

Subreddit Recommendation System

Table of Contents

01 Problem Statement

02 Data

03 Modeling Process

04 Recommendation System Walkthrough

05 Conclusion and Future Steps



01

Problem Statement

According to Reddit's blog they had 199 million posts in 2019 alone and approximately 138,000 active subreddits. With so much information, how do they make sure people are posting in the right subreddits? Well they don't. So, how can we ensure that posts are going into the proper subreddit's? Through a recommendation system!



02

Data

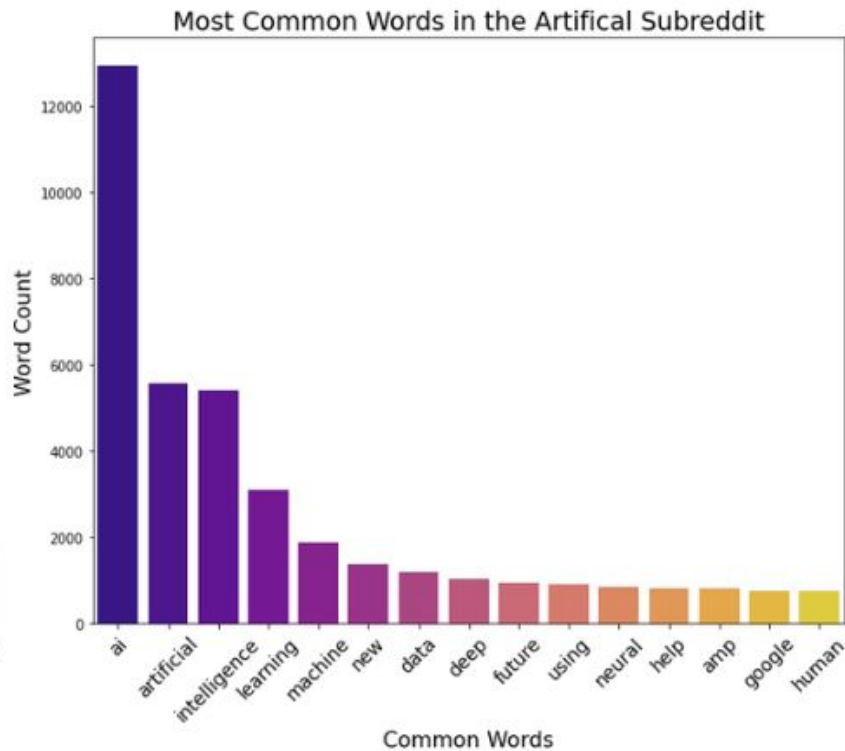
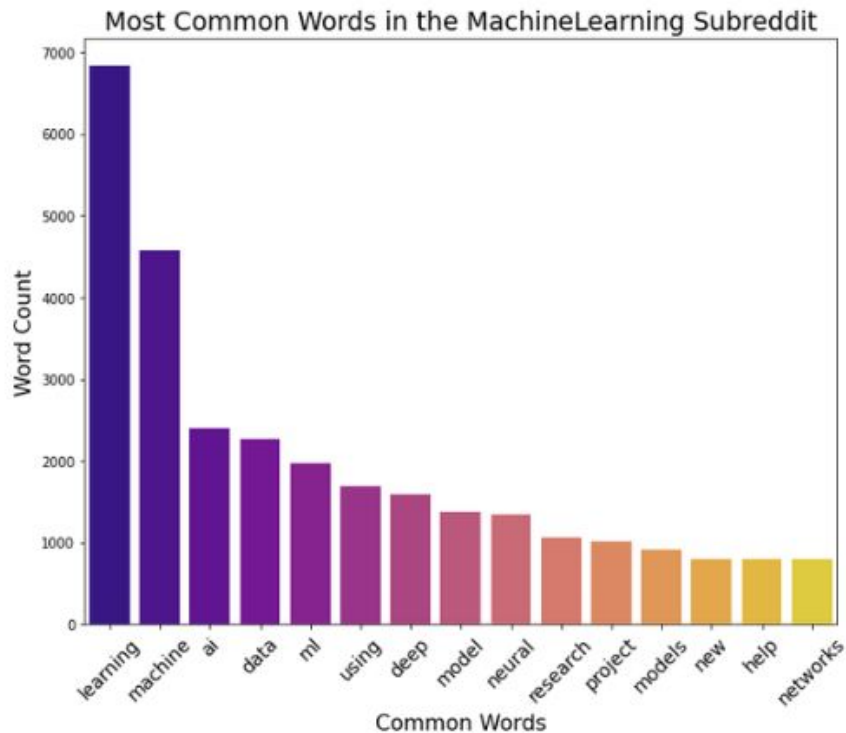
r/MachineLearning and r/artificial

