CREATING A SUBREDDIT RELEVANCE TOOL

r/science vs r/philosophy

GROUP 5 - PROJECT 3
JANET - JIE YONG - MARK - WAFIR

Who are we?

Science Subreddit Moderators

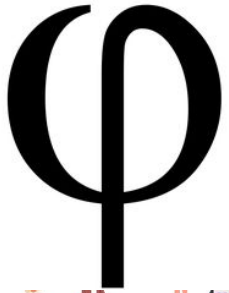
Janet **Mark**

Philosophy Subreddit Moderators

Jie Yong **Wafir**

Who are you?

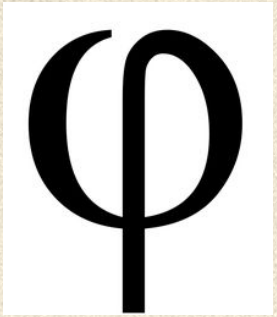
28.4 million



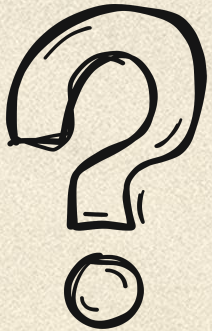
16.9 million

The Problem

Science



Philosophy



The Problem.

According to the nytimes, for roughly 98 percent of the last 2,500 years of Western intellectual history, philosophy was considered the mother of all knowledge.

Today, science, not philosophy, has taken up the mantle as the world's de-facto source of truth, with some no longer sure what philosophy is or is good for anymore.

This begs the question: In which subreddit, would your thought/article/post be best placed in today's context?
Science or **Philosophy**?

