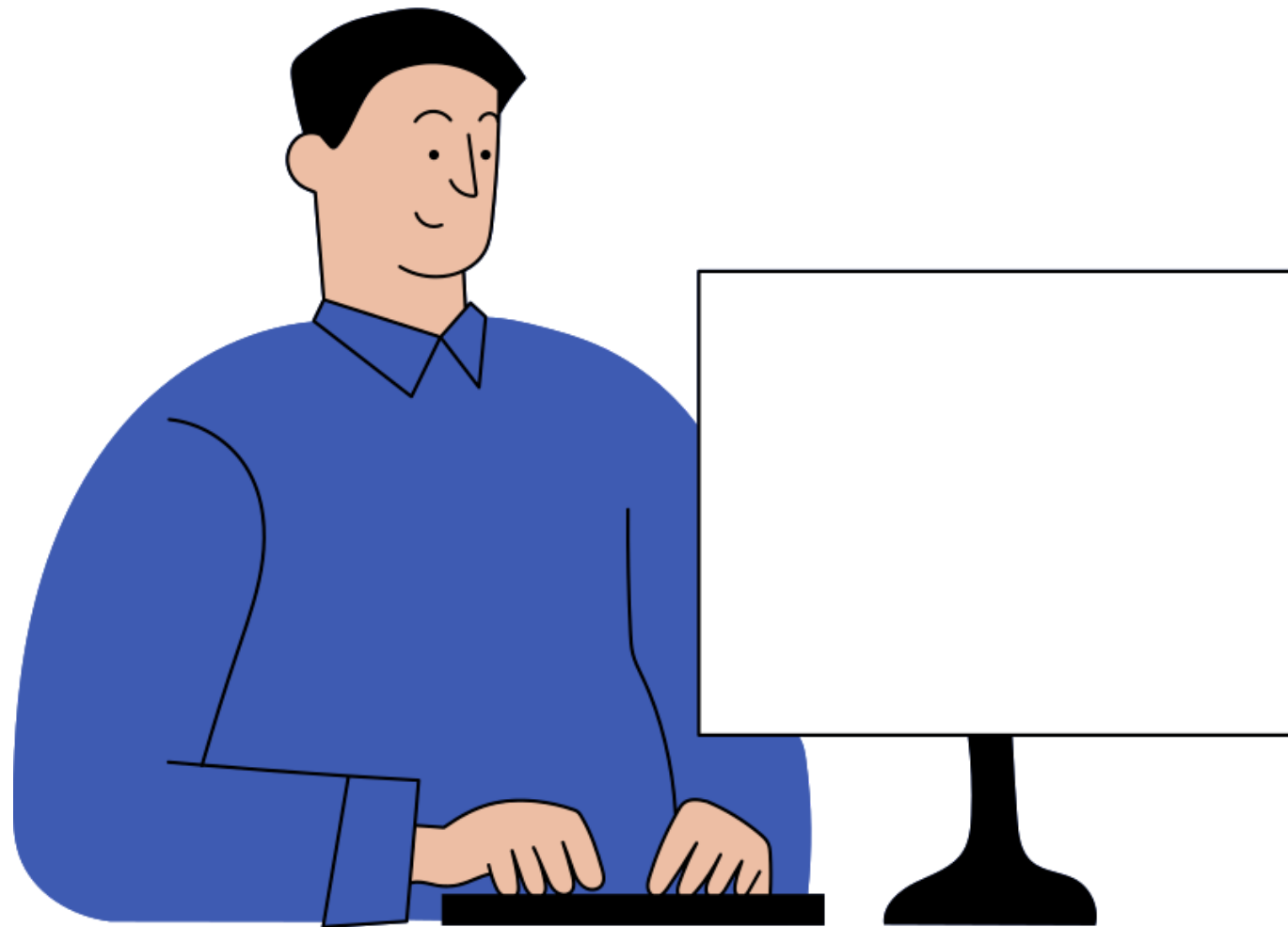


CE/CZ4034 Information Retrieval

Topic: 2023 Recession



Contents



Introduction

Data Crawling

Indexing

Classification

Conclusion



Introduction

2023 recession in the eyes of Reddit

- COVID-19 pandemic has caused significant disruptions in the global economy.
- IMF has predicted that the global economy will grow by 2.9% in 2023, which is lower than the growth rate of 6.0% in 2021.
- Analysing data on Reddit can help identify how people feel about the economy and their financial situation, in order to determine the general sentiment about the predicted Recession.