Amount of Posts From Each Subreddit

31,299
r/MachineLearning

31,299
r/artificial

Top Words in Each Subreddit



03

Modeling Process

Model Selection

1

Naïve Bayes

86.15%

2

LSTM RNN

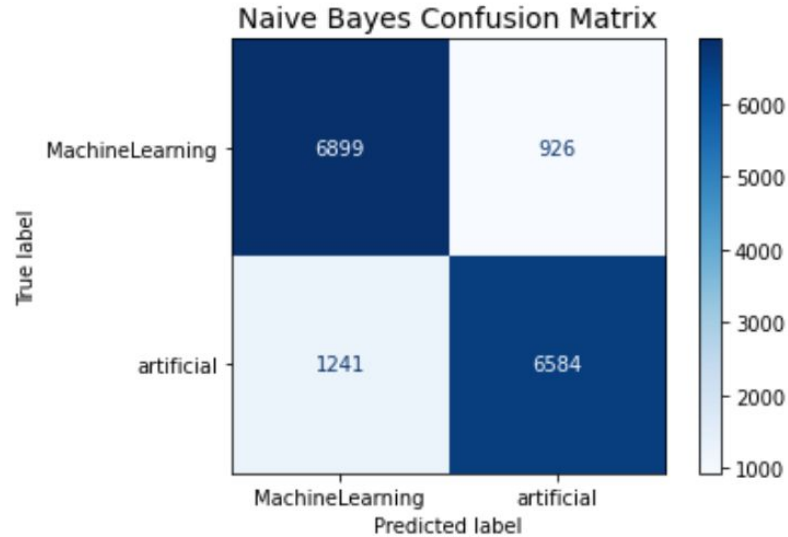
81.52%

3

**Logistic
Regression**

80.98%

Naïve Bayes Model



The Accuracy score is: 86.15%

The Missclassification rate is: 13.8%

The Sensitivity is: 84.14%

The Specificity is: 88.17%

The Precision is 87.67%

LSTM RNN as the Production Model

- RNN's have somewhat of a feedback loop which is it's "memory." This means that the past inputs leave a footprint in the network.
- An LSTM extends this idea by adding a short term and a long term memory to the network.
- LSTM RNN's are great for classifying sequential data such as text like reddit posts!

04

Recommendation System Walkthrough

**Kelly just
wrote this
great blog
post now she
wants to
post it on
Reddit**

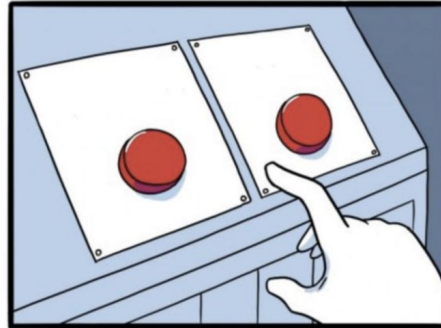
Decision Trees: Understanding the Basis of Ensemble Methods

A great base model for machine learning, but not a great final model.



Kelly Slatery Following

Mar 8 · 9 min read



JAKE-CLARK.TUMBLR

<https://imgflip.com/memegenerator/Two-Buttons>

Create a post DRAFTS 0

Choose a community

Post Images & Video Link Poll

Decision Trees: Understanding the Basis of Ensemble Methods 59/300

B *i* **A^** [Markdown mode](#)

Text (optional)

+ OC **+ SPOILER** **+ NSFW** **FLAIR**

CANCEL **POST**


☒ Send me post reply notifications

[Connect accounts to share your post](#) ⓘ

She begins writing her post but doesn't know which subreddit to post it in.












Instead of Kelly guessing the best subreddit for her post this recommendation system gave her a suggested subreddit based on the post she is writing.

Create a post DRAFTS 0


Choose a community  r/MachineLearning
1,256,130 members

[Post](#) [Images & Video](#) [Link](#) [Poll](#)

Decision Trees: Understanding the Basis of Ensemble Methods 59/300


B *i*    **A^**         [Markdown mode](#)

Text (optional)

[+ OC](#) [+ SPOILER](#) [+ NSFW](#) [FLAIR](#) 

[CANCEL](#) [POST](#)

☒ Send me post reply notifications

[Connect accounts to share your post](#) 

Old Without Suggestions

Create a post DRAFTS 0

Choose a community ▼

Post Images & Video Link Poll

Decision Trees: Understanding the Basis of Ensemble Methods 59/300

B *i* **A** [Markdown mode](#)

Text (optional)

☐ OC ☐ SPOILER ☐ NSFW ☐ FLAIR ▼

☒ Send me post reply notifications
[Connect accounts to share your post](#) ⓘ

New With Suggestions

Create a post DRAFTS 0

Choose a community ▼ **r/MachineLearning**
1,256,130 members

Post Images & Video **Link** Poll

Decision Trees: Understanding the Basis of Ensemble Methods 59/300

B *i* **A** [Markdown mode](#)

Text (optional)

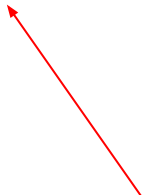
☐ OC ☐ SPOILER ☐ NSFW ☐ FLAIR ▼

☒ Send me post reply notifications
[Connect accounts to share your post](#) ⓘ

How the Model Classified Her Post

```
▶ new_post = ["Decision Trees: Understanding the Basis of Ensemble Methods"]  
  seq = tokenizer.texts_to_sequences(new_post)  
  padded = pad_sequences(seq, maxlen=MAX_SEQUENCE_LENGTH)  
  pred = model.predict(padded)  
  labels = ['MachineLearning', 'artificial']  
  print(pred, labels[np.argmax(pred)])
```

```
↳ [[0.73572147 0.26427853]] MachineLearning
```



**73% probability this post would
belong to r/MachineLearning**

05

Conclusions and Future Steps

Conclusions

Adding this tool has a lot of benefits such as:

- Cleaner subreddits
- More accurate advertising
- Reducing the workload for reddit moderators
- Easier posting for users
- Accumulating massive amounts of text data

Although I got a pretty decent accuracy scores for these two subreddits it is very important to remember that this model was only trained on only these two subreddits.

Future Steps

- Instead of only using the title for the recommendation system, the model will include the content of the post as well. This will add a lot more data to the model.
- I will continue to optimize the LSTM RNN and finding the best hyperparameters for this problem.
- I will be scraping 2,000 posts from the active 138,000 subreddits, adding up to 276 million posts. This should be enough data to begin the steps of implementing a full subreddit recommendation system.

Questions?