Agenda

01 **Introduction** *(Jie Yong)*

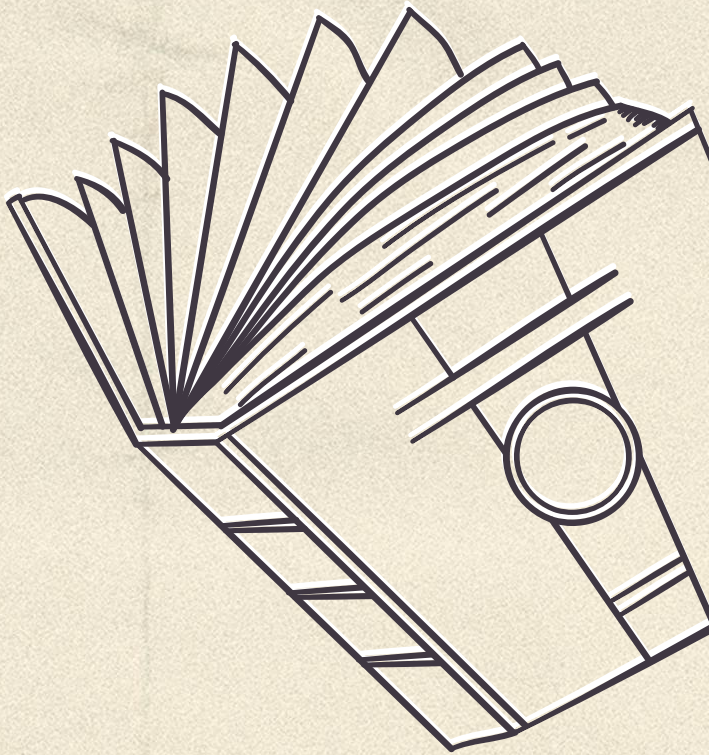
02 **Data Extraction** *(Jie Yong)*

03 **Exploratory Data Analysis** *(Wafir)*

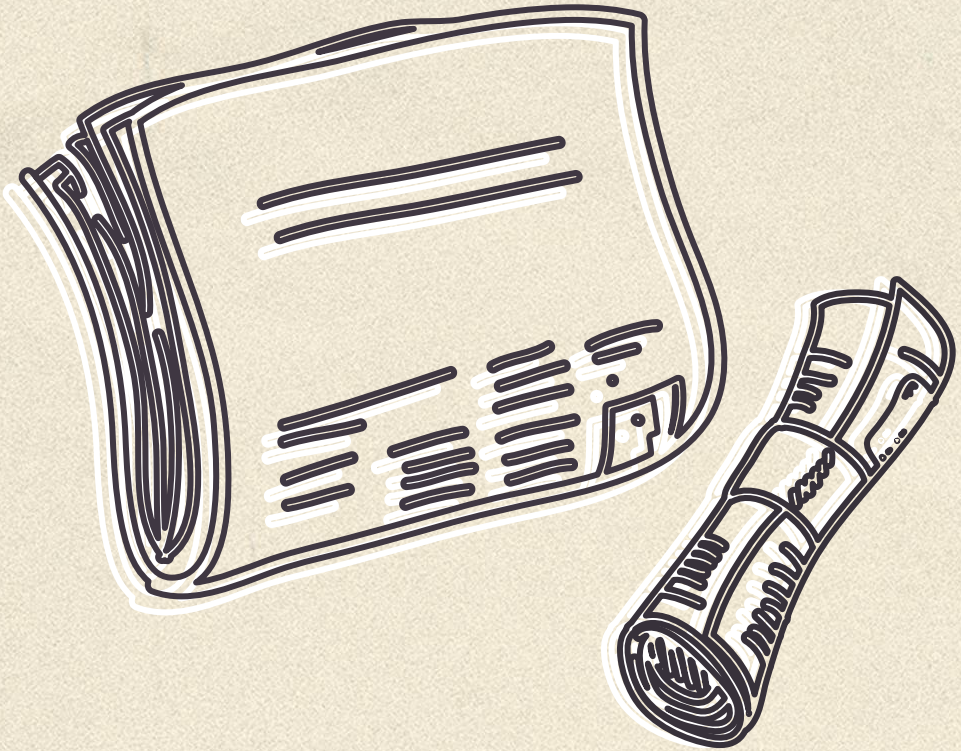
04 **Modelling** *(Mark)*

05 **Conclusions** *(Janet)*

06 **Next Steps** *(Janet)*



Data Extraction



Pushshift API

