

# Project 3

Classifying Reddit Posts by NLP

---

Tan Si Hao

Zack Chia

Goh Wen Xuan

# Problem Statement / Business Case

We are developing a new wellness app that prompts the User to write a short journal entry, which will be analysed to determine the User's philosophical inclination and return a relevant message or thought of the day.

Our app focuses on two philosophical beliefs - Stoicism and Buddhism.

To understand the topics from each philosophical group, data is extracted from two subreddits - r/Stoicism and r/Buddhism. We will use the data to attempt to train a classifier model to predict the User's philosophical inclination based on the journal entry.

We believe we can adapt the model in the future into other subreddits.

# Process

- Data Extraction
- Preprocessing
- EDA
- Modelling and Model Selection
- Model Evaluation
- Conclusions

# Data Extraction

## pushshift.io API

- Flexible API, able to extract and search comments and submissions with various parameters
- Data extracted from r/Buddhism and r/Stoicism
- Approximately 10,000 posts were extracted from each subreddit
- Limited to 100 posts per request, multiple requests handled using for-loop

# Preprocessing

## Cleaning Data

### Remove special characters, symbols, punctuations and emojis

- []@#\$\$%^&, etc

### Remove unwanted words

- www, https://, com

### Retain only english characters

## Binary Classification

Stoicism : 1 and Buddhism : 0

title	selftext
Can not cohabiting be considered stoicism?	If a segment of stoicism is choosing not to be affected by things that are uncontrollable and focusing on things that are then can not cohabiting be considered a stoic stance as opposed to fighting the un-winnable fight of changing biased laws?
Ancient Stoic astronomer Cleomedes gets a shout-out in /r/AskHistorians	<p>[The question] (<a href="https://www.reddit.com/r/AskHistorians/comments/os4aiw/during_his_lifetime_julius_caesar_had_been_to/">https://www.reddit.com/r/AskHistorians/comments/os4aiw/during_his_lifetime_julius_caesar_had_been_to/</a>) During his lifetime, Julius Caesar had been to both Egypt and Britain on military campaign. In summer months, a day in London lasts about 2hrs longer than one in Alexandria. Did Caesar or any of his contemporaries have an explanation for this? Did they even notice it?</p> <p>The top answer begins: Yes, this was a well known phenomenon and was correctly understood as caused by the spherical shape of the earth, and the seasonal variation as caused by the angle between the earth's equator and the plane of the ecliptic. The most to-the-point discussion in ancient sources is in Cleomedes' "On the heavens", ch. 1.4: he was writing later than Caesar, but his work is based on material going back to long before Caesar's time.</p> <p>There's some debate over when [Cleomedes] lived, though he was clearly influenced by [Posidonius]. Cleomedes is featured prominently in [Alexander Jones]'s chapter ("The Stoics and the Astronomical Sciences") in the "Cambridge Companion to the Stoics".</p>

*Symbols and unwanted text in extracted posts*

# Preprocessing

## Tokenization

Separate pieces of text into smaller pieces based on word.

## Lemmatizing

Normalise text into their root form  
(i.e. playing, plays, played -> play)

Why not Stemming?

- While Stemming might be faster to process, it cuts off the word without knowing the context of the word, resulting in less accuracy.
- Given that both subreddits are text-heavy, context is important and the meaning of the word should be preserved as much as possible.
- The dataset is relatively small hence the performance difference is negligible

if karma carries over from life to life what is the  
point of death it seems to me that death and its  
process would serve as a clean slate obviously this is  
wrong view but why



Tokenization

[if, karma, carries, over, from, life, to, life, what, is, the,  
point, of, death, it, seems, to, me, that, death, and, its,  
process, would, serve, as, a, clean, slate, obviously, this, is,  
wrong, view, but, why]



Lemmatizing

[if, karma, carry, over, from, life, to, life, what, is, the, point, of,  
death, it, seems, to, me, that, death, and, it, process, would,  
serve, a, a, clean, slate, obviously, this, is, wrong, view, but,  
why]

# Preprocessing

## TfidfVectorizerization

Transforms the text into a vector based on the number of times a word appears in the text to the total number of words in the text.

