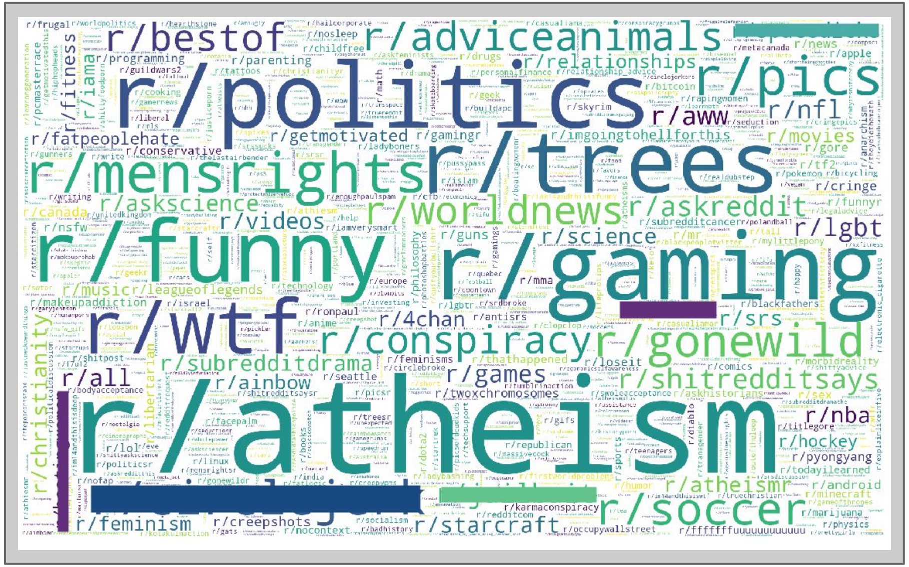
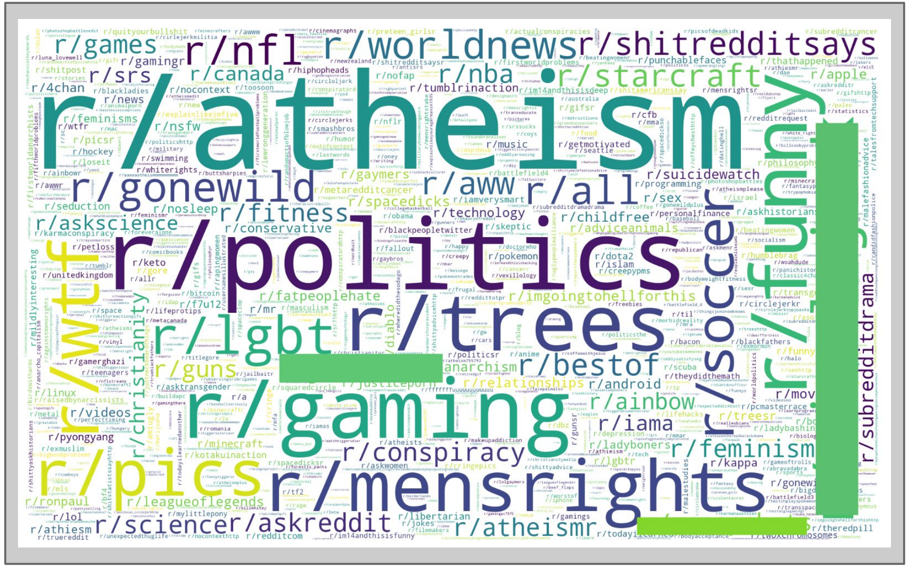
Predicting Reddit Comment Upvotes

FlatIron School- Module 4 Final Project

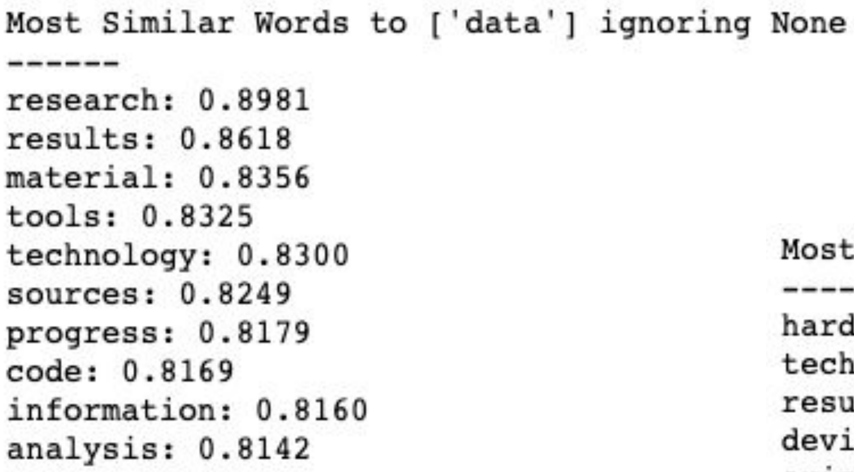
Purpose and Process

Can we predict Reddit Comment "scores" *from* the context *of* the comment?