API used to download from reddit

25,000

Posts downloaded from Science and Philosophy Subreddits each

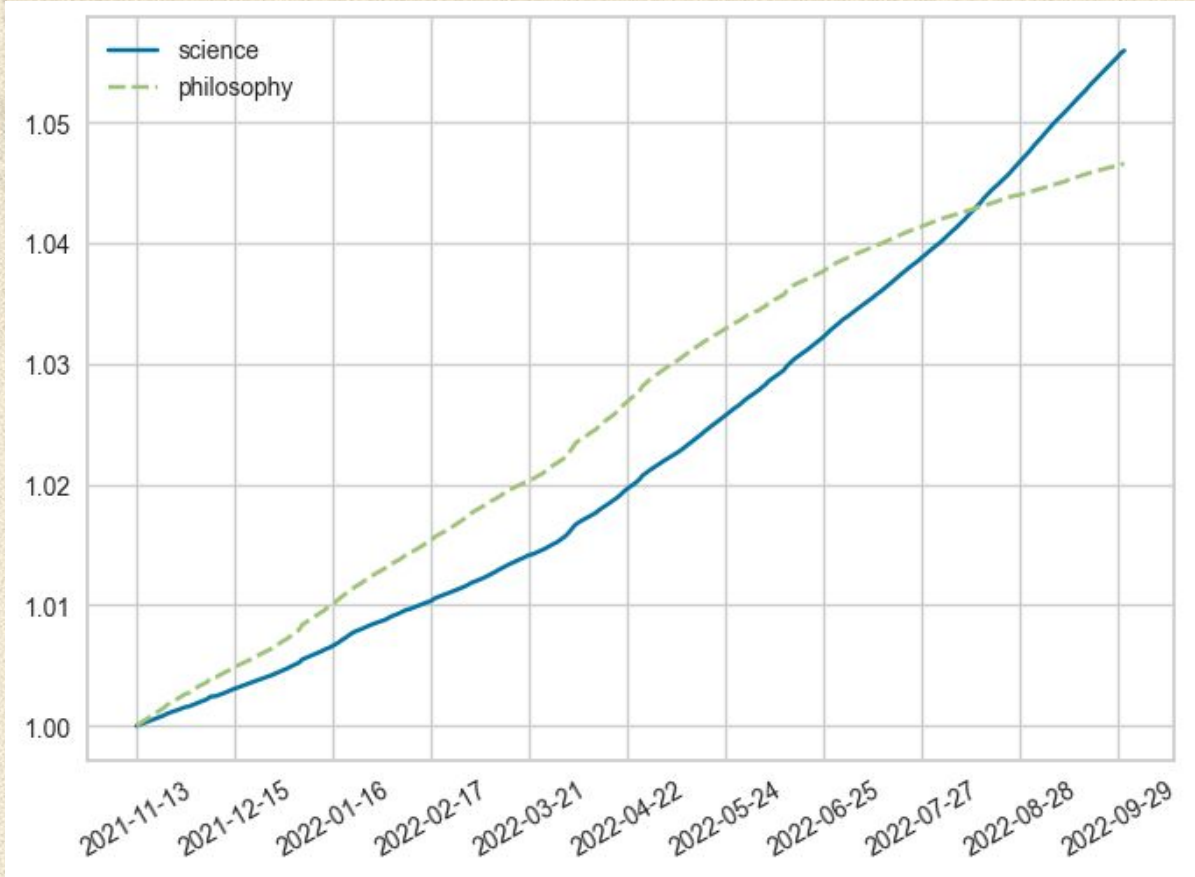
4 Oct 22, 12:00am

Datetime scrapping was done

03

Exploratory Data Analysis

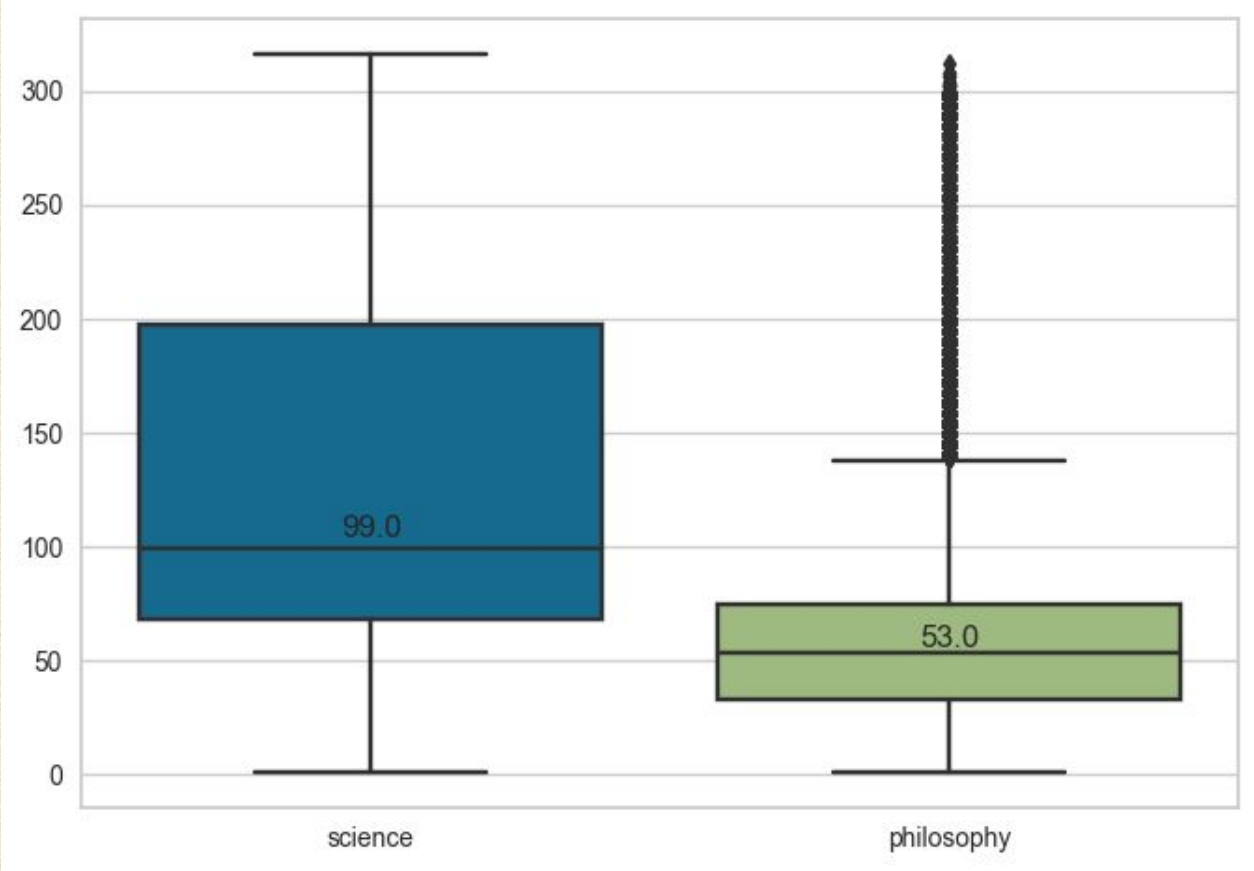
r/Science vs r/Philosophy



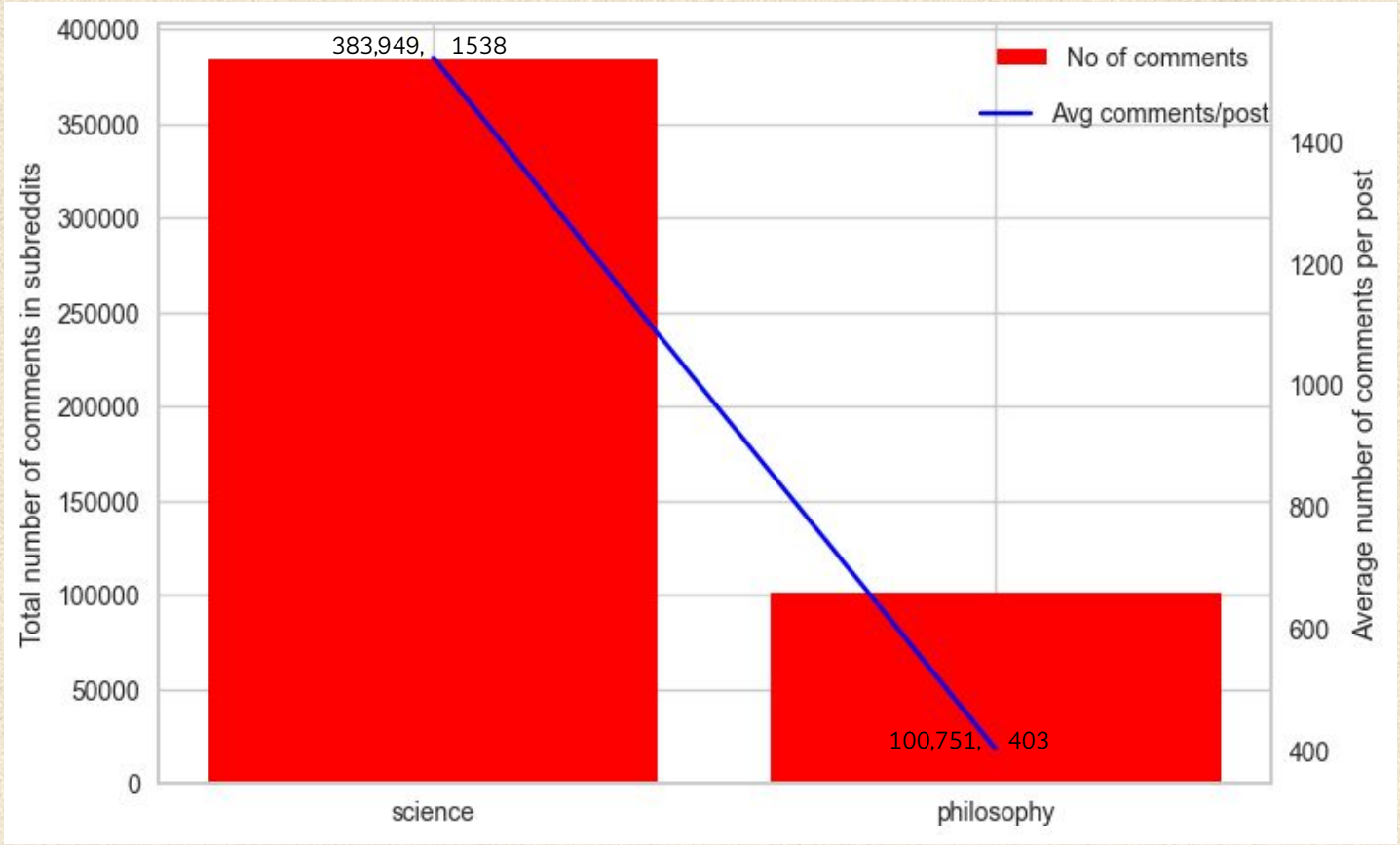
Member Count:

r/Science - 28.4m
r/Philosophy - 16.9m

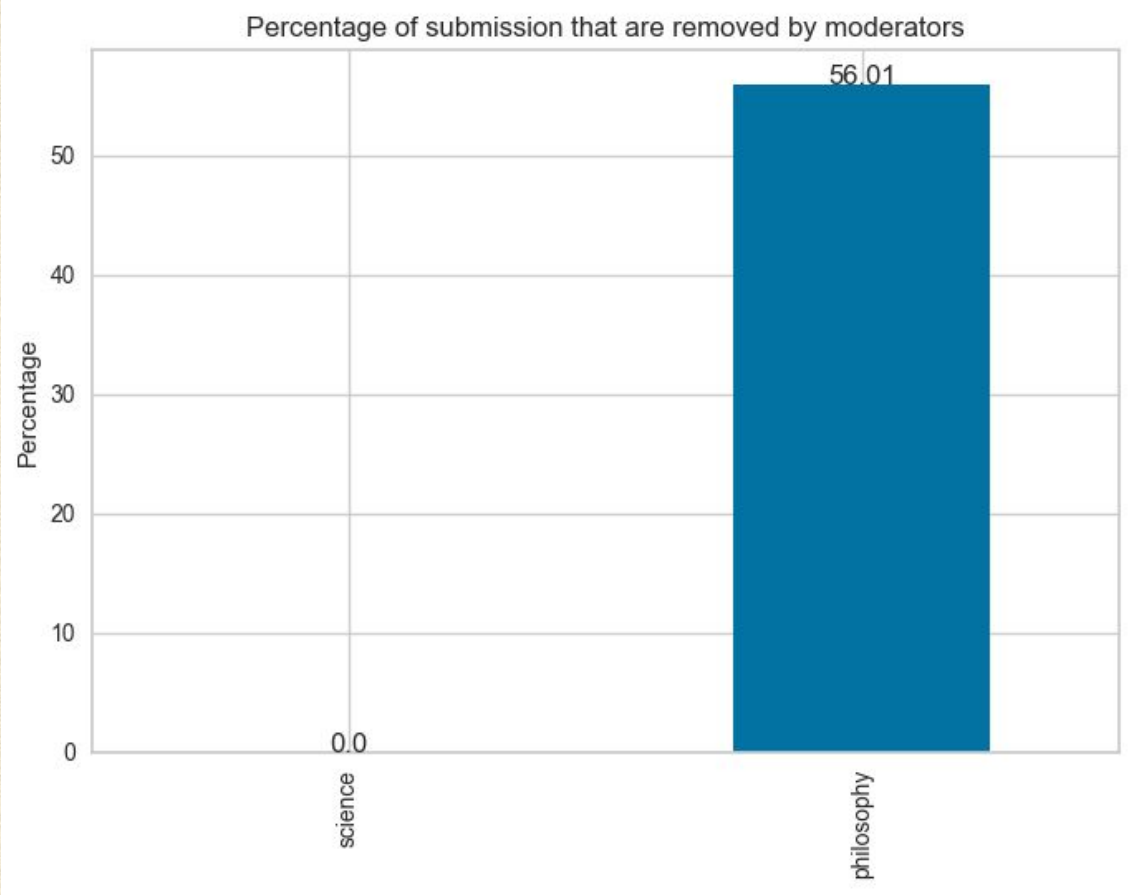
Submission Length



Number of Comments

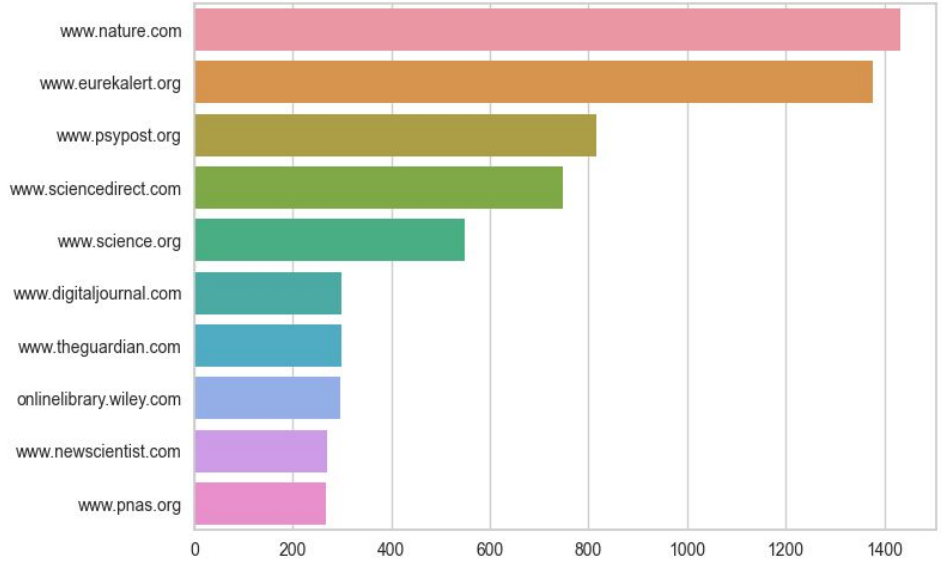


Subreddit Moderation

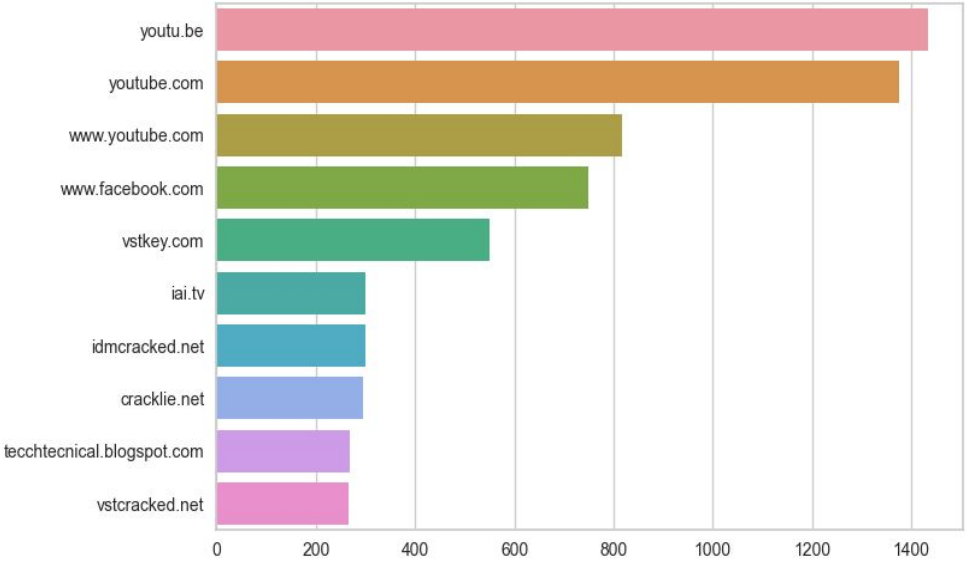


Top 10 Domain Shared

r/Science



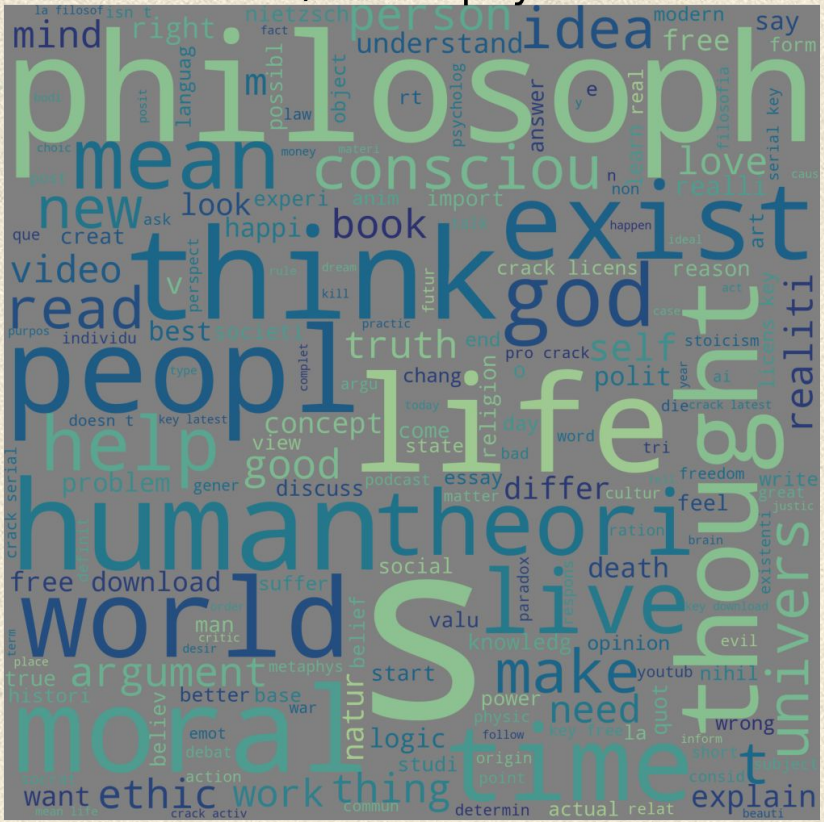
r/Philosophy



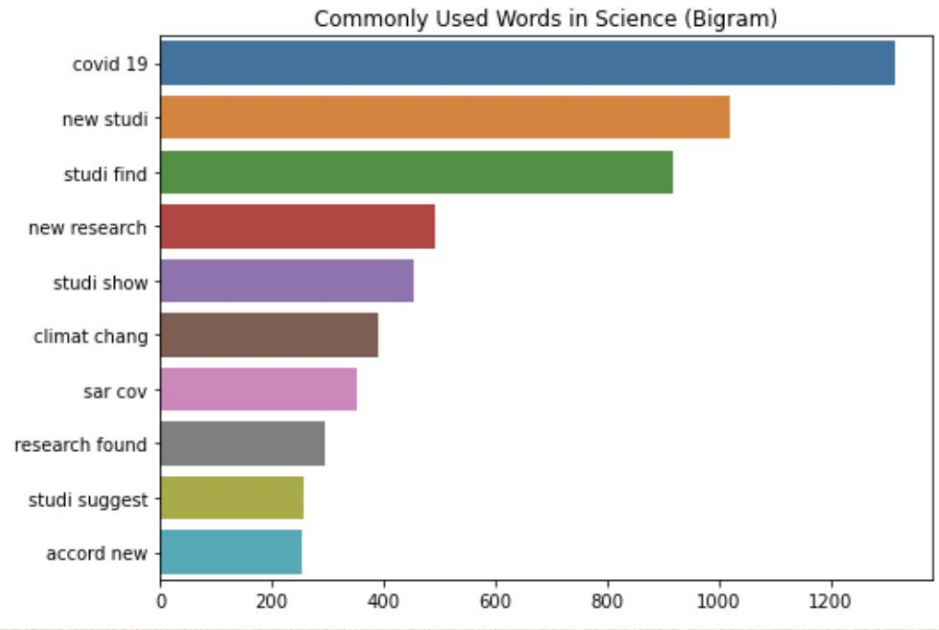
Top Frequent n-grams

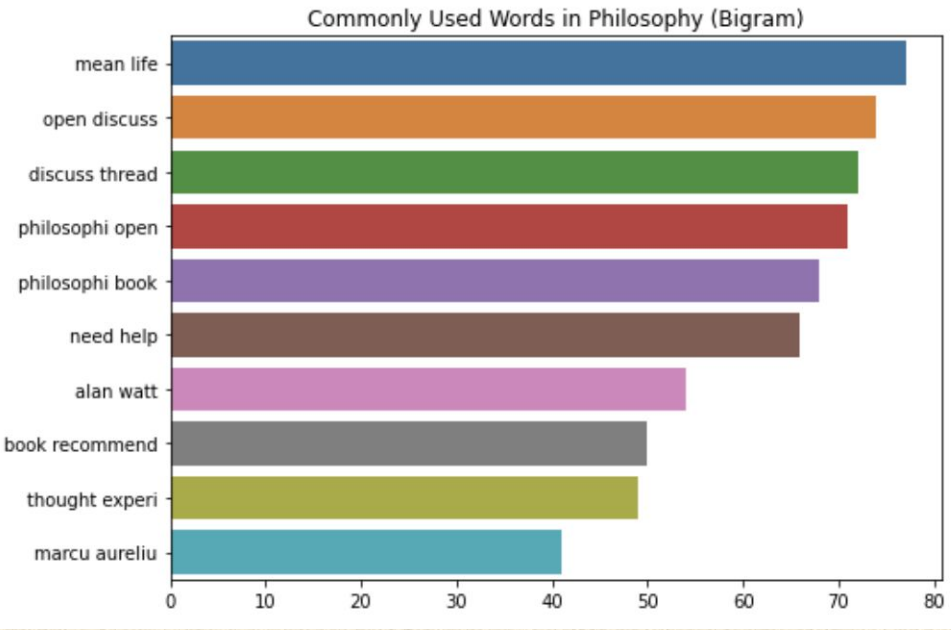
r/Science

r/Philosophy

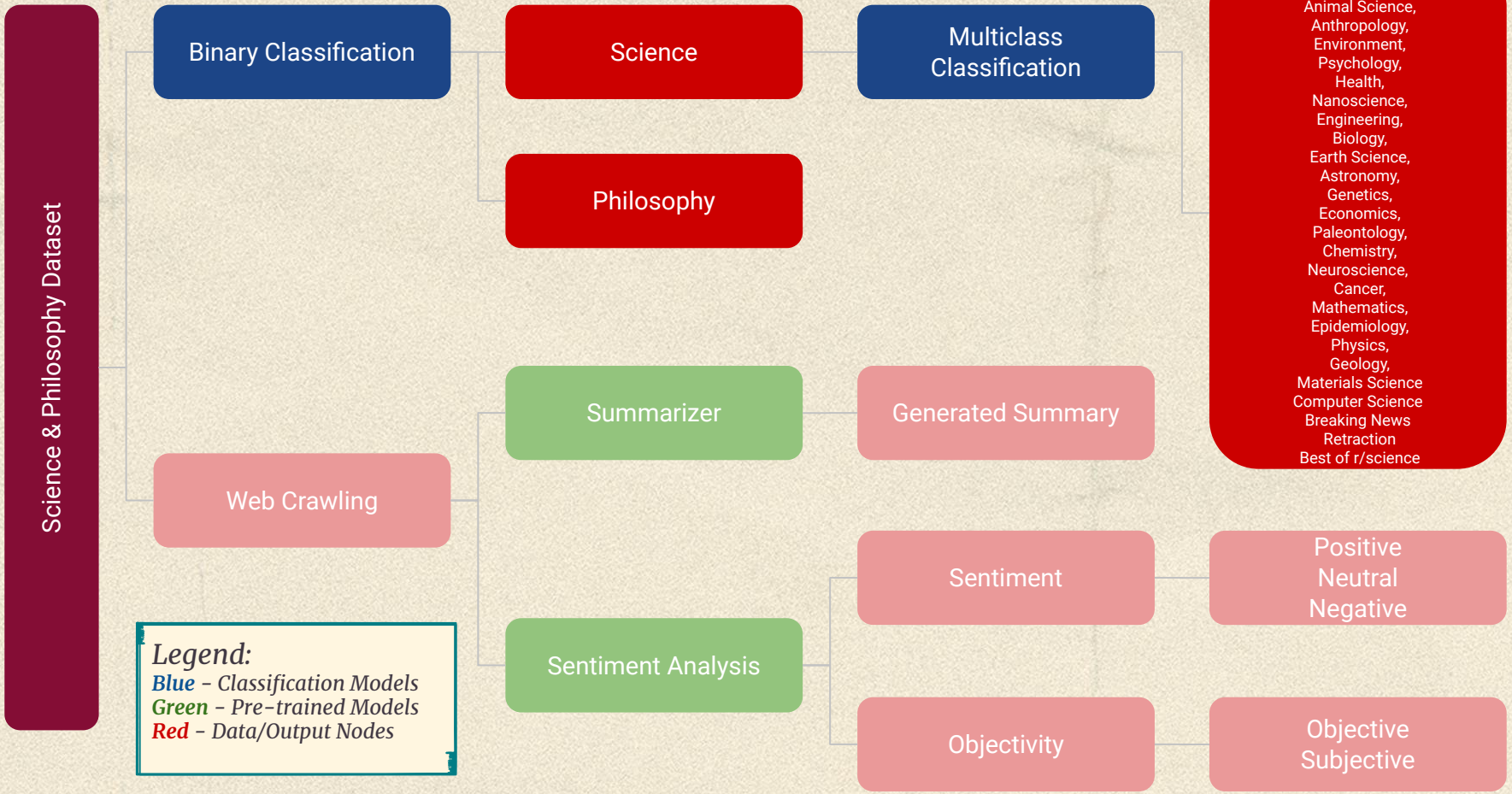


Top 10 Topics





04 part 1
Modelling


Science & Philosophy Dataset
Binary Classification
Science
Philosophy
Multiclass Classification
Medicine
Social Science,
Animal Science,
Anthropology,
Environment,
Psychology,
Health,
Nanoscience,
Engineering,
Biology,
Earth Science,
Astronomy,
Genetics,
Economics,
Paleontology,
Chemistry,
Neuroscience,
Cancer,
Mathematics,
Epidemiology,
Physics,
Geology,
Materials Science
Computer Science
Breaking News
Retraction
Best of r/science
Web Crawling
Summarizer
Generated Summary
Sentiment Analysis
Sentiment
Positive
Neutral
Negative
Objectivity
Objective
Subjective
Legend:
Blue - Classification Models
Green - Pre-trained Models
Red - Data/Output Nodes