The parameter 'stop\_words' is added to remove stop words during the process.

- Stop words are words that occur frequently but does not provide any useful information (i.e. 'and', 'its', 'will', 'this', etc)
- Additional stop words are added to further reduce any unwanted words (noise)

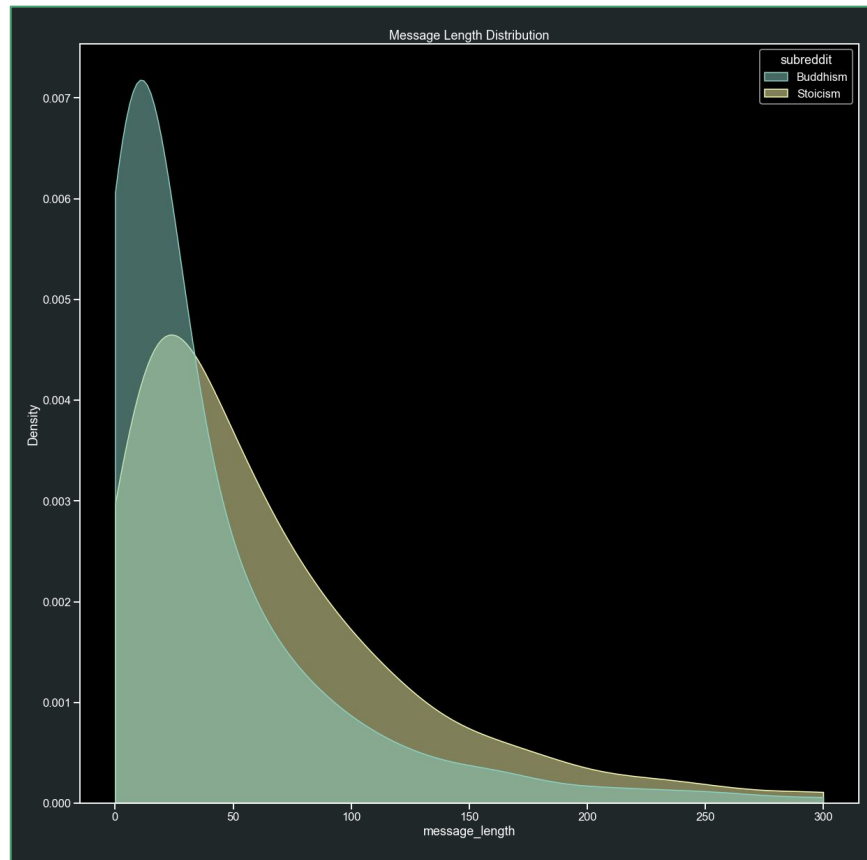
```
add_stop_words = ['wa', 'im', 'dont', 'ha', 'ive', 'doe']  
stop_words = text.ENGLISH_STOP_WORDS.union(add_stop_words)
```

Adding words into stop\_words

# EDA

## Message Length

- Buddhism post length tend to be lower than their Stoic users.





# EDA

## GenSim Topic Extraction

- Purpose: Helps us understand our target audiences better and what topics are relevant to them
- Latent Dirichlet Allocation (LDA) algorithm
- After tokenisation, we have to create a dictionary(maps words and their integer ID) and a corpus to fit into our model.
- Identify best number of topics based on coherence score.

