Project 3: Subreddit Classification

Sam Waldner

Problem Statement:

Misinformation on the internet is found just about everywhere. This project seeks to explore two prominent informational subreddits to determine general bias toward types of questions. Then we will build a model that can help determine which one is better suited to an individual's question.

Background

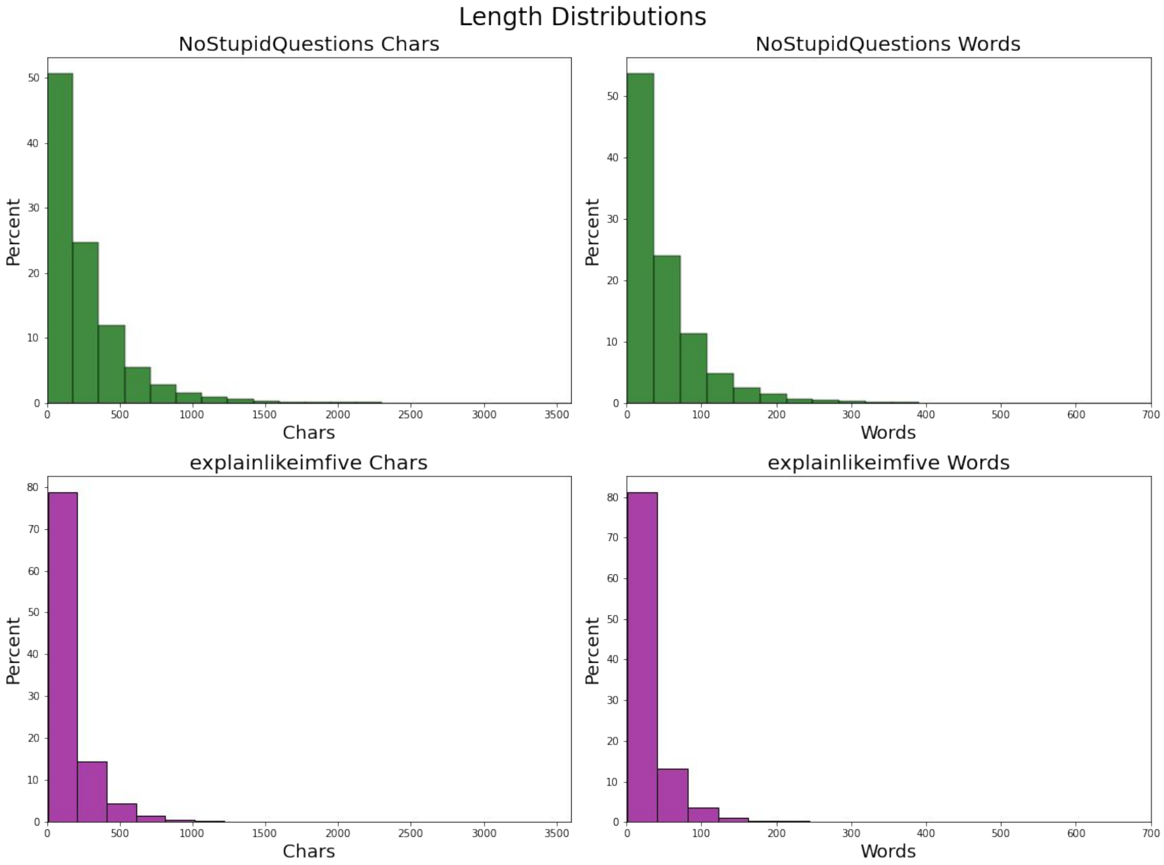
- “4% of U.S. adults report using the site…70% of Reddit users say they get news there.” - Pew Research Center, 2016
- “Reddit was the only other platform polled about that experienced statistically significant growth during this time period – increasing from 11% in 2019 to 18% today.” - Pew Research Center, 2021
- r/NoStupidQuestions
  - 2.5m Members
  - Created Feb 2, 2013
- r/explainlikeimfive
  - 20.0m Members
  - Created Jul 28, 2011