Introduction


2023 recession in the eyes of Reddit

- Social media platforms allow us to share thoughts and opinions freely.
- Machine learning can be used to crawl and analyze Reddit posts and comments, providing invaluable insights into societal sentiments and human emotions.
- Our team aims to create an information retrieval system that allows users to enter specific keywords and generate a vast collection of Reddit posts and comments for further analysis and classification.

Getting the Data

Subreddits
r/FinancialCareers, r/cscareerquestions, r/medicine
r/singapore, r/india, r/usa, r/china, r/unitedkingdom

Keywords
layoff, recession, economic downturn, unemployment, debt, slowdown, financial crisis, poverty

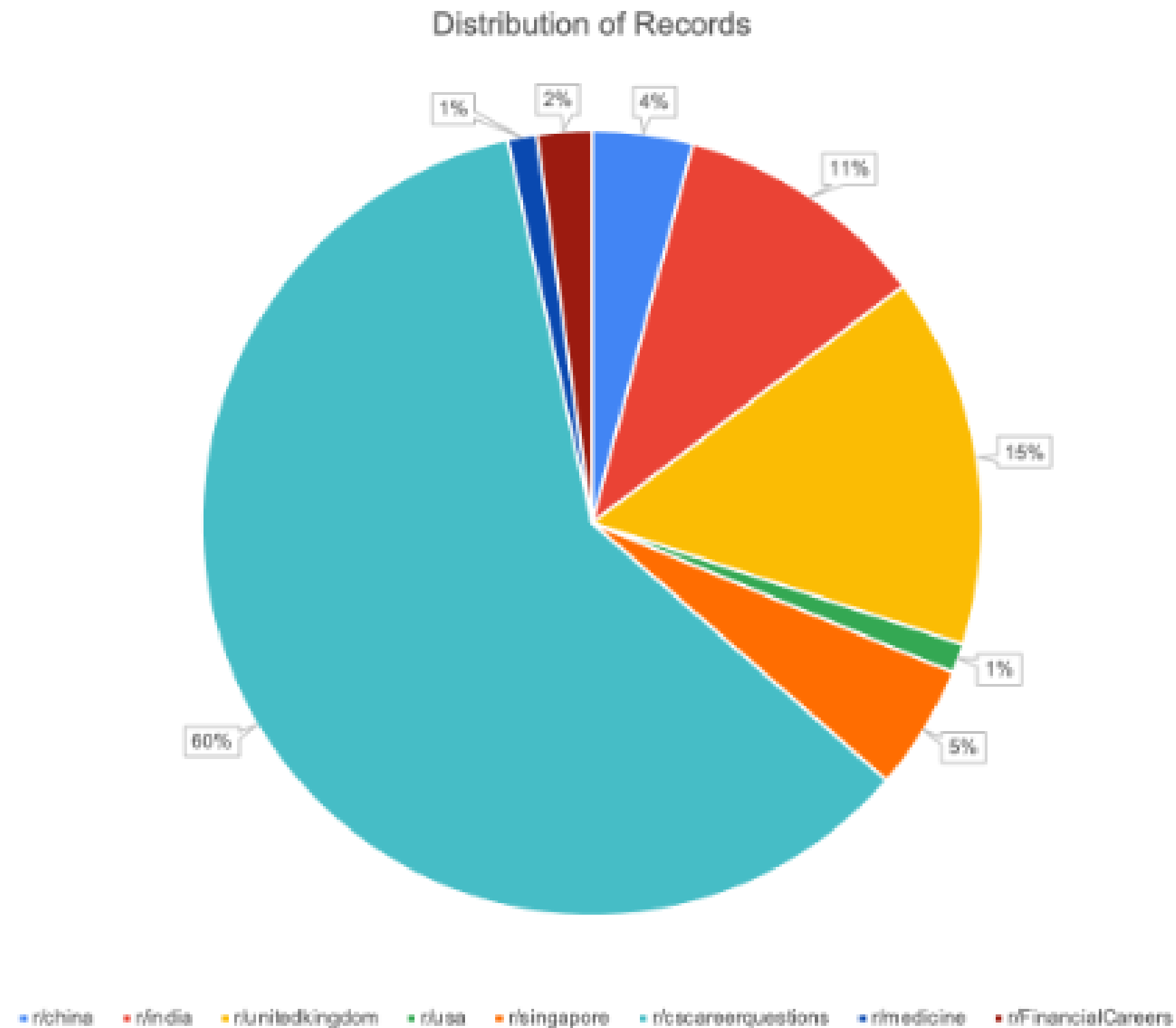
Tools


20,559 posts and comments

955,882 words

27,806 unique words

Distribution of Data



- Unequal subreddit membership sizes.
- Ranges from 44.9k to 1.4M members.
- Subreddits contribute unequal data.

Example Queries

- Find from the corpus the general sentiment of Singapore residents towards the recession
- Find from the corpus which country is feeling the most positive about the recession
- Find from the corpus which industry is feeling the most positive about the recession
- Find from the corpus how many people working in the healthcare industry have been laid off.
- Find from the corpus the unemployment rate in the financial sector.

User Goals

- Users may wish to compare the impact of the recession across different countries
- Users may be prospective employees considering how safe it is to migrate to the listed industries in the midst of a recession.
- Users may be part of the industries and may wish to know the future of the industry.
- Users may be interested the general sentiment that the public have on the recession.

Team Check-in

01. Click the more tab in the editor side panel.

02. Click the more tab in the editor side panel.

03. Click the more tab in the editor side panel.

Choose an Emoji, GIF, or image from a mood meter that best represents how you feel at the moment.

How are you feeling?



Icebreaker

**What did you have
for breakfast today?**



Add your
idea here



**Before we start with the
session, let's warm up a little
with an icebreaker question:**

01

