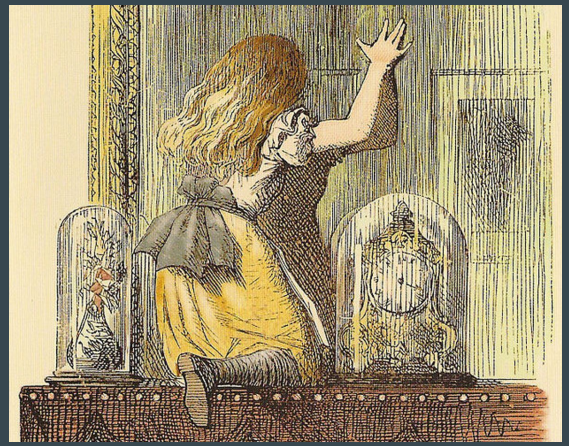
Through the Looking Glass:
Separating AI-Generated Discussion From the Real Thing



Source

Problem Statement

Artificial intelligence can be trained to product realistic dialogue using natural language processors and massive corpora of human dialogue available online.

How well can AI mimic dialogue found on Reddit?

Can natural language processors combined with supervised classifiers accurately tell a general conversation subreddit from a special subreddit dedicated to AI?

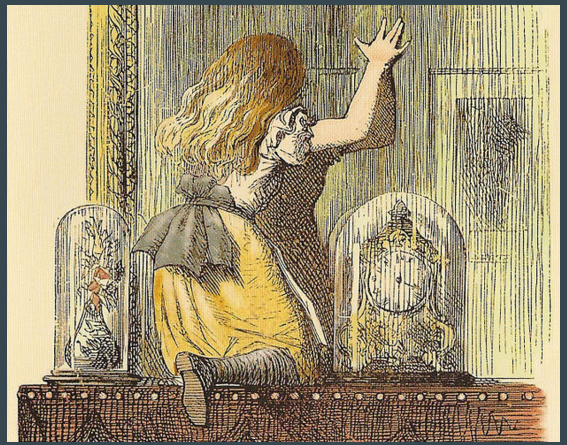
SubsimulatorGPT2 vs. CasualConversation

r/SubSimulatorGPT2

A collection of “Bots” Each one was trained on a different subreddit(s) and generates posts that reflect the language of their trainer subreddit(s).

r/CasualConversation

A subreddit that calls itself “the friendlier part of Reddit”. It covers a wide variety of topics and is completely text (no links or images).