Comparing Length of Posts by Percent

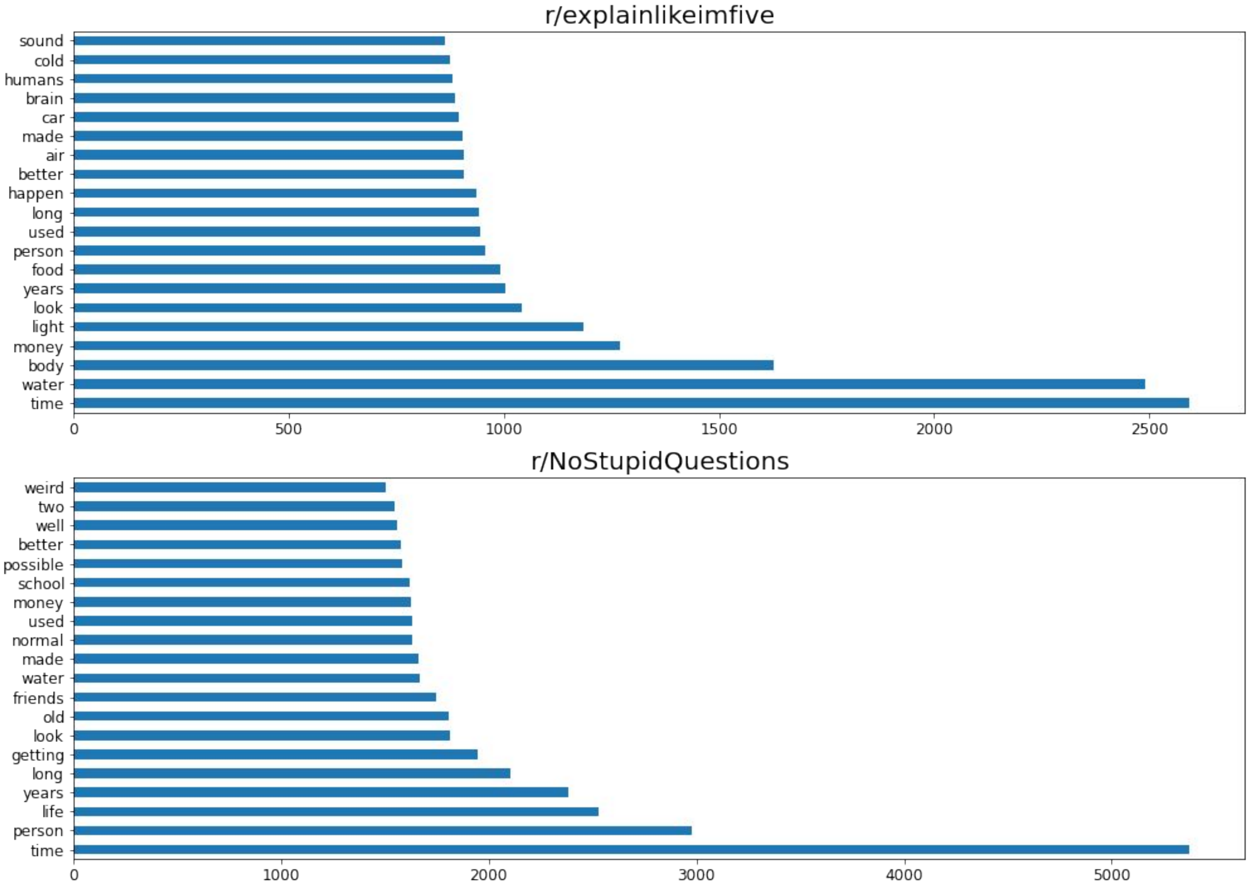
- “Questions” ~30% fewer 0-100 word posts than “Explain”

- Posts over 500 words:

NoStupidQuestions	71
explainlikeimfive	3



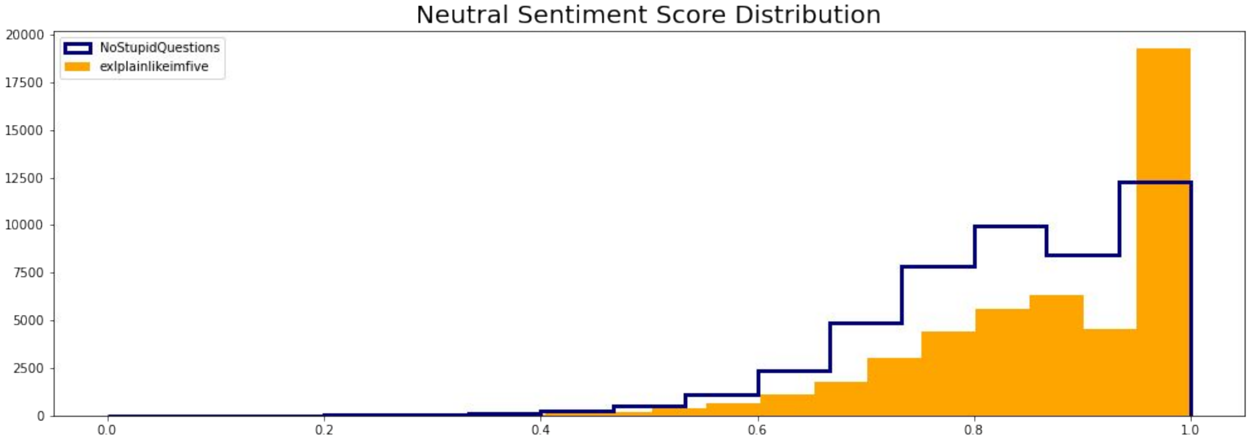
Top 20 Words



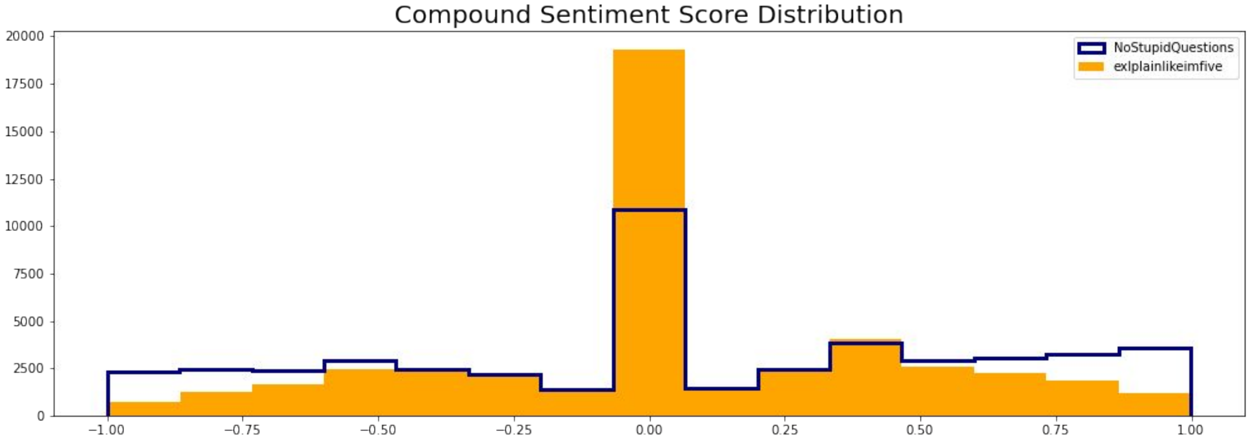
- Subjective words vs Objective words

1. “How does time dilation work?” -explainlikeimfive

2. “Why is oversharing a red flag? I overshare all the time. Oops.” -NoStupidquestions

