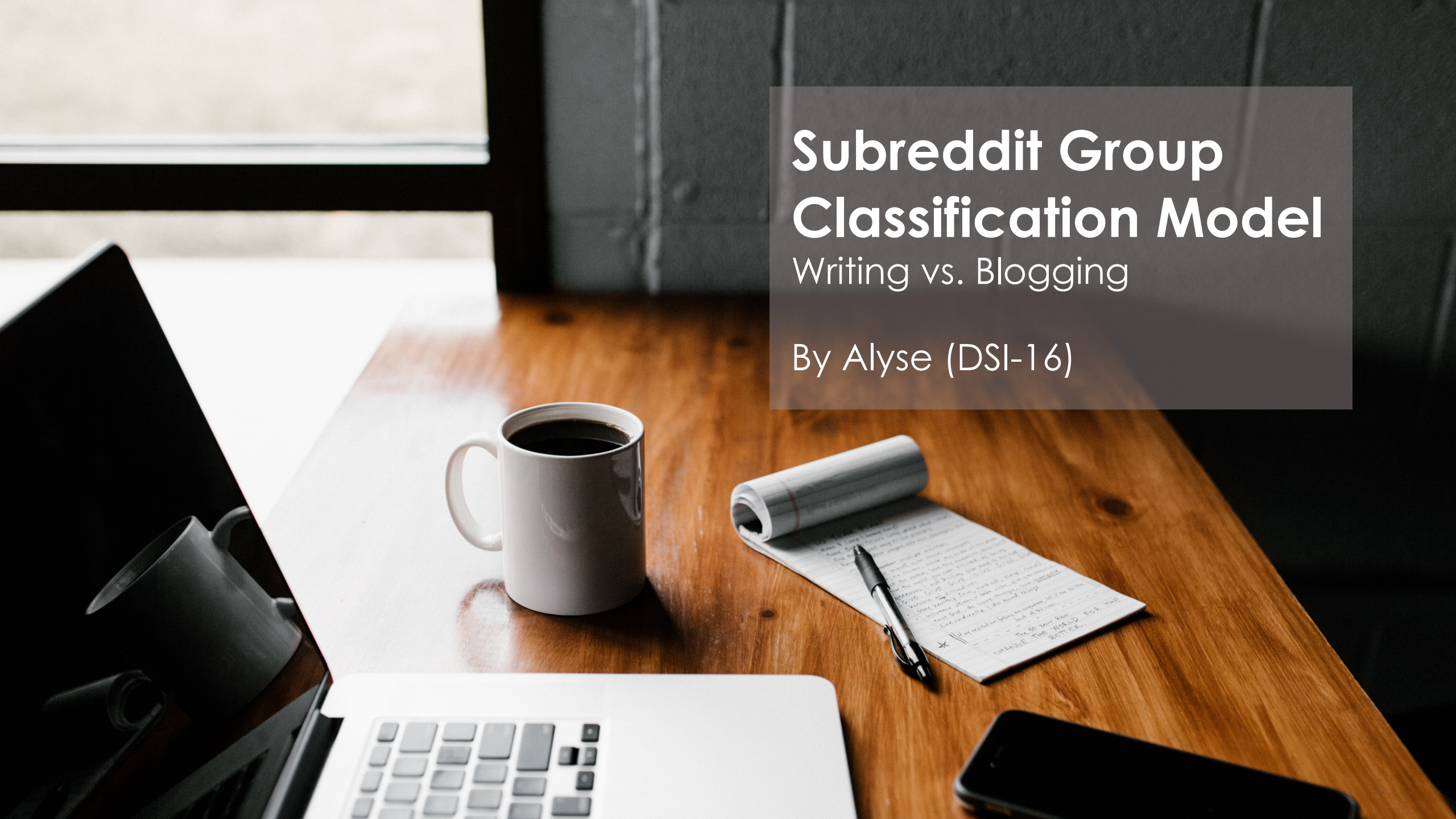


Subreddit Group Classification Model

Writing vs. Blogging

By Alyse (DSI-16)

Introduction

Providing context to our business problem

Key Clients

- Aspiring Bloggers

Pain Points

- Writing technique/style
- Blogging content
- Increasing traffic flow
- Optimizing SEO
- …. *And many more!*

Troubleshoot

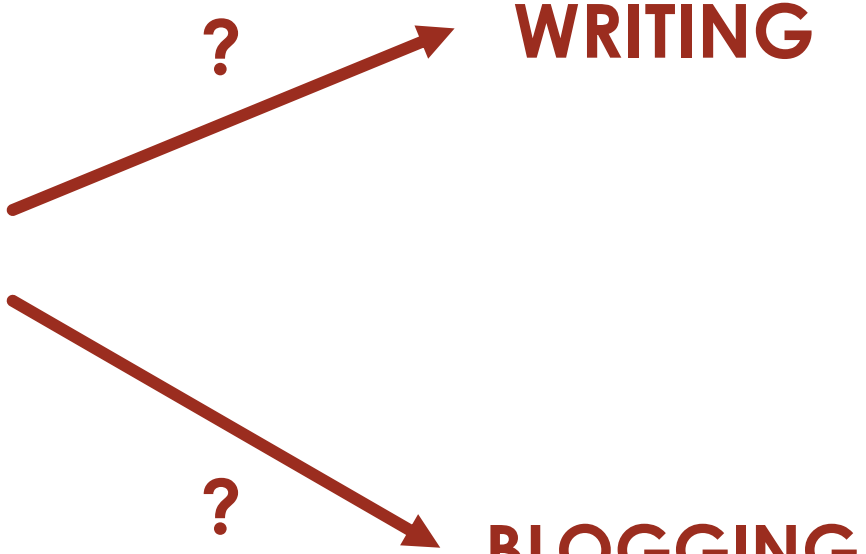
Many different platforms that bloggers seek help. One common platform that is very active and popular in the West is **Reddit.**

Business Case

Text classifier tool to optimize allocation of reddit posts to "Writing" or "Blogging"

Pain Points

- Writing technique/style
- Blogging content
- Increasing traffic flow
- Optimizing SEO
- …. And many more!