Modeling Data Selection

Title

Title of post

Selftext

Content of post

Vectors for classification

Url

Shared url in post

Target

Subreddit

Science or Philosophy subreddit

Flair

Subcategory within subreddit

Binary Classification - Science v Philosophy

PyCaret Top Model Results

Model	Preprocessing	CV Accuracy
MultinomialNB	Manual (all submissions)	0.9550
MultinomialNB	Manual (remove deleted submissions)	0.9408
MultinomialNB	CountVectorizer (all submissions)	0.9653 *
MultinomialNB	CountVectorizer (remove deleted submissions)	0.9441
MultinomialNB	TfidfVectorizer (all submissions)	0.9640
MultinomialNB	TfidfVectorizer (remove deleted submissions)	0.9406
Random Forest	CountVectorizer (all subm	

```
{'cvec__binary': True,
 'cvec__max_df': 1.0,
 'cvec__max_features': None,
 'cvec__min_df': 1,
 'cvec__ngram_range': (1, 2),
 'cvec__stop_words': 'english',
 'cvec__token_pattern': '\\w+'}
```

Top performing models for PyCaret Binary Classification

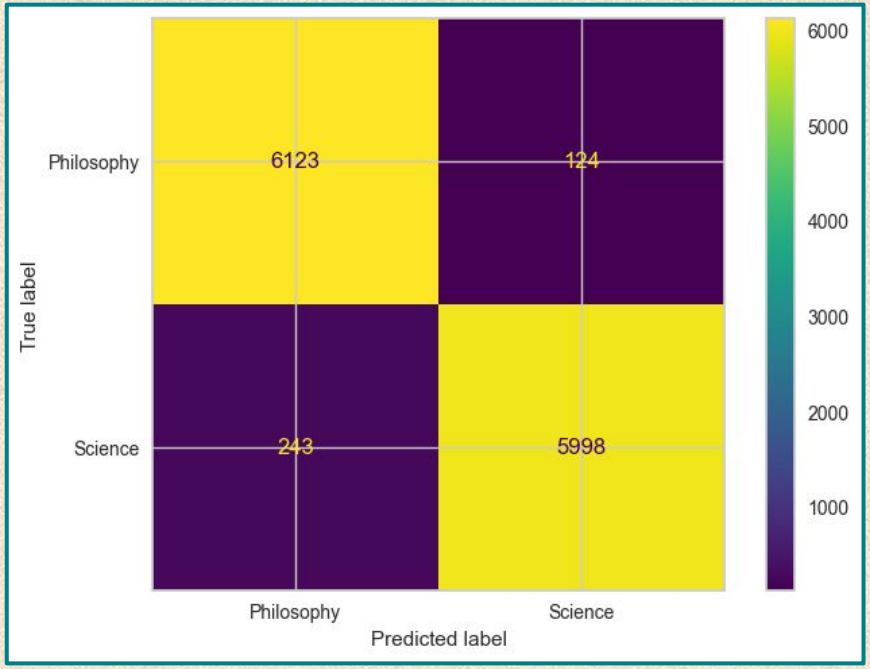
Multibinomial Naive Bayes
With Count Vectorization

Relevant Metrics:

Accuracy - 0.965327
Precision - 0.979745
Recall - 0.961064
F1 Score - 0.970315

Binary Classification - Science v Philosophy

PyCaret Top Model Confusion Matrix



Multibinomial Naive Bayes

With Count Vectorization

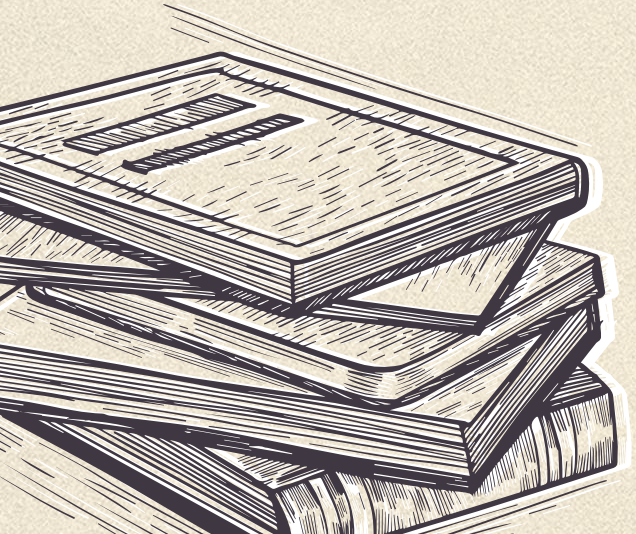
Relevant Metrics:

Accuracy - 0.965327
Precision - 0.979745
Recall - 0.961064
F1 Score - 0.970315

Multiclass Classification - Science

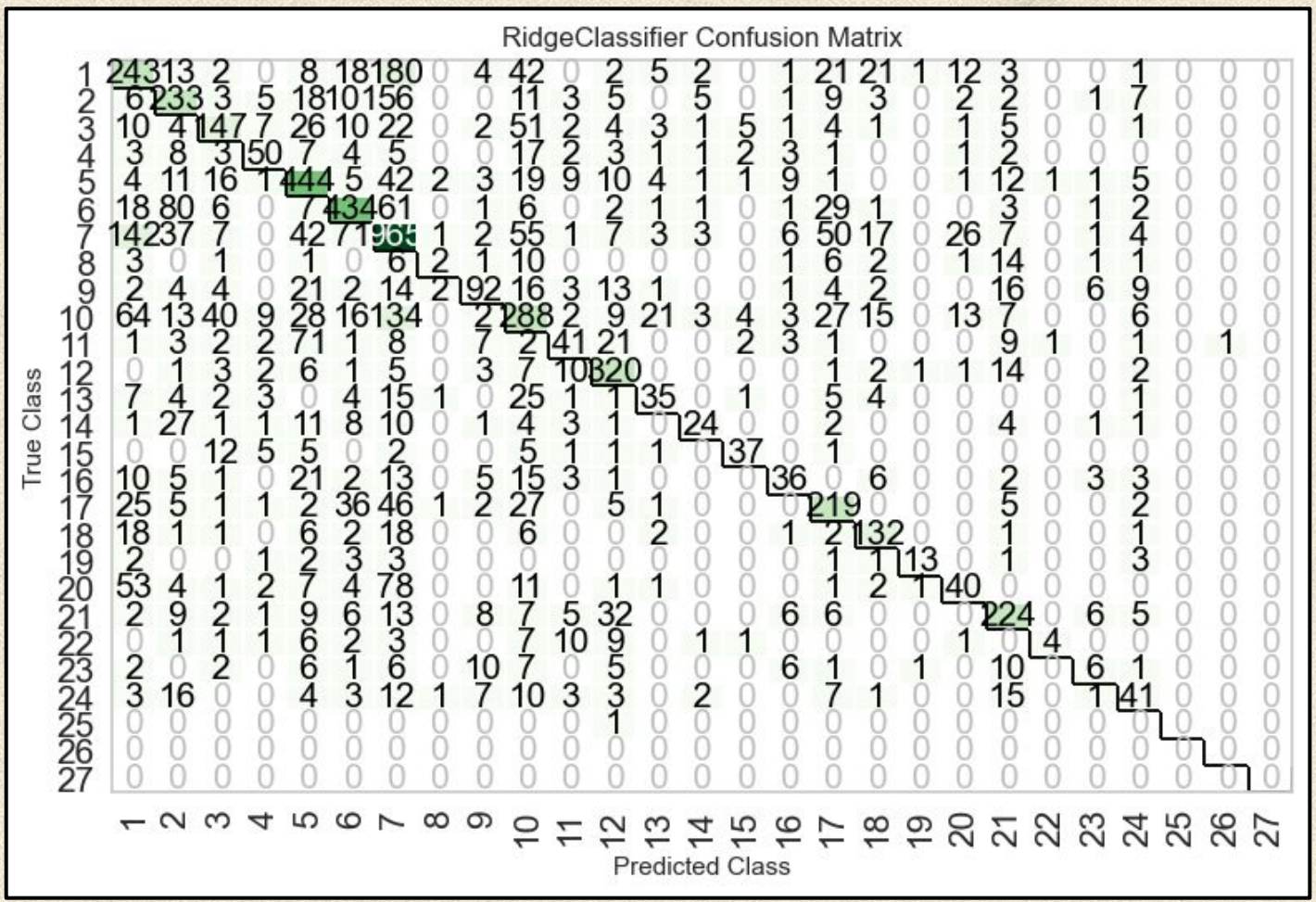
PyCaret Top Model Results

Ridge Classification Model
With TF-IDF Vectorization

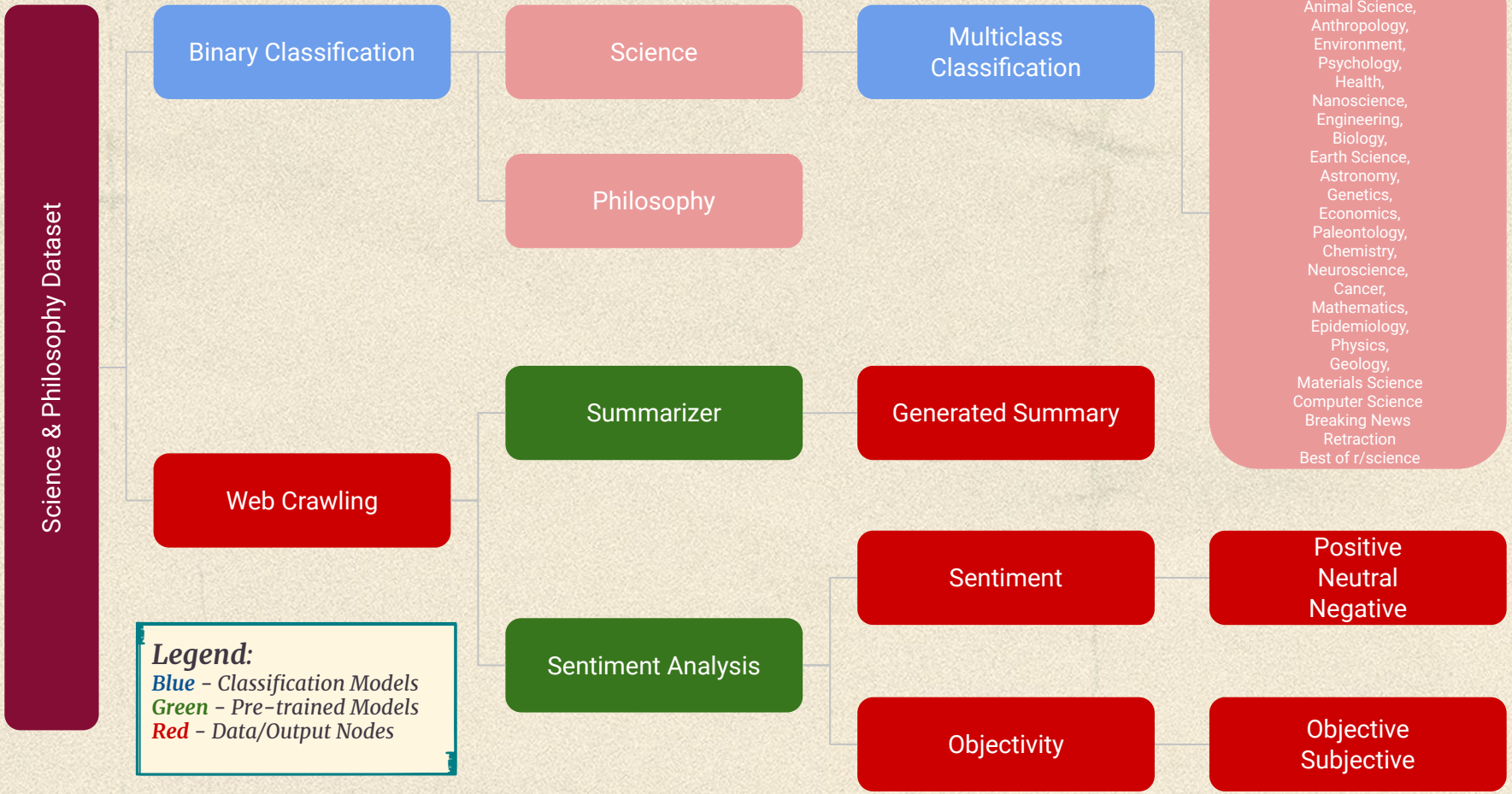


Model	Accuracy	Recall	Precision	F1
Tfidf Pycaret	0.5252*	0.3795	0.5151	0.5089*
Countvec Pycaret	0.4858	0.3476	0.4841	0.4770
facebook/bart-large-mnli	0.2400	0.2400	0.2650	0.2444
valhalla/distilbart-mnli-12-1	0.3000	0.3000	0.5429	0.3487
valhalla/distilbart-mnli-12-3	0.3400	0.3400	0.4547	0.3585
typeform/distilbert-base-uncased-mnli	0.3200	0.3200	0.4313	0.3019
Narsil/deberta-large-mnli-zero-cls	0.4400	0.4400	0.5163	0.4599