What was the last thing you purchased online?

02

What's your morning routine?

03

What's a random act of kindness you did for a stranger?

04

What's the most challenging thing you've done in life?

05

What's your favorite dessert?



Add your
idea here



Add your
idea here



Add your
idea here

Indexing

Full-Stack Search Engine

- Data Indexing
- User Querying

APACHE
LUCENE

Solr



Backend



AnyChart

Frontend

Apache Solr

Data Fields

Field	Description
dataset	Dataset the Reddit data belong to ie. train/test
category	Category the Reddit data belong to ie. comment/post
subreddit	Subreddit the Reddit data belong to

Field	Description
author	Author of the Reddit data
created_date	Created date of the Reddit data
score	Reddit score of the Reddit data
text_clean	Pre-processed content text of the Reddit data
final_label	Sentiment of the Reddit Data ie. Positive/Neutral/Negative



Apache Solr

Configurations

```
733 <!-- Primary search handler, expected by most clients, examples and UI frameworks -->
734 <requestHandler name="/select" class="solr.SearchHandler">
735   <lst name="defaults">
736     <str name="echoParams">explicit</str>
737     <int name="rows">10</int>
738     <str name="spellcheck.dictionary">default</str>
739     <str name="spellcheck">on</str>
740     <str name="spellcheck.count">10</str>
741   </lst>
742   <arr name="last-components">
743     <str>spellcheck</str>
744   </arr>
745 </requestHandler>
```

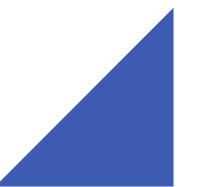
Spell Checking

- Configures default dictionary based on terms found in document
- Returns top 10 suggestions for incorrect query

```
24 <filter>
25   <filter-name>cross-origin</filter-name>
26   <filter-class>org.eclipse.jetty.servlets.CrossOriginFilter</filter-class>
27   <init-param>
28     <param-name>allowedOrigins</param-name>
29     <param-value>*</param-value>
30   </init-param>
31   <init-param>
32     <param-name>allowedMethods</param-name>
33     <param-value>GET,POST,OPTIONS,DELETE,PUT,HEAD</param-value>
34   </init-param>
35   <init-param>
36     <param-name>allowedHeaders</param-name>
37     <param-value>origin, content-type, accept</param-value>
38   </init-param>
39 </filter>
40
41 <filter-mapping>
42   <filter-name>cross-origin</filter-name>
43   <url-pattern>*</url-pattern>
44 </filter-mapping>
```