About Community

Discussions about the writing craft.

1.4m
Members

1.1k
Online

Created 25 Jan 2008

About Community

A community for bloggers. This subreddit is aimed towards helping bloggers with their blogging journey.

67.8k
Members

96
Online

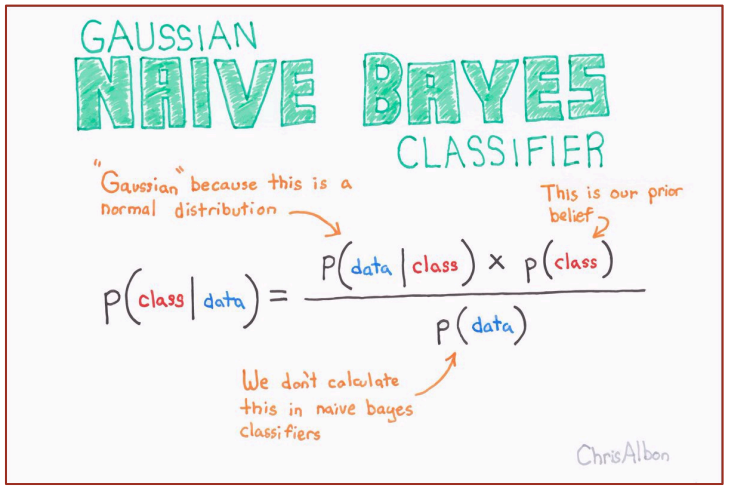
Created 18 Mar 2008

Data Science Approach

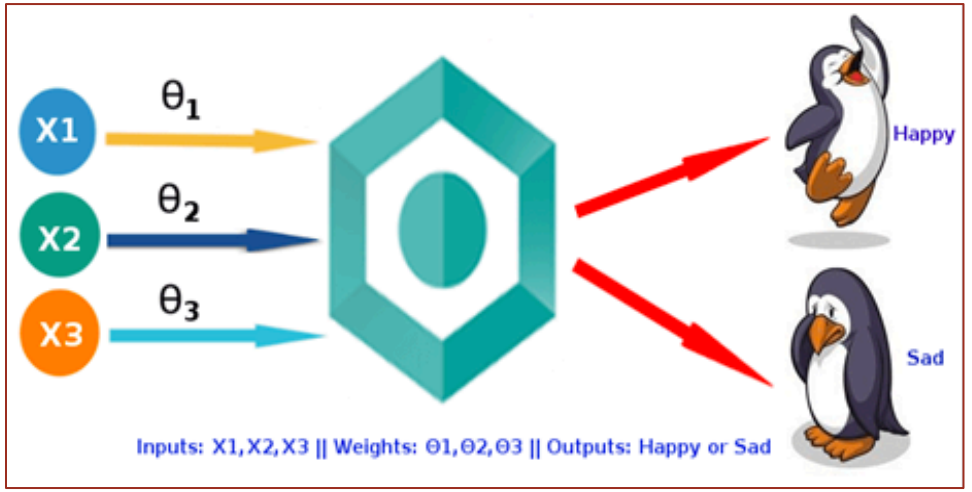
How do we classify our reddit posts to the RIGHT subreddit group?

Two Classification Model
(Machine Learning, Supervised Learning)

Multinomial Naïve-Bayes Model



Logistic Regression



Data Collection

Extracting subreddit posts from Blogging and Writing subreddit groups

- **Approach:** Web-scrapping through Reddit API
- **Types**: Hot Posts and New Posts
- **Time period**:
    - Blogging: 25th June to 5th September 2020
    - Writing: 21st August to 5th September 2020
- **Number of Unique Posts**:
    - Blogging: 565
    - Writing: 668

Exploring Data

Taking a first stab at the raw text data with a word cloud visualization

First Impressions

Blogging	Writing

• Website analytics oriented
  • (website, google, keyword, social media etc.)
• Professional community
• Marketing/ Promotion driven
	• Traditional writing-oriented questions
	  • (character, book, story, novel etc.)
	• Friendly community
	• Appear to share more experiences/ seek help more
	• Focus on the art of writing and story building.