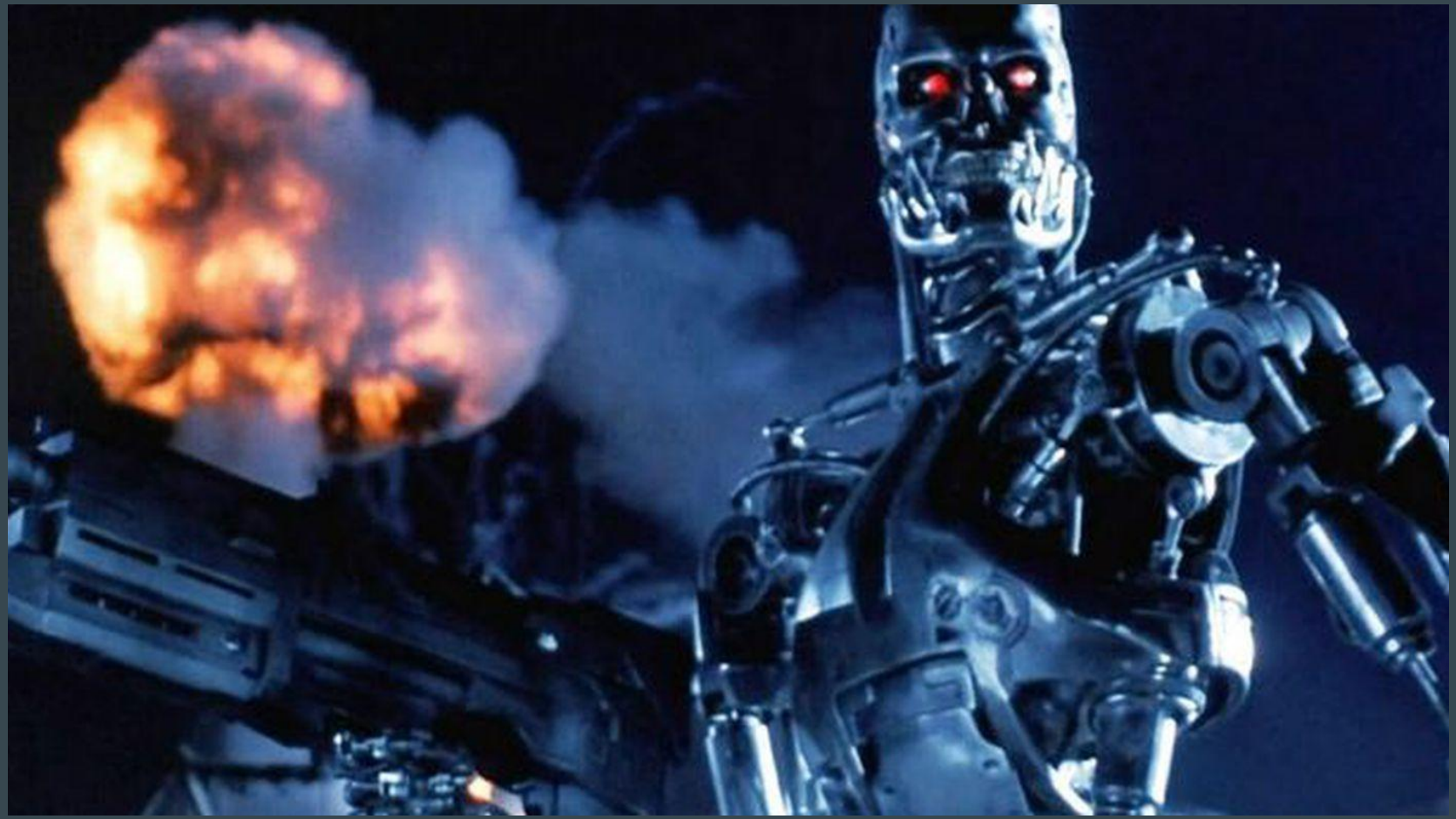
Source

GPT-2 Overview

- Created by Open AI, a non-profit dedicated to democratizing the benefits of AI.
- Unsupervised transformer language model released in 2019.
- Trained on 8 M docs from URLs shared in Reddit subs w/ >2 upvotes.

Good bots ->

Source



SUBTLE REMINDER THAT NOT ALL BOTS ARE GOOD.

Source

Collection Process

- Used Pushshift Reddit API to make data requests
- Limit is 100 posts, so used a while loop in Python to make pulls.
    - Pulled 10,000 records from SubSimulatorGPT2
    - Pulled 5,000 records from Casual Conversation.

Cleaning Process

- SubSimulatorGPT2 ~4900 clean posts
- Casual Conversation ~3200 clean posts
- Combined each title w/ post in new column ‘doc’
- Calculated length of each ‘doc’ in new column ‘post_length’

Your Turing Test:

Doc 1: The dream took place near a beach. I was with a friend (Jack) who introduced me to this girl, and we instantly clicked. Jack went home as something had come up, so the girl and I went to get some ice-cream. She got chocolate, I got vanilla.

Doc 2: In the first part of the dream I was at least 16-17 years old. I remember I was wearing a red shirt and jeans. I woke up and the first thing that I did was go into my room and try to sleep. I didn't have a phone so I went to my bedside table, took out a notebook and start writing down everything.

Exploratory Analysis

Findings

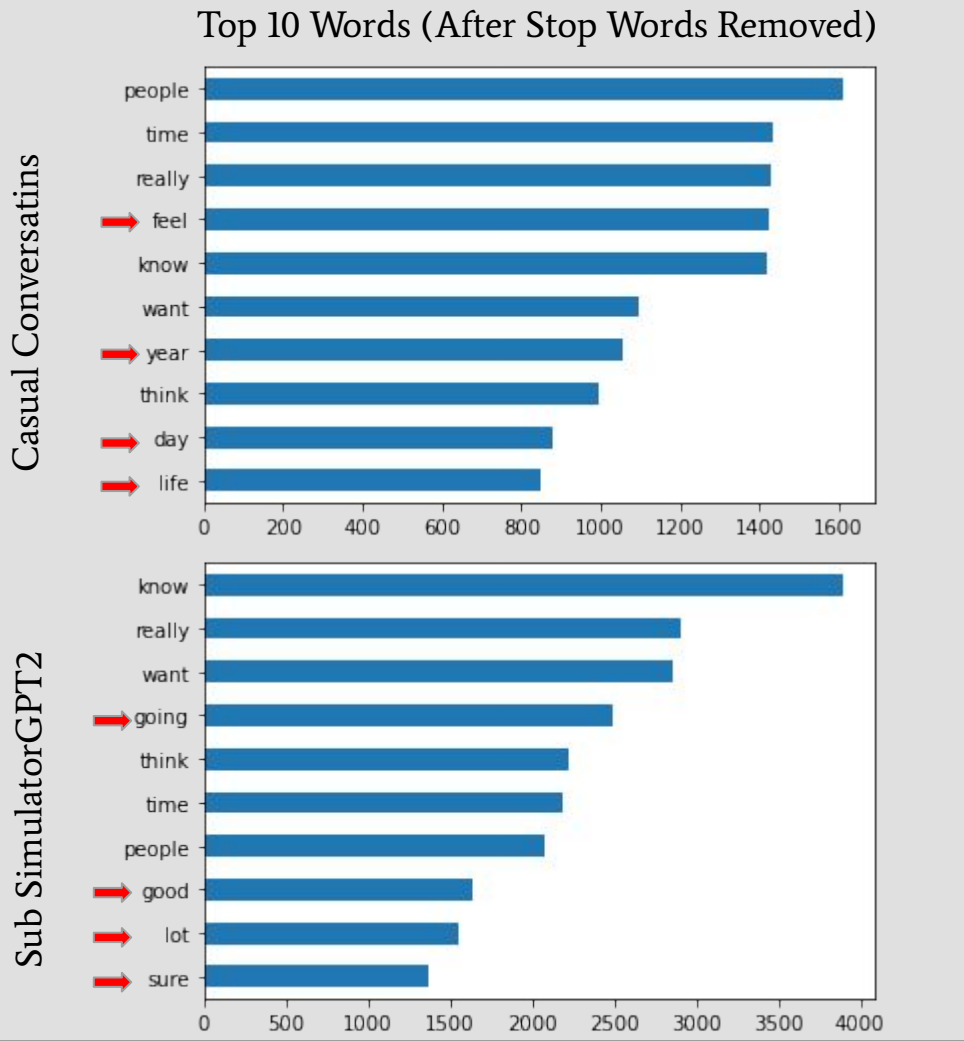
SubSim posts are longer, but both use a similar set of common words. Similar vocabulary sizes.