REST API Communication

- /update: for POST request to upload document to database
- /select: main query handler for GET request to retrieve information



Apache Solr

Querying

Query field:

Example	Description
<code>&q=text_clean%job</code>	find data with term 'job' in 'text_clean' field

Filter query:

Example	Description
<code>&fq=final_label%positive</code>	filter query results for positive sentiments only

```
http://localhost:8983/solr/reddit/select?indent=true&q.op=OR&q=text_clean%3Ajob&spellcheck.build=true&spellcheck=true&useParams=

{
  "responseHeader":{
    "zkConnected":true,
    "status":0,
    "QTime":3,
    "params":{
      "q":"text_clean:job",
      "indent":"true",
      "spellcheck":"true",
      "q.op":"OR",
      "spellcheck.build":"true",
      "useParams":"","
      "_":"1680450985575"}},
  "command":"build",
  "response":{"numFound":2780,"start":0,"numFoundExact":true,"docs":[
    {
      "dataset":["train"],
      "category":["comment"],
      "subreddit":["r/FinancialCareers"],
      "author":["foodVSfood"],
      "created_date":["10/13/2022 18:14"],
      "score":[1],
      "text":["Finding a job is much easier when you have a job. Treat the job search like you don't have a job right now."],
      "text_clean":["finding a job is much easier when you have a job treat the job search like you dont have a job right now"],
      "label_1":["POSITIVE"],
      "label_2":["POSITIVE"],
      "label_3":["NEUTRAL"],
      "final_label":["POSITIVE"],
      "text_clean_stopword":["finding job much easier job treat job search like dont job right"],
      "id":["0e1e522-299a-4dd5-ba41-10ea5a5cf6c2"],
      "_version_":1762076364452659200},
    {
      "dataset":["train"],
      "category":["comment"],
      "subreddit":["r/cscareerquestions"],
      "author":["Countrydeepfrypie"],
      "created_date":["12/1/2022 2:58"],
      "score":[3],
      "text":["200 job applications in one night, 5 second interviews, 3 job offers, one job successful. 4 weeks looking. 3 years experience."],
      "text_clean":["job applications in one night second interviews job offers one job successful weeks looking years experience"],
      "label_1":["POSITIVE"],
      "label_2":["POSITIVE"],
      "label_3":["NEUTRAL"],
      "final_label":["POSITIVE"],
      "text_clean_stopword":["job applications in one night second interviews job offers one job successful weeks looking years experience"],
      "id":["0e1e522-299a-4dd5-ba41-10ea5a5cf6c2"],
      "_version_":1762076364452659200}
  ]}
}
```



Apache Solr

Querying

Re-rank query:

Example	Description
<code>&rq={!rerank reRankQuery=\$rq reRankDocs=1000 reRankWeight=3}&rq q=(job)</code>	for top 1000 query results, re-order by re-computing scores using new weight of 3 for those containing term 'job'