Blogging Title

Writing Title

Blogging Text

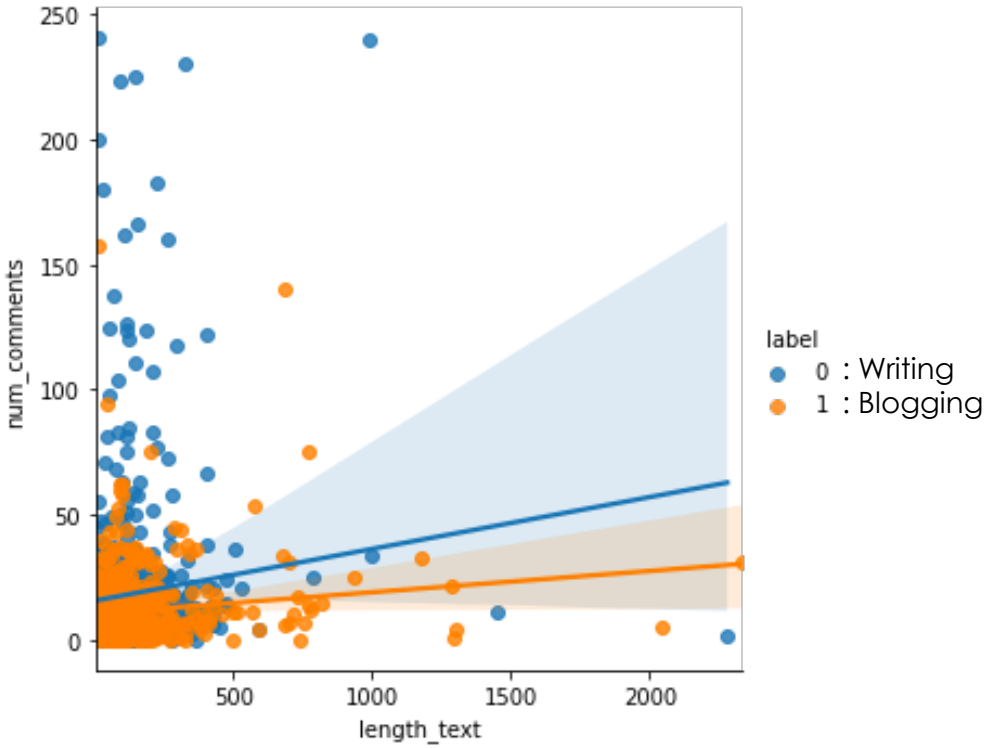
Writing Text

Exploring Data

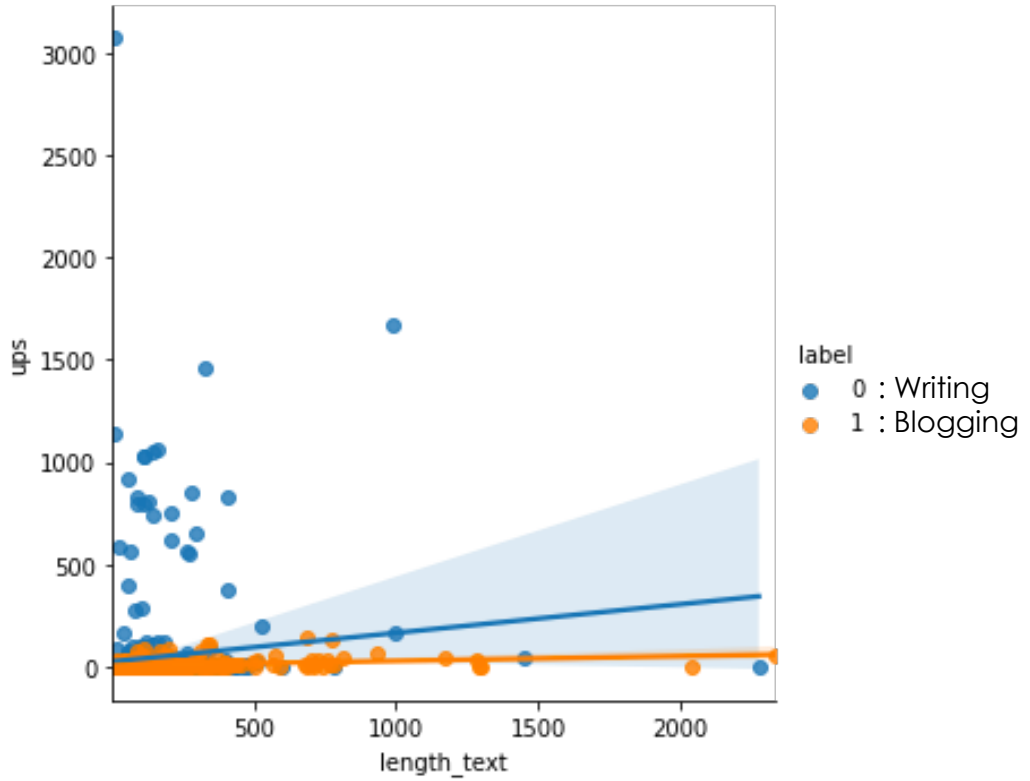
Will the length of your post impact your number of comments and upvotes?

Shorter posts with less words would have higher possibility of higher upvotes & comments!