# Topic Modelling Results

## r/Buddhism

Dominant_Topic	Num_Posts	Perc_Posts
0	1	1898
1	2	1648
2	6	1587
3	11	1490
4	0	1076
5	10	635
6	14	450
7	8	434
8	9	229
9	4	75
10	12	68
11	7	67
12	3	58
13	5	49
14	13	43

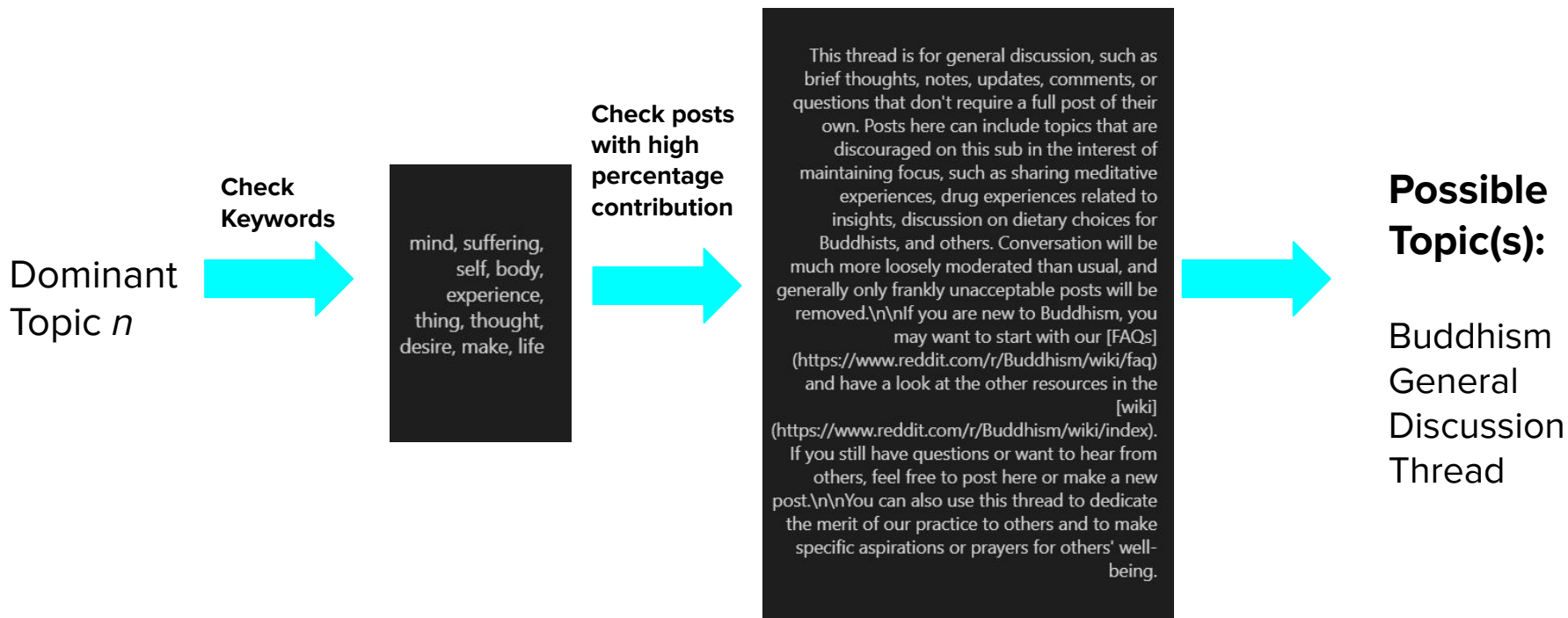
Top 5 topics occupy  
~75% of posts

Top 3 topics occupy  
~95% of posts

## r/Stoicism

Dominant_Topic	Num_Posts	Perc_Posts
0	0	4614
1	9	2890
2	5	1987
3	8	232
4	10	25
5	6	24
6	11	18
7	3	12
8	7	11
9	1	10
10	2	8
11	4	4

# General flow to understand topics classified



# Topics Buddhism subreddit users talk about

Religion vs Buddhism

I want to learn more  
about Buddhist  
Philosophy and  
practices

Need help cultivating  
long-term Buddhist  
habits

Moral Dilemmas and  
Philosophical  
Questions?