r/atheism
r/politics
r/trees
r/funny
r/gaming
r/mensrights
r/wtf
r/worldnews
r/pics
r/bestof
r/adviceanimals
r/conspiracy
r/gonewild
r/askreddit
r/soccer
r/subredditdrama
r/shitredditsays
r/askscience
r/videos
r/science
r/games
r/nfl
r/aww
r/lgbt
r/nba
r/4chan
r/feminism
r/starcraft
r/all
r/ainbow
r/fatpeoplehate
r/relationships
r/getmotivated
r/imgoingtohellforthis
r/movies
r/christianity
r/leagueoflegends
r/music
r/makeupaddiction
r/hockey
r/pyongyang
r/todayilearned
r/android
r/atheismr
r/creepshots
r/nocontext
r/canada
r/nsfw
r/cringe
r/fitness
r/iama
r/programming
r/conservative
r/guns
[redacted]

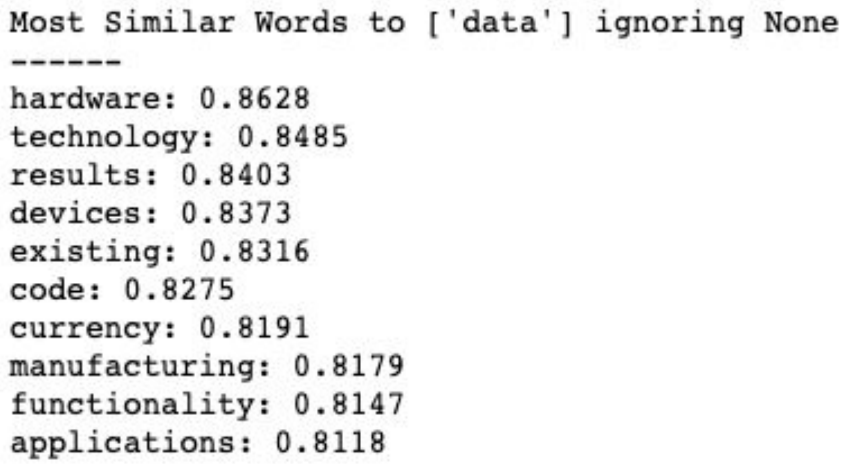

r/atheism
r/politics
r/trees
r/gaming
r/mensrights
r/worldnews
r/shitredditsays
r/nfl
r/games
r/starcraft
r/gonewild
r/aww
r/all
r/fitness
r/askscience
r/lgbt
r/wtf
r/pics
r/conspiracy
r/bestof
r/soccer
r/funny
r/subredditdrama
r/sciencer
r/askreddit
r/atheismr
r/feminism
r/ainbow
r/iama
r/guns
r/christianity
r/fatpeoplehate
r/imgoingtohellforthis
r/spacedicks
r/technology
r/childfree
r/adviceanimals
r/canada
r/nba
r/srs
r/nsfw
r/apple
r/leagueoflegends
r/anarchism
r/conservative
r/sex
r/suicidewatch
r/minecraft
[redacted]