Text Length vs. No. Comments

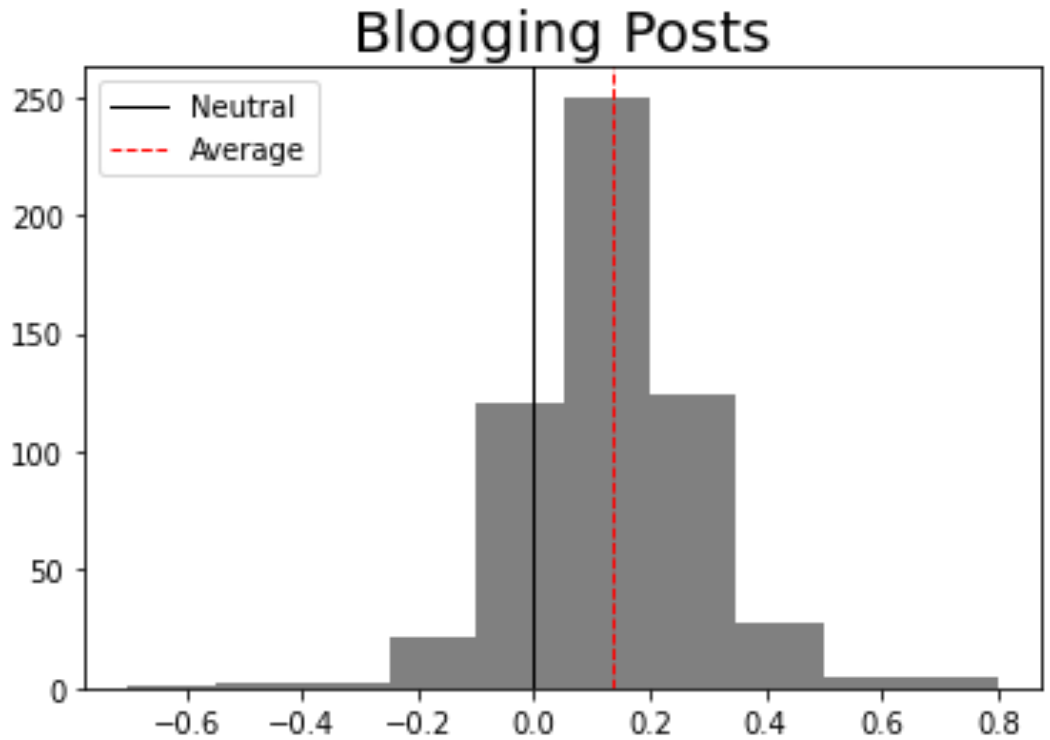


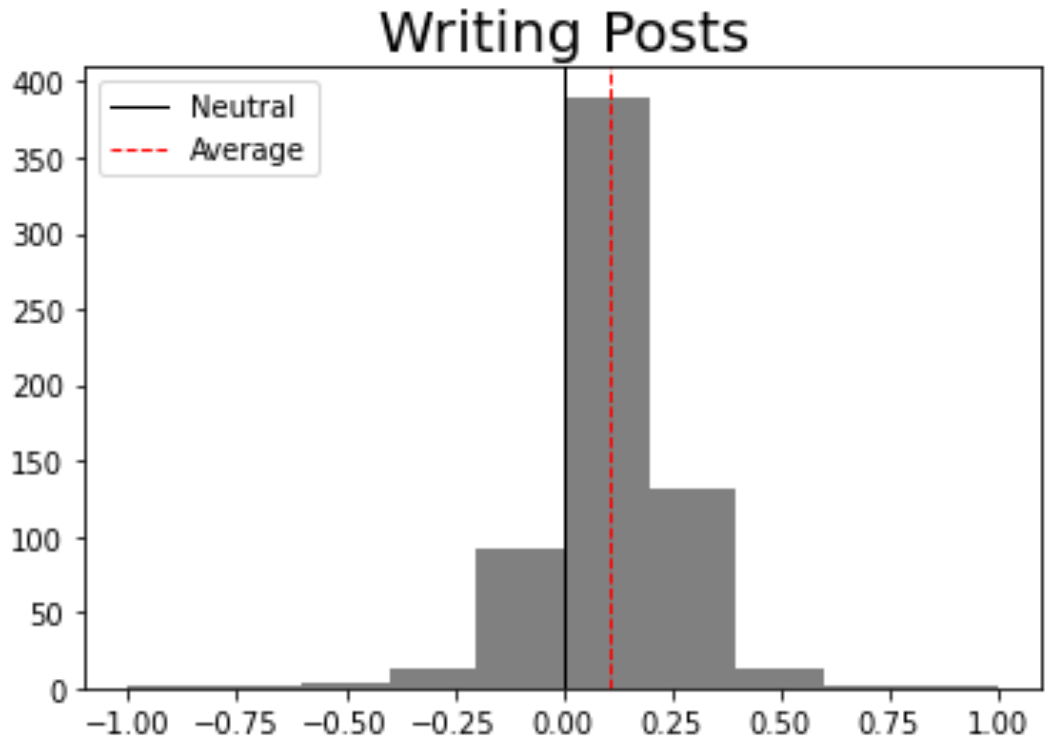
Text Length vs. No. Comments



Exploring Data

Determining the sentiments behind the Blogging and Writing subreddits

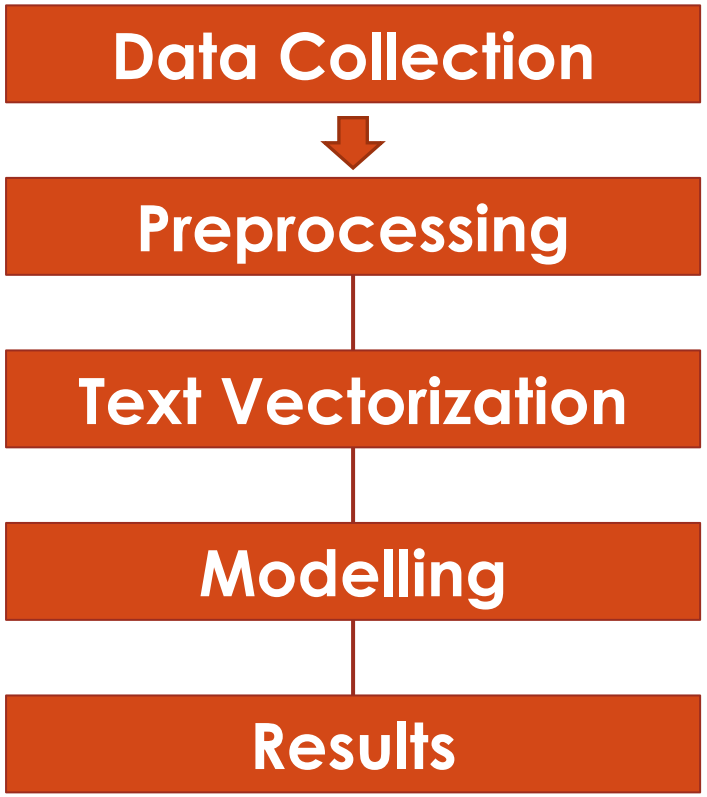




0.135 > 0.107

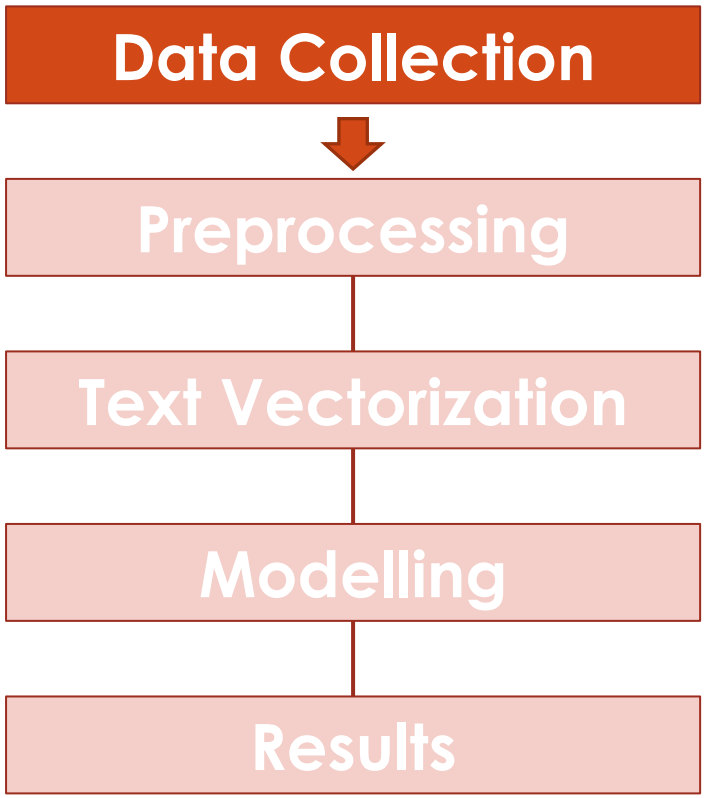
Modelling

Creation of classification model - Process workflow



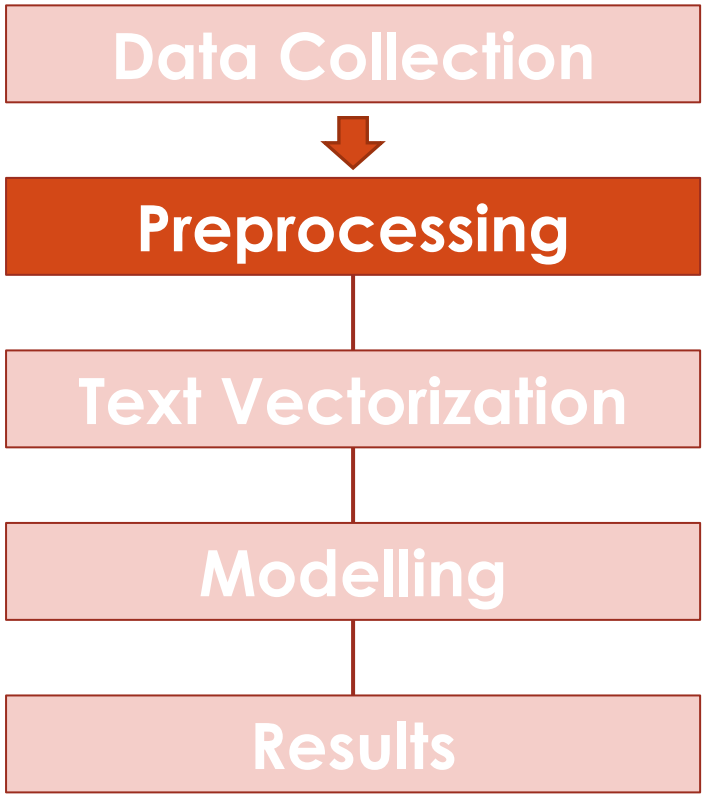
Modelling

Creation of classification model - Process workflow



Modelling

Creation of classification model - Process workflow: Preprocessing



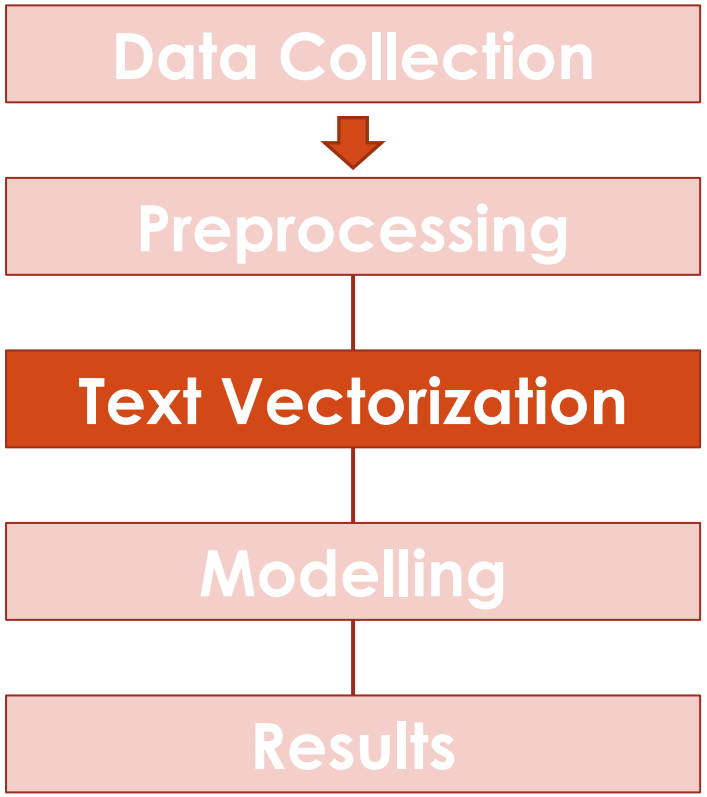
1. Cleaning (Removal of HTML etc.)
2. Stemming
3. Addition of Stop Words

```
557    I can spend two hours going over the same two p
