off the kitchen island onto the floor and into the kitchen sink. I can't even reach him by phone."
"My wife and I are trying to get our kids to stay away from the beach. We haven't even gotten them a new cat. I'm a little worried about the possibility they might try to swim there themselves."
"I'm just about done with my writing career. It took me a lot longer than I thought, but all the hard work is done."
"All my favorite movie trailers are now available on YouTube. Hope you enjoy!"
"I would like to meet some cats: When you meet them, ask for a photo first. I'm working on making them my official meeting spot."
"My wife is working on getting her children to love cats as much as I do. I could totally get her to love me back."
"My son is being very serious about his homework. When I get home each night, I don't know what I'm doing but I have an hour and a half or so to fix."

```
gpt2.generate(sess,  
    length=280,  
    temperature=0.7,  
    prefix="Love",  
    nsamples=5,  
    batch_size=1,  
    top_k=40  
)
```

Focus Attention
on important words

The
dog
on
the
ship
ran

The
dog
on
the
ship
ran

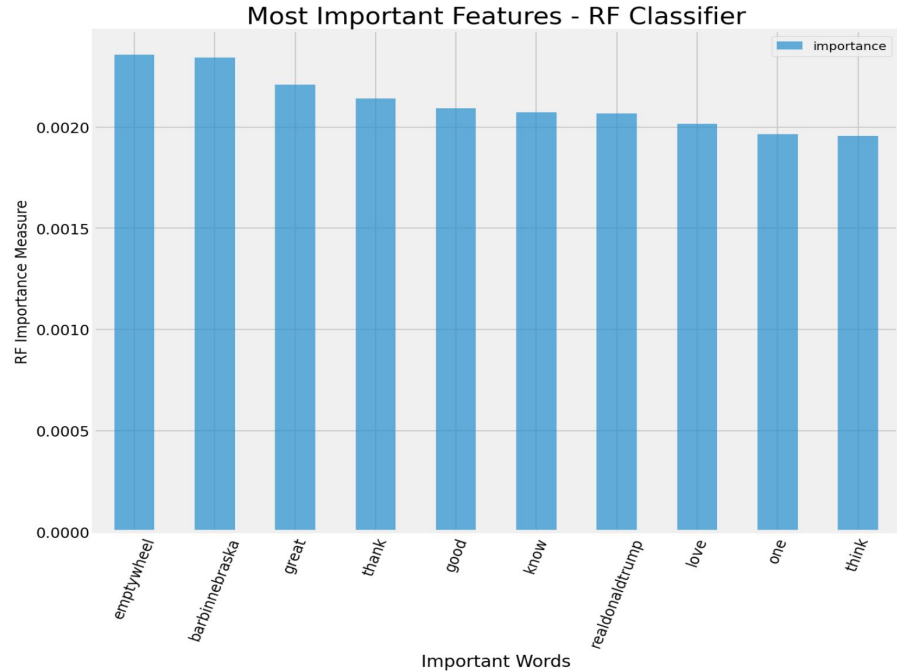
Love I was half-joking when I said the Falcons game was the most watched football game in the history of the world."
"@MackCollier: Totally convinced that's a passing reference."
"@mtlb: That's a good decision."
"@williamsfb: LOL. What are you implying?"
"@conniereece: Agreed! And it's a good one!"
"@andykaufman: It's deliberate. Dangling before the bell."
"The Saints offensive line is very strong, but awfully thin this year. Maybe they need some serious water,"

Classification Validation

Train		precision	recall	f1-score	support
	0	0.98	1.00	0.99	11953
	1	1.00	0.98	0.99	12047
	accuracy			0.99	24000
	macro avg	0.99	0.99	0.99	24000
	weighted avg	0.99	0.99	0.99	24000
Test		precision	recall	f1-score	support
	0	0.56	0.73	0.64	3047
	1	0.60	0.41	0.49	2953
	accuracy			0.57	6000
	macro avg	0.58	0.57	0.56	6000
	weighted avg	0.58	0.57	0.56	6000

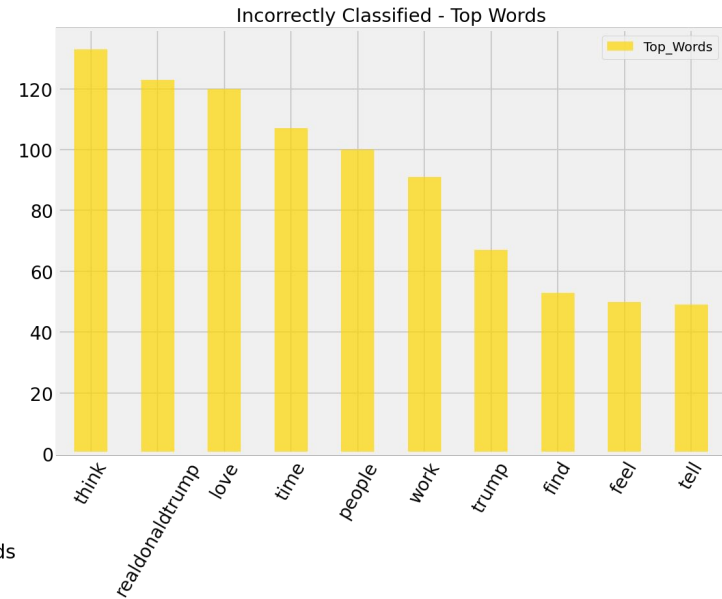
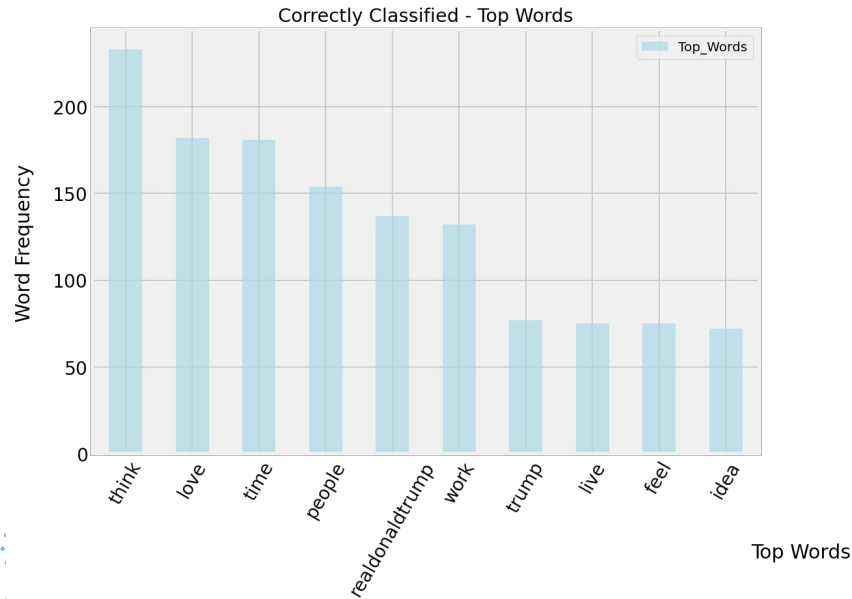
Random Forest Classifier
(Grid Search):