Target Comments



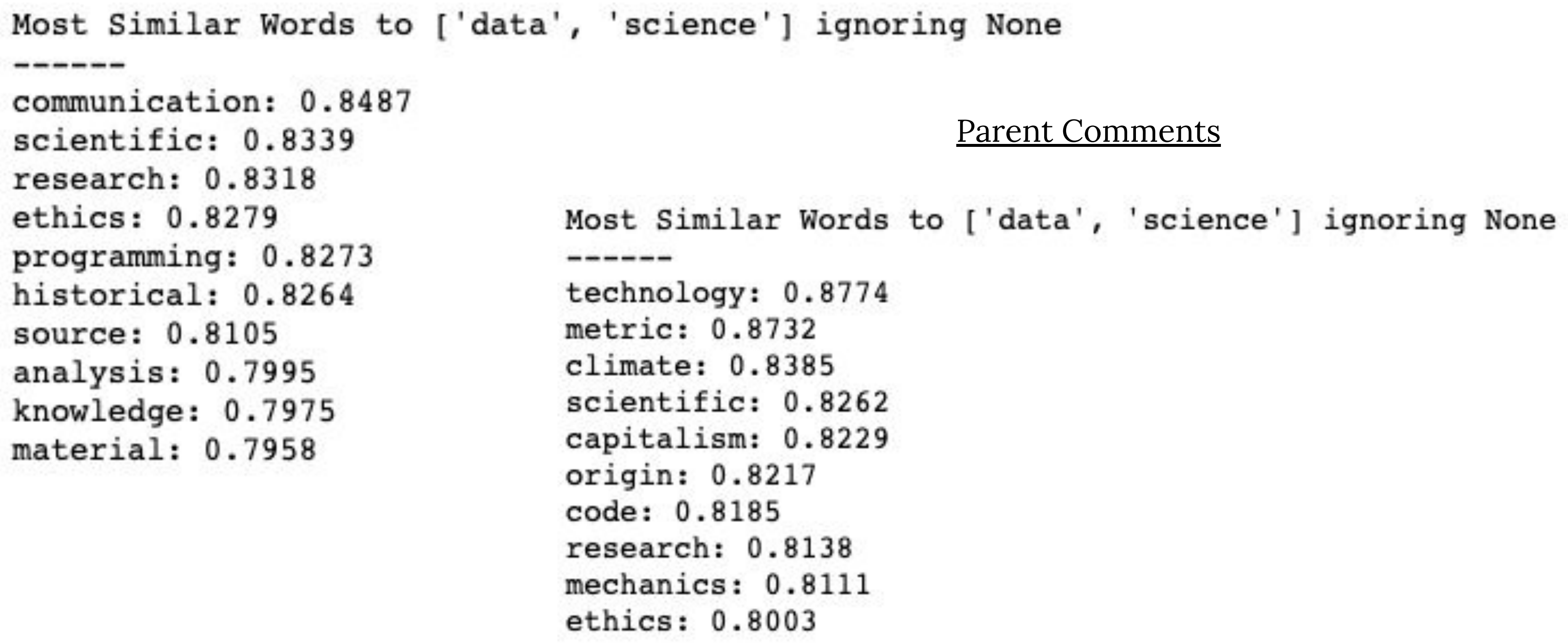
```
Most Similar Words to ['data'] ignoring None
-------
research: 0.8981
results: 0.8618
material: 0.8356
tools: 0.8325
technology: 0.8300
sources: 0.8249
progress: 0.8179
code: 0.8169
information: 0.8160
analysis: 0.8142
```

Parent Comments



```
Most Similar Words to ['data'] ignoring None
------
hardware: 0.8628
technology: 0.8485
results: 0.8403
devices: 0.8373
existing: 0.8316
code: 0.8275
currency: 0.8191
manufacturing: 0.8179
functionality: 0.8147
applications: 0.8118
```

Target Comments

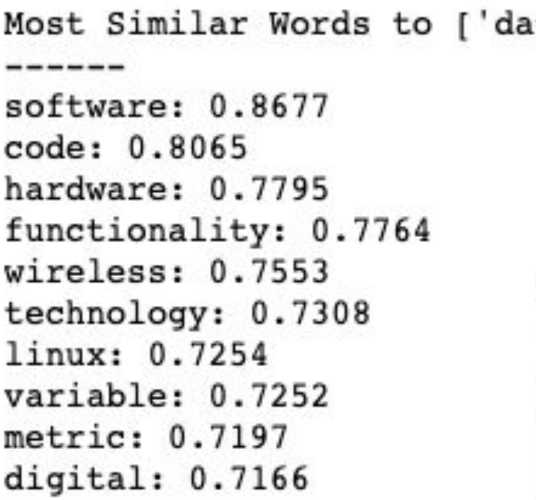


```
Most Similar Words to ['data', 'science'] ignoring None
-------
communication: 0.8487
scientific: 0.8339
research: 0.8318
ethics: 0.8279
programming: 0.8273
historical: 0.8264
source: 0.8105
analysis: 0.7995
knowledge: 0.7975
material: 0.7958
```