# Topics Stoicism subreddit users talk about

Quotes to help  
enlighten others

Relationship  
problems / Lifestyle  
changes. Need  
advice

New to Stoicism.  
Resources to  
recommend?



# Summary

## Similarities

- Attract intellectually curious people
- Users gravitate to the subreddits as they seek help to apply the respective philosophies to their life circumstances

*Insight: Possible target audience when marketing our app*


## Differences

- Control and Positive Thinking for Stoic users, Meditation and Rituals for Buddhism
- Buddhist users are less verbose versus Stoic users.
- Buddhism users has a tendency to focus on religion compared to Stoicism users
- Buddhism users are more outward looking, and more 'general' in their thinking (e.g. questions about the world). Stoic users are more introspective and personal, focusing on one's own circumstances.

*Insight: Possible differences that can help distinguish between the two audiences during our classification process*

# Three criteria are used for model selection

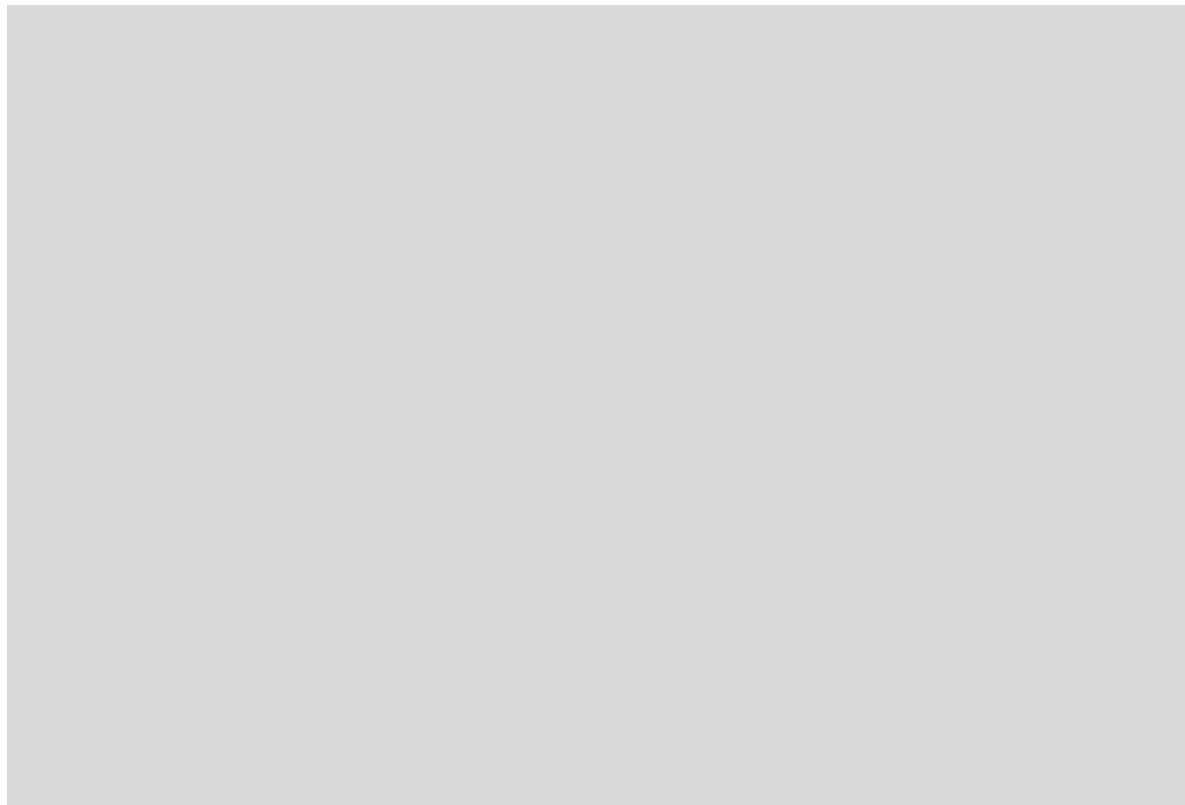
Model	Accuracy
Light Gradient Boosting Machine	0.8442
Extra Trees Classifier	0.8391
Logistic Regression	0.8374
Extreme Gradient Boosting	0.8353
Ridge Classifier	0.8343
Random Forest Classifier	0.8177
Linear Discriminant Analysis	0.8145
Gradient Boosting Classifier	0.7970
Naïve Bayes	0.7951