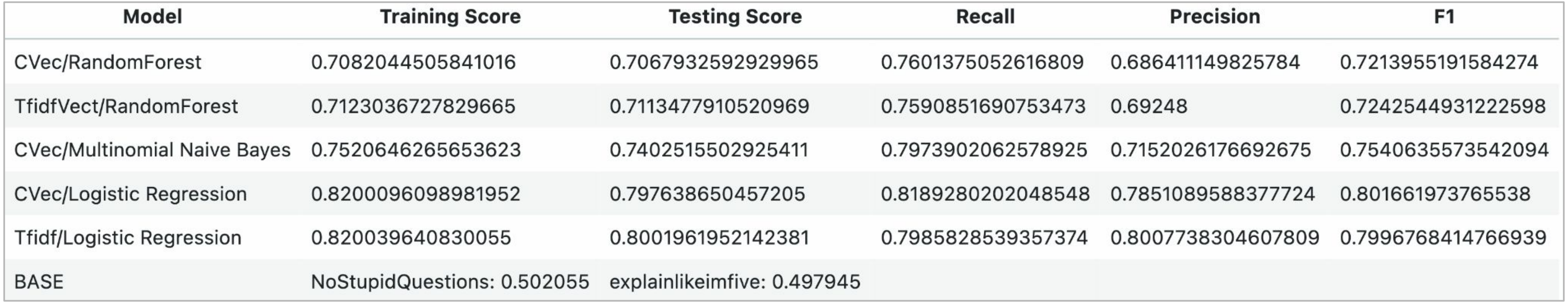


Neutral Sentiment Score Distribution



Compound Sentiment Score Distribution

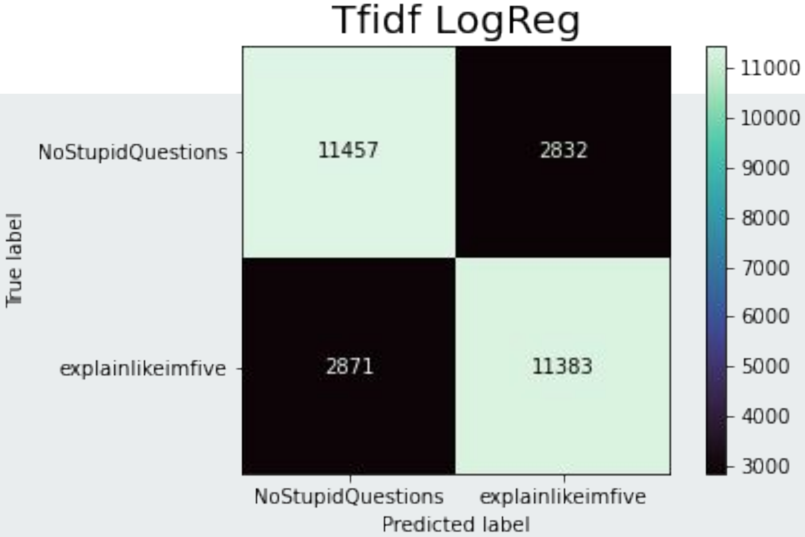
- Highlighting Neutrality in Sentiment Analysis