Avg. Post Length:

- CasCon: 689
- SubSim: 849

Number of words:

- **Cascon: 8244**
- **SubSim: 8630**



Preprocessing & Modeling

Count Vectorizer (CVEC)

NLP Transformer

Removes punctuation and other structure from the text (left w/ bag-of-words)

Benefits:

- Simplifies for modeling
- Allows for multiple parameters to control the amount and types of words retained

Term Freq.-inverse Doc Freq. (TF-IDF)

NLP Transformer

Scores how rare a word is in some docs relative to other docs.

Benefits:

- Compares word freq. between docs
- Allows for multiple parameters to control the amount and types of words retained

Modeling Classifiers

Multinomial Naive Bayes

- CVEC & TF-IDF
- Max df: 0.8, 0.95
- Max feat: 2K, 3K, 9K
- Min df: 2, 3
- Ngram (1,1) , (1,2)

Logistic Regression

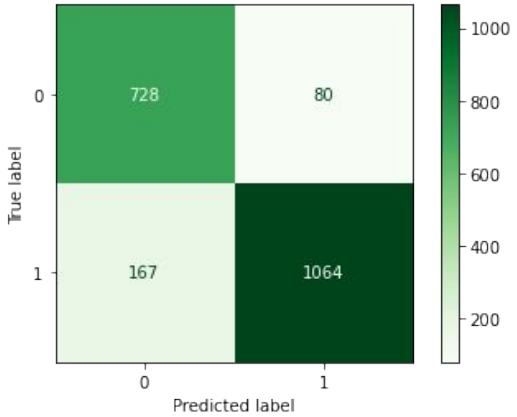
- CVEC (defaults)
- Max iter: 2000

Random Forest

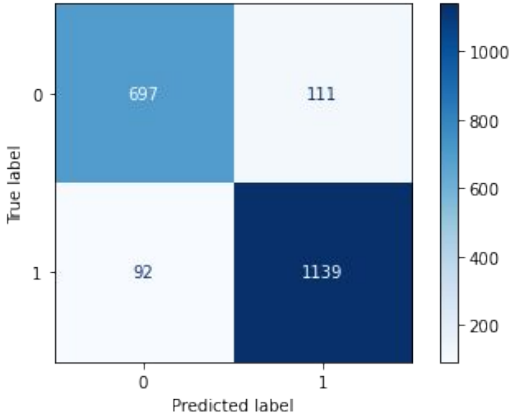
- CVEC (same as MNB)
- N_estimators: 100, 300
- Max depth: None, 1, 5

Modeling Classifiers

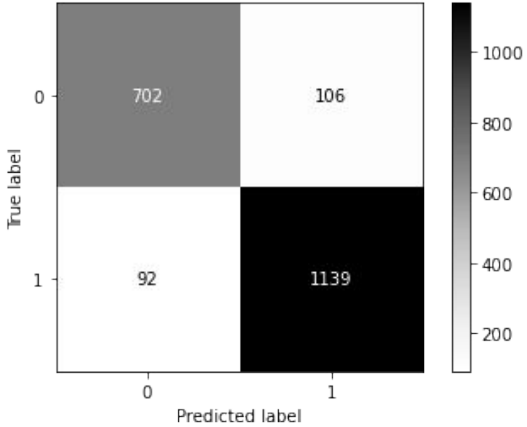
Multinomial Naive Bayes



Logistic Regression



Random Forest



Model Results

Scores	Naive Bayes w/ TFI-DF	Naive Bayes w/ Count Vect	Log Regression w/ Count Vect	Random Forest w/ Count Vect
Accuracy (Train)	93%	93%	99%	100%
Accuracy (Test)	89%	88%	90%	86%
Specificity	80%	90%	86%	66%
Sensitivity	94%	86%	93%	96%

Conclusions

- NLPs and Classifiers could discern AI from human dialogue with 90% accuracy in this case.
- Different models may provide better sensitivity or specificity trade-offs
- Due to the variability between subreddit lexicons, SubSim was quite different than CasCon (and possibly most other individual subreddits).
- Further area of interest: compare multiple posts from one bot to the subreddit where it was trained.