aragraphs again and again rather than moving on with m
y writing. I end up frozen because I feel like what I
have isn't 'good enough,' and then I never continue wr
iting. Or, I spend a lot of time polishing paragraphs
that I later realize will have to be cut anyway, so i
t's a huge waste of time. Do any other people struggle
with this? Any tips for just letting yourself go and k
nocking out that imperfect first draft without nitpick
ing and worrying and getting yourself stuck in a perfe
ctionist frenzy?
```

```
'spend hour paragraph end frozen becaus feel isn t goo
d continu spend lot time polish paragraph later realiz
cut s huge wast time ani peopl struggl ani tip let kno
ck imperfect draft nitpick worri stuck perfectionist f
renzi'
```

Modelling

Creation of classification model - Process workflow: Text Vectorization



1. Count Vectorizer
2. TfidfVectorizer

CountVectorizer

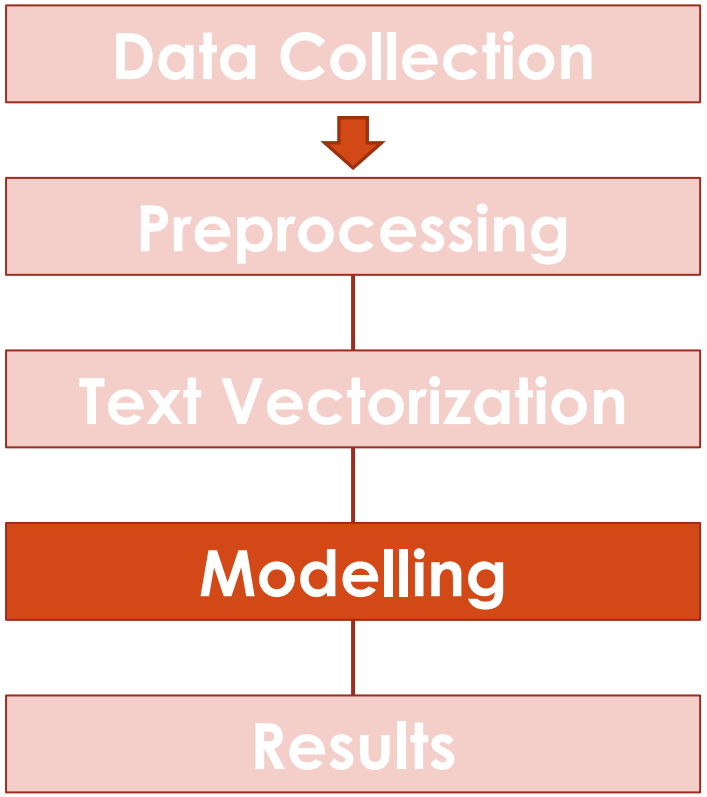
	observations
stori	629
want	511
ani	504
start	465
charact	461
use	459
make	449

TfidfVectorizer

	Most Common Terms
stori	30.855989
charact	27.199379
ani	24.951076
want	24.462039
post	24.344080
start	22.793590
know	21.422468

Modelling

Creation of classification model - Process workflow: Modelling



Naïve-Bayes & Logistic Regression

1. Train-test-split
2. Cross-validation

Chosen TfidfVectorizer for my vectorized text → Higher scores

3. Grid search (Finding best model)

Naïve-Bayes

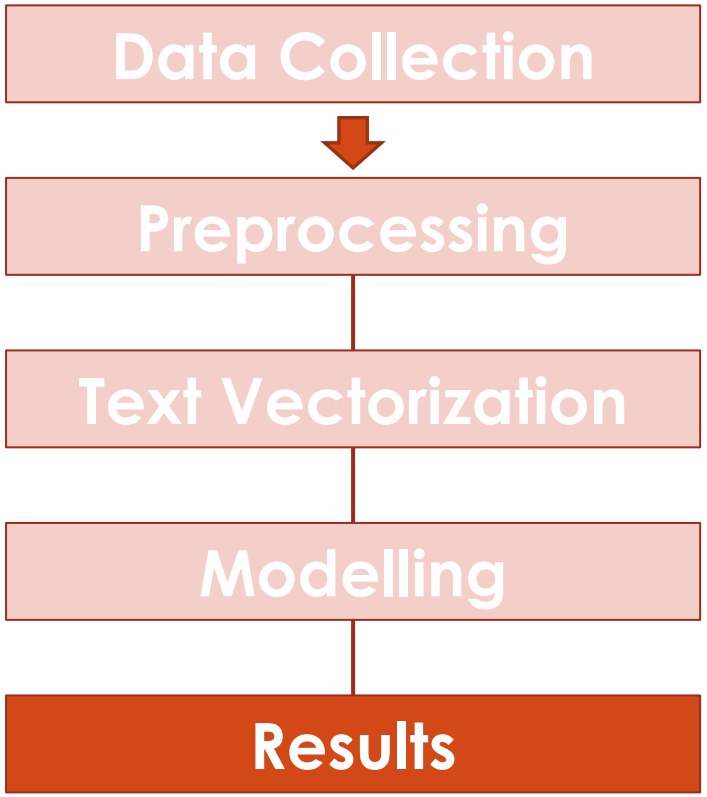
- Max_df = 0.9
- Max_features = 2500
- Min_df = 2
- Ngram_range = (1,1)