Top performing models for PyCaret Multiclass Classification


RidgeClassifier Confusion Matrix
True Class
Predicted Class
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27

04 part 2
Web Crawling


Science & Philosophy Dataset
Binary Classification
Science
Philosophy
Multiclass Classification
Medicine
Social Science,
Animal Science,
Anthropology,
Environment,
Psychology,
Health,
Nanoscience,
Engineering,
Biology,
Earth Science,
Astronomy,
Genetics,
Economics,
Paleontology,
Chemistry,
Neuroscience,
Cancer,
Mathematics,
Epidemiology,
Physics,
Geology,
Materials Science
Computer Science
Breaking News
Retraction
Best of r/science
Web Crawling
Summarizer
Generated Summary
Sentiment Analysis
Sentiment
Positive
Neutral
Negative
Objectivity
Objective
Subjective
Legend:
Blue - Classification Models
Green - Pre-trained Models
Red - Data/Output Nodes

99% post only have content + link

Posted by u/Meatrition 2 hours ago

Vote

Health Time-restricted eating with or without low-carbohydrate diet reduces visceral fat and improves metabolic syndrome: A randomized trial

cell.com/cell-r...

10 Comments

Posted by u/Additional-Two-7312 9 hours ago

675

Astronomy Our moon has been slowly drifting away from Earth over the past 2.5 billion years, research finds

pnas.org/doi/fu...

140 Comments Award Share Save ...

Posted by u/ADefiniteDescription Φ 14 hours ago

8

Article [PDF] Mackie and the Meaning of Moral Terms

jhaponline.org/jhap/a...

1 Comment Award

Posted by u/ADefiniteDescription Φ 2 days ago

0

Blog Informed Consent and the Joe Rogan Experience

prindleinstitute.org/2022/0...

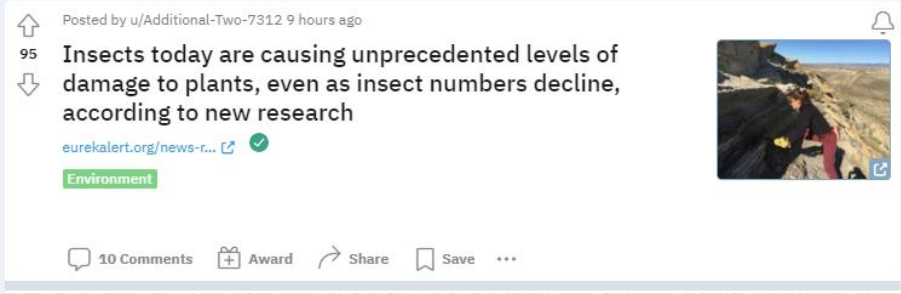
1 Comment Award Share Save ...

Summarizer

Hugging Face

sshleifer/distilbart-cnn-12-6

Reddit Post



Article Link

https://www.eurekalert.org/news-releases/967297

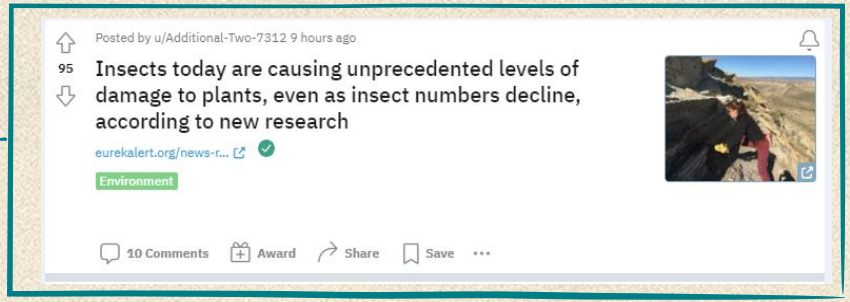
Summary

First-of-its-kind study compares insect damage of modern-era plants with that of fossilized leaves from as far back as the Late Cretaceous period, nearly 67 million years ago . Scientists say a warming climate, urbanization and introduction of invasive species likely have had a major impact .

Sentiment Analysis

spacytextblob

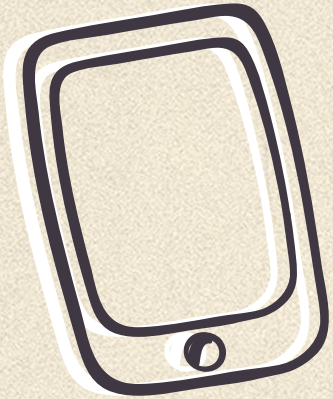
Reddit Post



+

https://www.eurekalert.org/news-releases/967297

Article Link



Outputs

Sentiment	Subjectivity
Neutral	Objective
0.15	0.47

05
Summary

Summary

Statistics

r/Science
28.4 million members
2000 posts & comments / day
100 words / post

r/Philosophy
16.9 million members
200 posts & comments / day
50 words / post
Substantial spam posts

Goals

Guide our users to:

1. Identify which Subreddit to posts at
2. Write a neutral & objective posts
3. Write a summary for the external links

Modelling

- **Binary Classification**
  Multinomial Naive Bayes
  Accuracy: 0.965
- **Multiclass Classification**
  Ridge Classifier
  Accuracy: 0.525
- **Sentiment Analysis**
  spacytextblob
- **Summarizer**
  HuggingFace


WEEK 28
JULY
MONDAY
9
TUESDAY
10
THURSDAY
12
2018
FRIDAY
13
SATURDAY
14
SUNDAY
15
NOTES
JULY
The Solution:
An App for That
Figure out where and what to put in a post

Post Title

Insects today are causing unprecedented levels of damage to plants, even as insect numbers declin

Post Content

URL

https://www.eurekalert.org/news-releases/967297

Inspect

Post Title

You are not obliged to vote, and maybe shouldn't. An argument from analogy.

Post Content

URL

https://wonderandaporia.substack.com/p/should-you-vote-an-argument-from?r=1l11lq&utm_cam

Inspect

06

Next Steps

Philosophy – Multiclass Classification

- **Epistemology** — Knowledge & Truth
- **Metaphysics** — Reality & Being
- **Logic** — Argumentation & Reason
- **Axiology** — Aesthetics & Ethics
- **Political** — State & Government

Additional Features

Word Correction
Suggestion to user to potentially improve comments & upvotes

Popularity Prediction
Using regression analysis to predict the number of comments & upvotes of the posts

Auto Post
From streamlit, auto-post to reddit after user is satisfied with the results

Spam Detection
Detect non-related posts

Expand to other fact based Subreddits

- r/Economics
- r/Astronomy
- r/History
- r/Geography
- r/Health

THANK YOU

Try out our beta!

https://tinyurl.com/subreddit-rel