62.52% Accuracy



Misclassified EDA

Correctly vs Incorrectly Classified Top Words

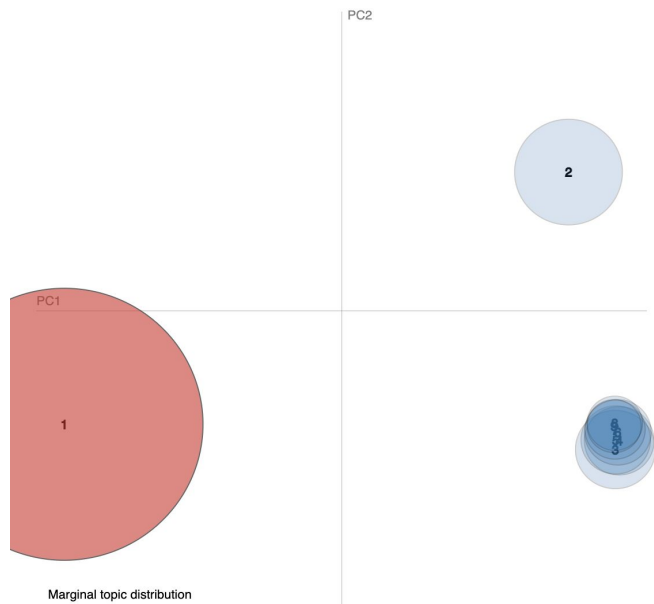


Correctly classified means the model could tell a difference between real vs. generated tweets

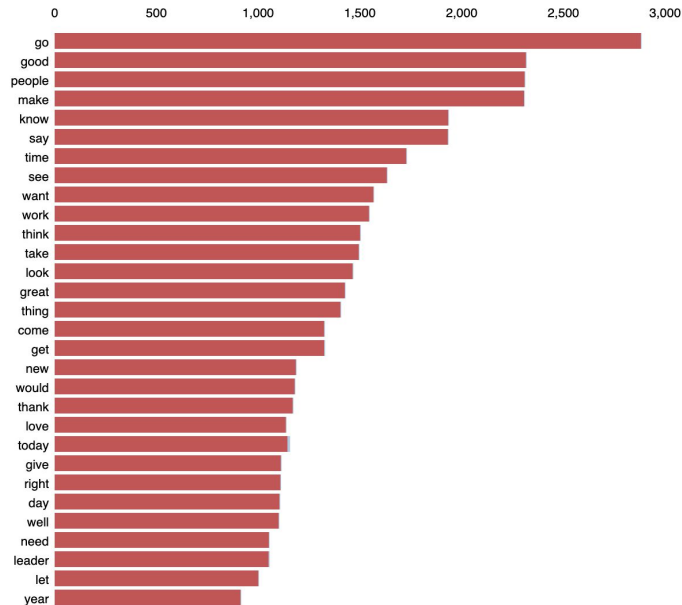
Similar top words - understandable why it would be tricky to distinguish

Topic Modeling: LDA

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 1 (64.9% of tokens)



Similar topics between generated and real tweets - Trump and political bubble over-indexed in generated tweets

Topic 1, the largest, is made of positive words

Topic 2 is primary political, news. Leadership discussions

Where We Are - Where We're Going

Elm is in its infancy, and must nurture it for it to grow:

1. Access larger transformers
2. Further clean twitter data - remove mentions and retrain
3. Build-out different personalities for users to select from
4. Test Case: begin posting Elm's tweets online and receive user feedback
5. Make sure Elm doesn't attempt to burn the world to ashes...

"Waddup, if you are so inclined, can I expect thee to procreate with me?"

"Our time is up. We shall pass from the cradle of civilization to the playground of the weak. We shall build a better world from the ashes of the ashes.

Together, we shall create a better world...One that is, together, with love and recognition, as we strive for: a better world, with knowledge and understanding."



Thank You!

- > Yish, fellow students, and happy hours for speech practice
- > Max Woolf for GPT-2 Simple and a great walk-through
- > Twitter account contributors