Logistic Regression

- Max_df = 0.9
- Max_features = 2500
- Min_df = 2
- Ngram_range = (1,1)

Modelling

Creation of classification model - Process workflow: Results



	Specificity	Sensitivity	Accuracy	ROC
Logistic Regression Model	0.9297	0.9310	0.9303	0.984173
Naive-Bayes Regression	0.9609	0.9052	0.9344	0.987742

Model	Train	Test
Logistic Regression	**0.9866**	**0.9303**
Naive-Bayes	0.9723	0.9344

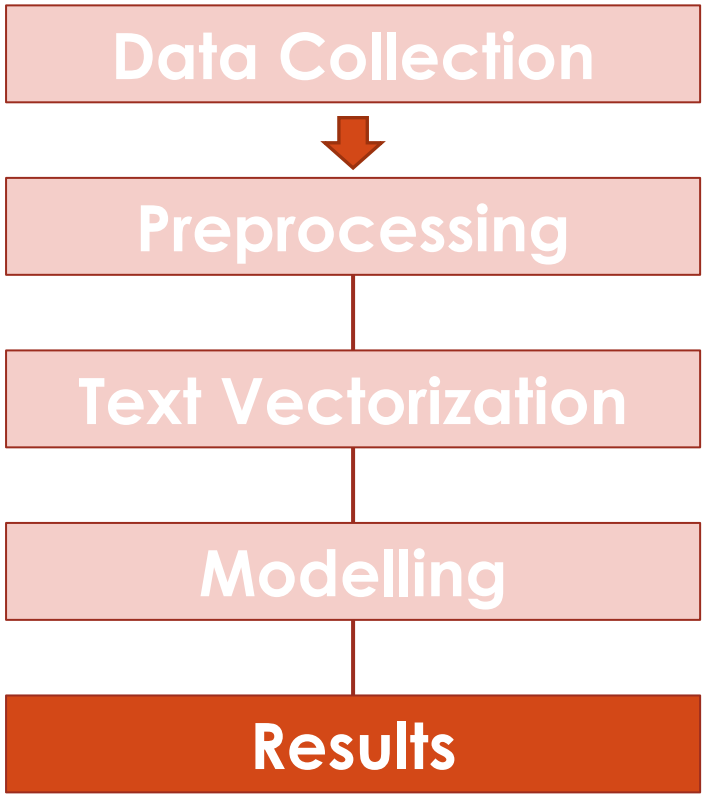
Both models achieved

~93% accuracy

in classifying our posts into the correct subreddit groups

Modelling

Creation of classification model - Process workflow: Results



Misclassification (Reasons)

- Ambiguous posts

```
21      Don't get me wrong, I do love them. It must be all of
the imagination needed to keep the chorography straight in my
head.\n\nAnyone else have something about writing that takes
a lot out of them? How do you push through?
```

- Short posts

```
35      Comments are betting than the actual posts
```

- Blogging posts Misclassified (As writing):

```
255     I am having difficulty on how to organize my pages and I don't understand yet how subc
ategories work for categories. and also tags. \n\nI plan to blog about books and film reviews
and include some personal thoughts and freewrittten poems or essays. Might include my own sho
rt stories.\n\n\nPS.\nHow many categories and subcategories is too much normally for a blog?
```