- 
1. **Accuracy** in classification (baseline is 0.5)
  2. **Training Time** in the context of being deployed on a mobile app
  3. **Probabilistic classification** - an advantage for the app team

# Opening the black box with Shapley Values

What happens to people who kill animals? Do they lose part of mind (consciousness) over time when they slaughter? Whatever the reason may be for money, food or occupational hazard. Is there anyway they can redeem?

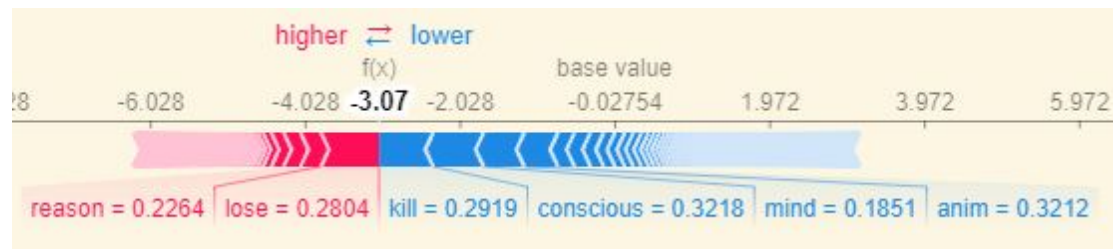
How can I stop being so sensitive? I struggle everyday with with the same toxic comments that I had in my life a lot. Many people just mock me constantly and it's hard to move ahead in life at times. I feel like trash and I am extremely sensitive when it comes to insults and I struggle to take criticism. I get emotional and I try to ignore it but it's too hard. How can I build more emotional resilience and callous my mind?





# Opening the black box with Shapley Values

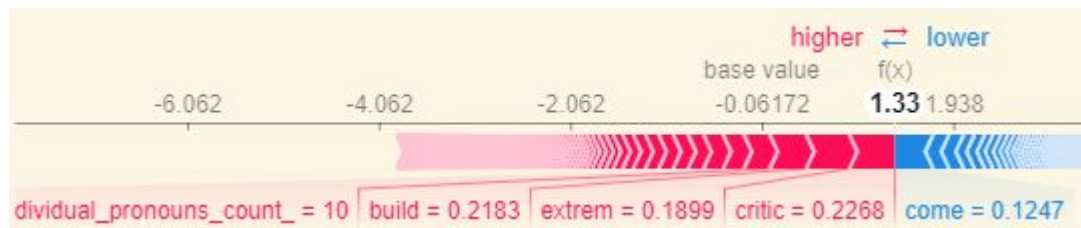
What happens to people who kill animals? Do they lose part of mind (consciousness) over time when they slaughter? Whatever the reason may be for money, food or occupational hazard. Is there anyway they can redeem?