Parent Comments

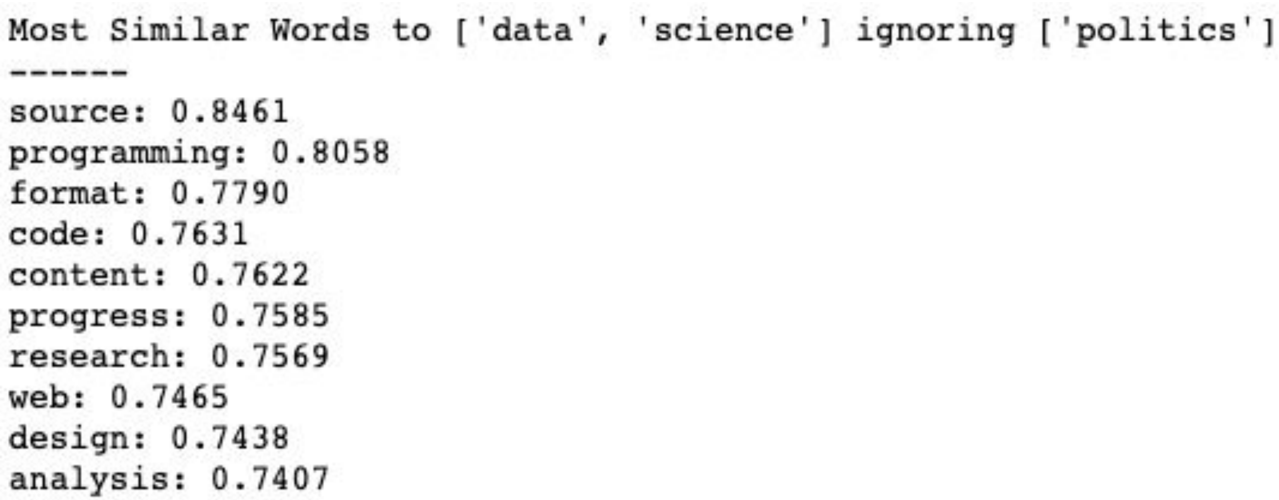
```
Most Similar Words to ['data', 'science'] ignoring None
-------
technology: 0.8774
metric: 0.8732
climate: 0.8385
scientific: 0.8262
capitalism: 0.8229
origin: 0.8217
code: 0.8185
research: 0.8138
mechanics: 0.8111
ethics: 0.8003
```

Target Comments



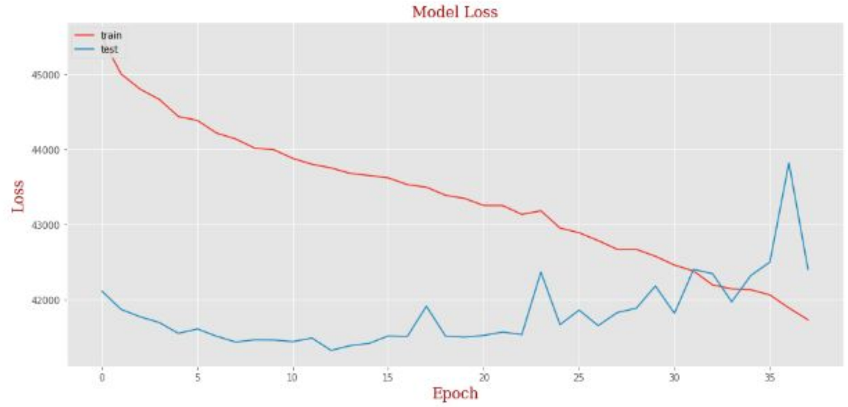
```
Most Similar Words to ['data', 'science'] ignoring ['politics']
------
software: 0.8677
code: 0.8065
hardware: 0.7795
functionality: 0.7764
wireless: 0.7553
technology: 0.7308
linux: 0.7254
variable: 0.7252
metric: 0.7197
digital: 0.7166
```

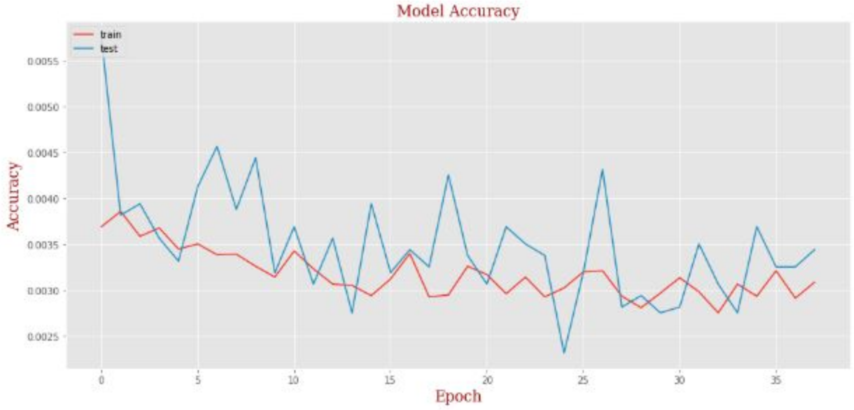
Parent Comments



```
Most Similar Words to ['data', 'science'] ignoring ['politics']
------
source: 0.8461
programming: 0.8058
format: 0.7790
code: 0.7631
content: 0.7622
progress: 0.7585
research: 0.7569
web: 0.7465
design: 0.7438
analysis: 0.7407
```