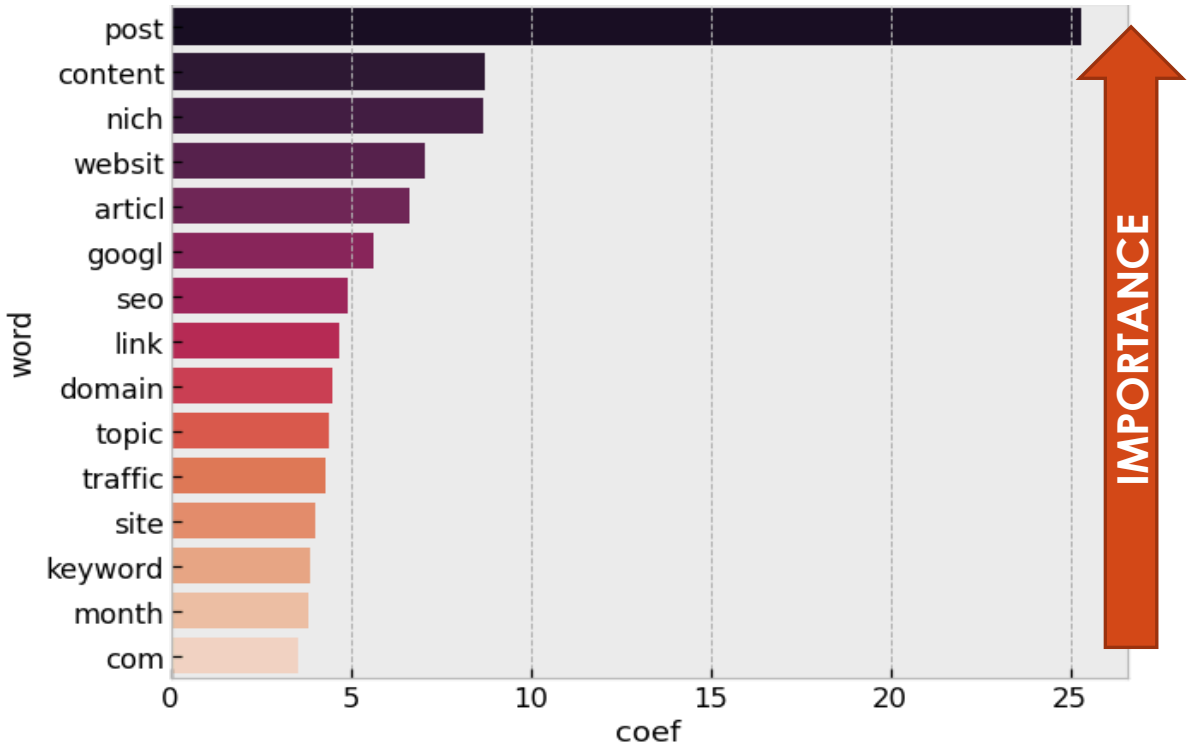
- Writing posts Misclassified (As blogging):

```
256     I ha e been wanting to try starting a blog for a few different rea
sons. One I really interested learning the aspects of seo and digital med
ia marketing. Trying out affiliate  marketing. Possibly learning how to m
ake money online. Overall I feel like it would be a really fun and intere
```

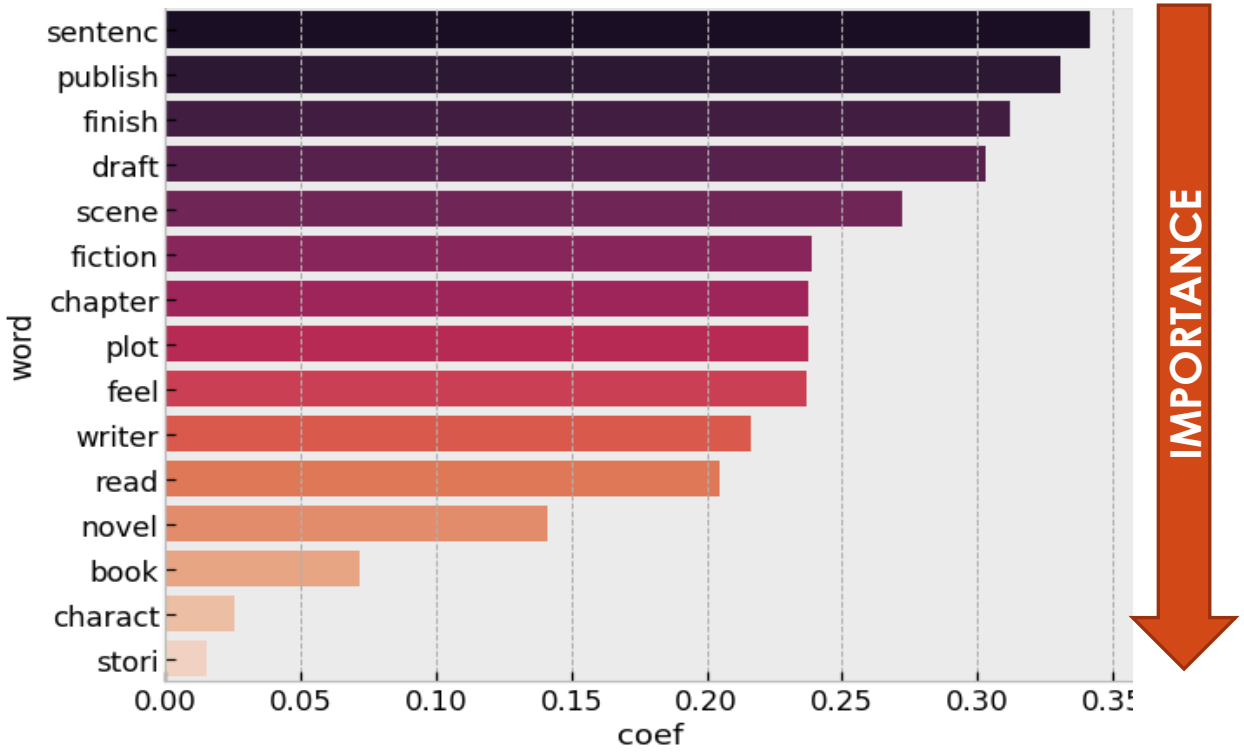
Modelling: Logistic Regression

Which words are more important in classifying our subreddit posts?

BLOGGING



WRITING



Conclusions

Our key takeaways…

Both Naïve-Bayes and Logistic Regression gives similar results in predictability of our test dataset with similar accuracy and model scores.

How we can improve on our model?

- With more time, we would like to scrape at least 5,000 posts from Reddit, so that we can have higher volume to increase our model's training test data.
- Moreover, we hope to scrape an addition 500 posts to set aside as a Test dataset.
- To improve the model by adding Title into our analysis.
- Look into misclassified words and decide if I should remove them.