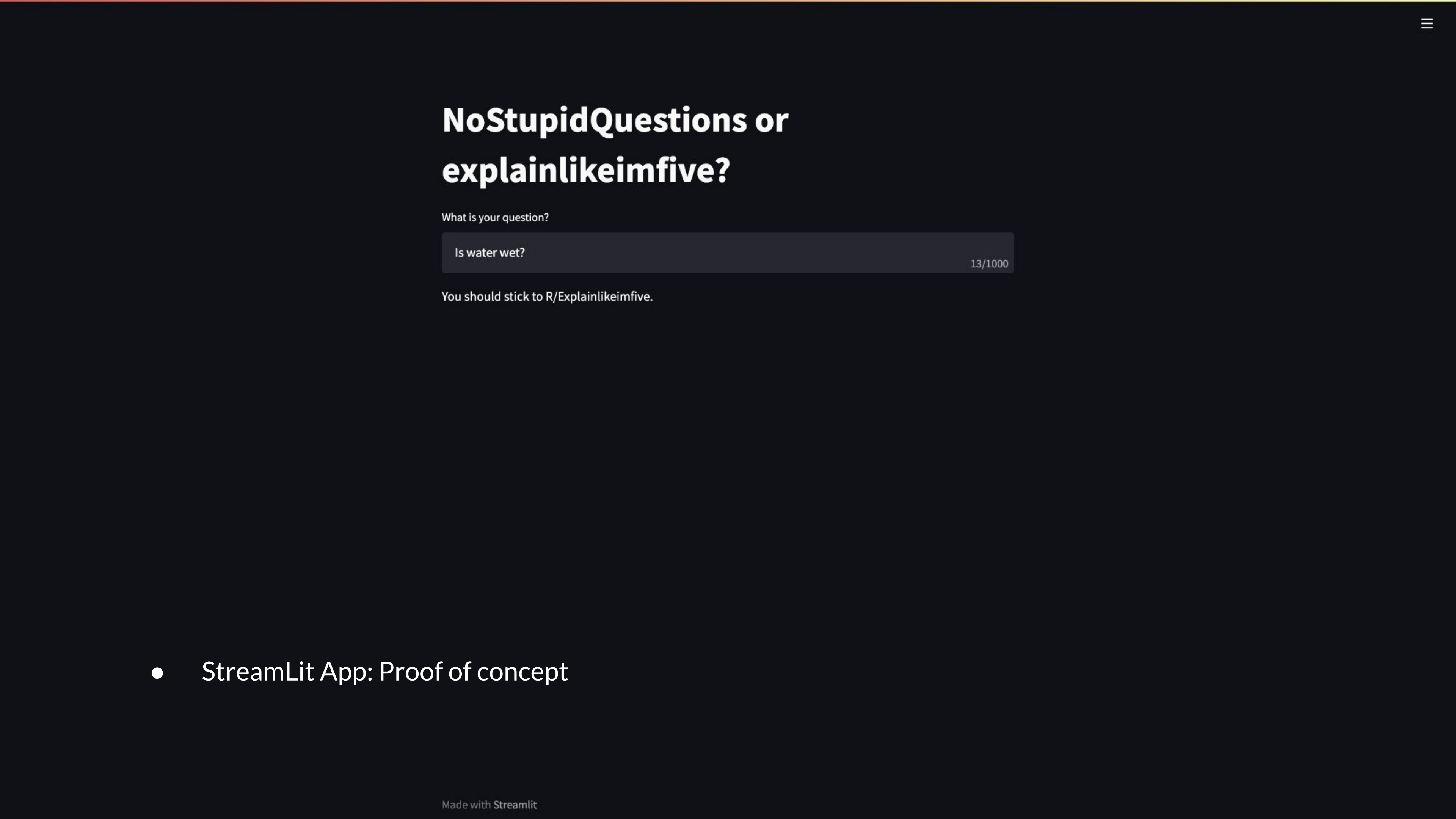


Model	Training Score	Testing Score	Recall	Precision	F1
CVec/RandomForest	0.7082044505841016	0.7067932592929965	0.7601375052616809	0.686411149825784	0.7213955191584274
TfidfVect/RandomForest	0.7123036727829665	0.7113477910520969	0.7590851690753473	0.69248	0.7242544931222598
CVec/Multinomial Naive Bayes	0.7520646265653623	0.7402515502925411	0.7973902062578925	0.7152026176692675	0.7540635573542094
CVec/Logistic Regression	0.8200096098981952	0.797638650457205	0.8189280202048548	0.7851089588377724	0.801661973765538
Tfidf/Logistic Regression	0.820039640830055	0.8001961952142381	0.7985828539357374	0.8007738304607809	0.7996768414766939
BASE	NoStupidQuestions: 0.502055	explainlikeimfive: 0.497945			

Models and Scores

- TfidfVectorizer Logistic Regression (L1):
  - Train Acc: 0.82
  - Test Acc: 0.80
- Baseline:
  - ~50/50





NoStupidQuestions or explainlikeimfive?

What is your question?

Is water wet?

13/1000

You should stick to R/Explainlikeimfive.

Made with Streamlit

- StreamLit App: Proof of concept

Conclusion

- Model minor success
  - ~80% effective at predicting correct subreddit
  - Test more params and model variations
- Subreddits
  - NoStupidQuestions: Emotional, Anecdotal, Bias
  - Explainlikeimfive: Neutral, Succinct, Impersonal

- App Integration
  - Could be flushed out and adapted to act as gateway to Reddit contributors, making sure any post is predicted to be appropriate to subreddit before allowing user to proceed

Citations

- Auxier, Brooke, and Monica Anderson. “Social Media Use in 2021.” *Pew Research Center: Internet, Science & Tech*, Pew Research Center, 9 Apr. 2021, https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/.

- Barthel, Michael, et al. “Seven-in-Ten Reddit Users Get News on the Site.” *Pew Research Center's Journalism Project*, Pew Research Center, 27 Aug. 2020, https://www.pewresearch.org/journalism/2016/02/25/seven-in-ten-reddit-users-get-news-on-the-site/.