Model Prediction = 0: from r/Buddhism

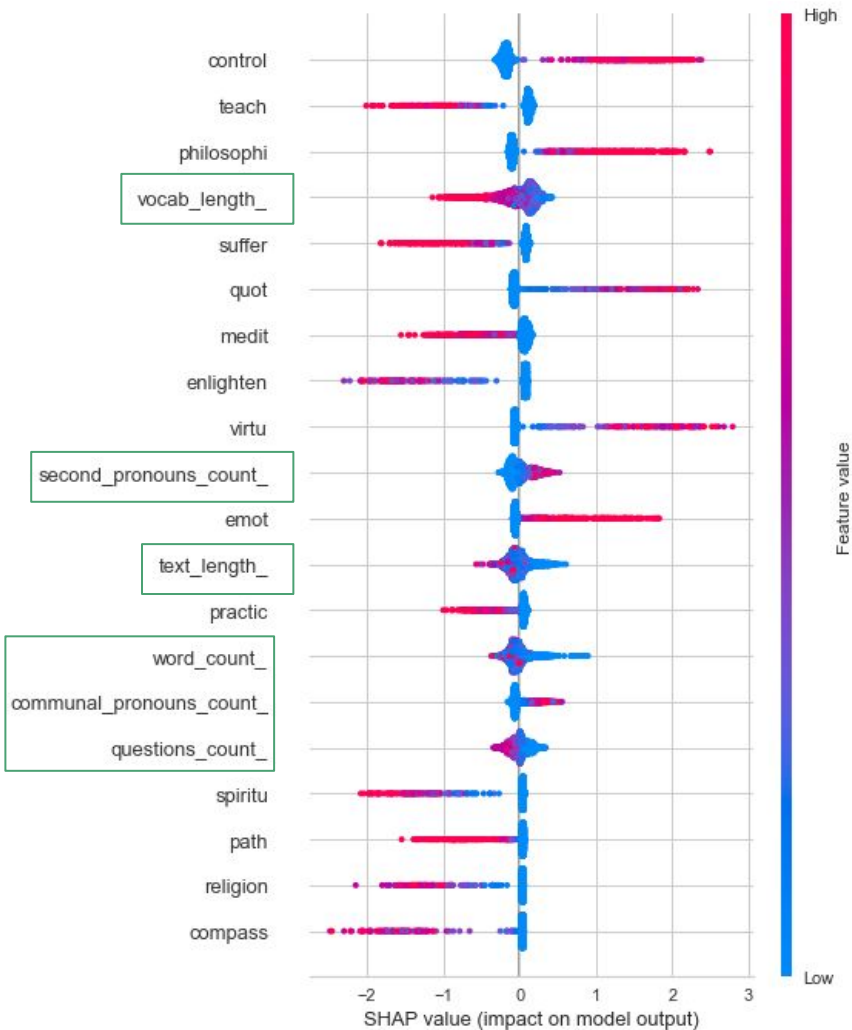
Actual Class = 0: text is really from r/Buddhism

How can I stop being so sensitive? I struggle everyday with with the same toxic comments that I had in my life a lot. Many people just mock me constantly and it's hard to move ahead in life at times. I feel like trash and I am extremely sensitive when it comes to insults and I struggle to take criticism. I get emotional and I try to ignore it but it's too hard. How can I build more emotional resilience and callous my mind?



Model Prediction = 1: from r/Stoicism

Actual Class = 1: text is really from r/Stoicism



## Global Feature Importance

- The color represents the feature value (red high, blue low).
- High tf-idf for words such as "control, quote, philosophy, virtue, emotion, situation" and a more frequent use of second-person pronouns, all lead to higher predicted probabilities of belonging to r/Stoicism.
- On the other hand, high tf-idf for words like "teach, meditate, suffer, path, enlighten, question" and use of longer vocabulary in general, all lead to higher predicted probabilities of belonging to r/Buddhism.

# Conclusion/ Recommendation

The project deliverable is a trained lightGBM classifier with a classification accuracy of 0.83 on a balanced dataset - and is able to discern the originating subreddit of a post simply by using the text within the post.

*The model has two strengths over similar approaches:*

1. The model is trained on the modified corpus where obvious keywords words are omitted.
2. The model leverages new features engineered from the original text (measures of verbosity, pronoun usage, and frequency of question statements).

Business insights:

1. We can focus on targeting our app to people who are intellectually curious, or are facing difficulties in life.
2. Topics we can focus on for our prompts can be about positive thinking, control, religion, meditation to help us better differentiate our users initial inclinations.