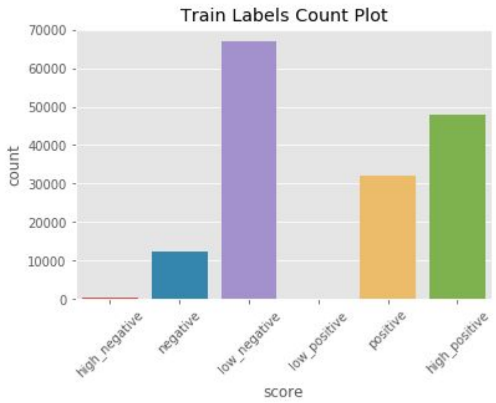
Linear Model

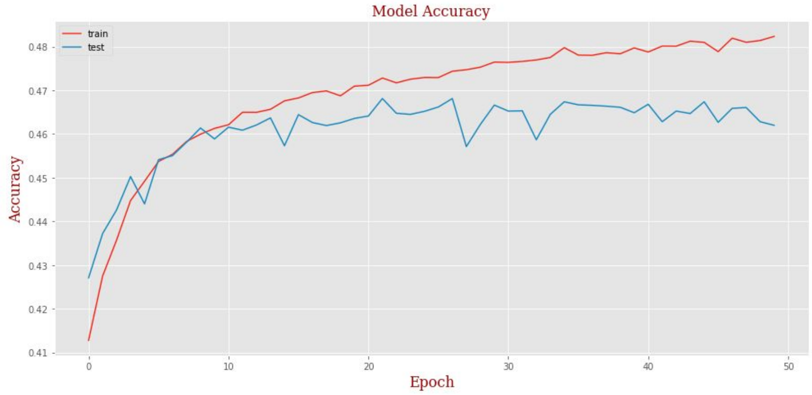

Model Loss

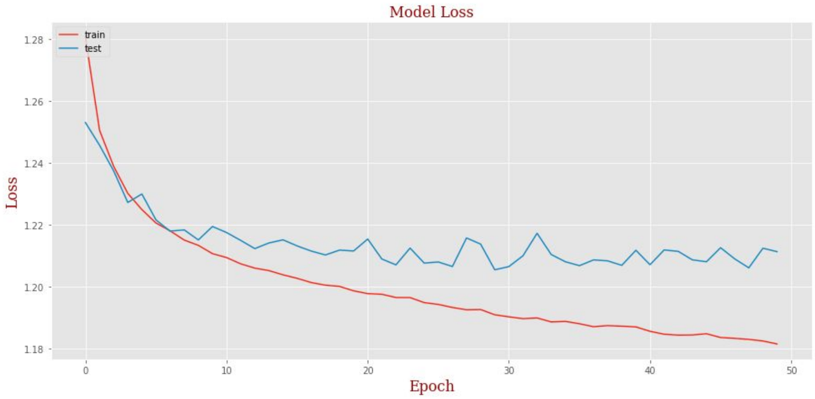

Model Accuracy

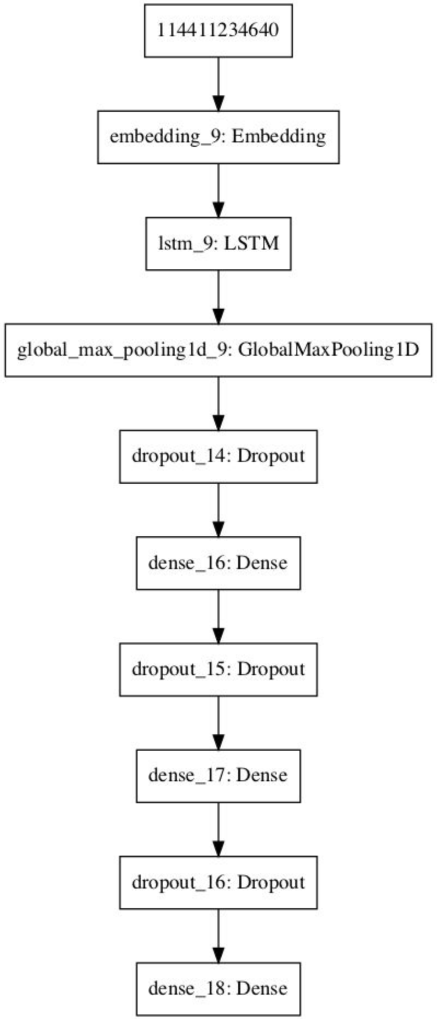
Categorical Model

≤ -100	High Negative
≤ -20	Negative
≤ 0	Low Negative
≤ 20	Low Positive
≤ 100	Positive
> 100	High positive









114411234640

embedding_9: Embedding

lstm_9: LSTM

global_max_pooling1d_9: GlobalMaxPooling1D

dropout_14: Dropout

dense_16: Dense

dropout_15: Dropout

dense_17: Dense

dropout_16: Dropout

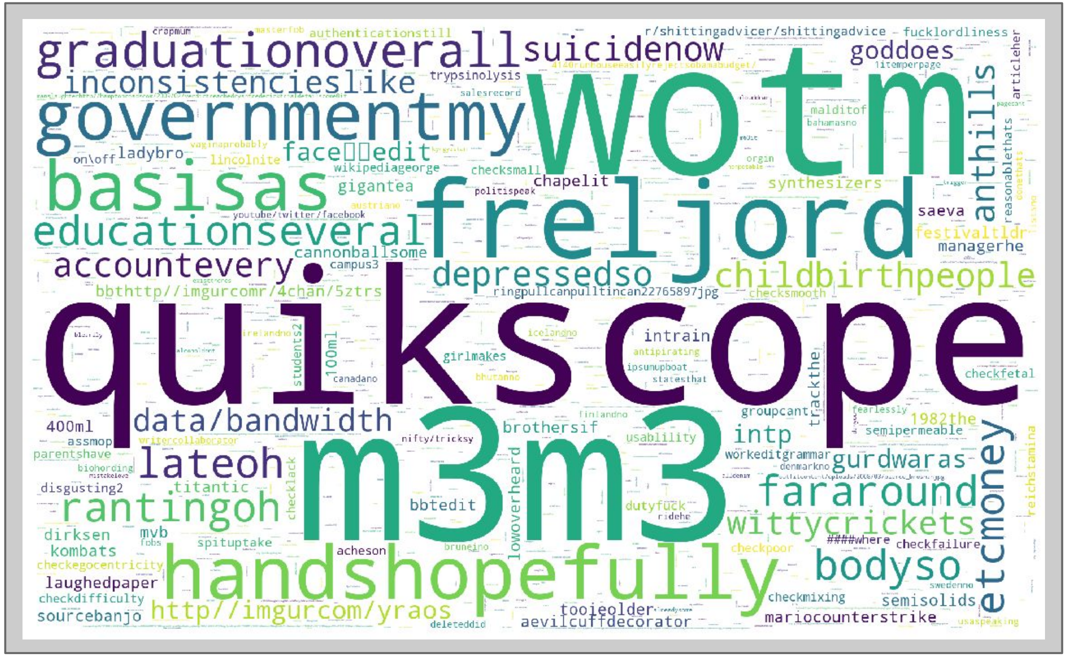
dense_18: Dense

Interpret

Model Loss: 1.1824
Model Accuracy: 48.23%

Validation Loss: 1.211
Validation Accuracy: 46.20%

Test Loss: 1.204
Test Accuracy: 47.14%



Interpret

Canada is in the data, but not on your chart. -2 internet points for you!

Current score = 125

High Negative: 0.18%
Negative: 7.06%
Low Negative: 36.3%

Low Positive: 0%
Positive: 22.5%
High Positive: 33.8%

Could there be a more useless, confusing, and stupid, representation of 'whatever the OP was attempting to display? Yeah, no.

Current Score= -25

High Negative: 0.15%
Negative: 5.9%
Low Negative: 30%

Low Positive: 0%
Positive: 24.7%
High Positive: 38.7%

Next Steps

1. Work to improve our cleaner function to improve our data.
2. Add more inputs to our model such as parent comments, subreddit and maybe parent score
3. Combine the texts for word embedding to improve my weights

Thank You!

Any questions?