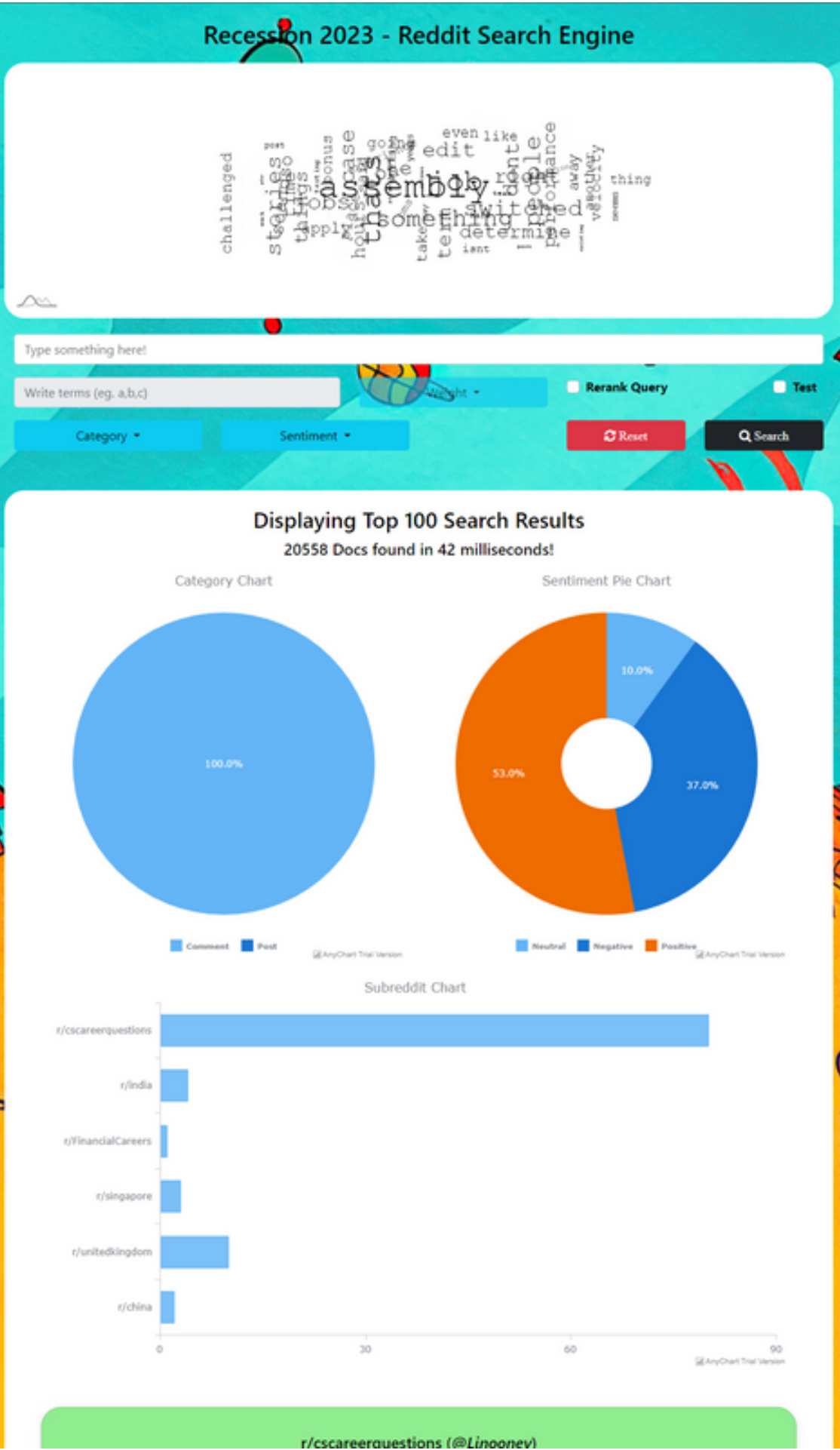
```
http://localhost:8983/solr/reddit/select?indent=true&q.op=OR&q=text_clean%3Ajob&spellcheck.build=true&spellcheck=true&useParams=

{
  "responseHeader":{
    "zkConnected":true,
    "status":0,
    "QTime":3,
    "params":{
      "q":"text_clean:job",
      "indent":"true",
      "spellcheck":"true",
      "q.op":"OR",
      "spellcheck.build":"true",
      "useParams":"","
      "_":"1680450985575"}},
  "command":"build",
  "response":{"numFound":2780,"start":0,"numFoundExact":true,"docs":[
    {
      "dataset":["train"],
      "category":["comment"],
      "subreddit":["r/FinancialCareers"],
      "author":["foodVSfood"],
      "created_date":["10/13/2022 18:14"],
      "score":[1],
      "text":["Finding a job is much easier when you have a job. Treat the job search like you don't have a job right now."],
      "text_clean":["finding a job is much easier when you have a job treat the job search like you dont have a job right now"],
      "label_1":["POSITIVE"],
      "label_2":["POSITIVE"],
      "label_3":["NEUTRAL"],
      "final_label":["POSITIVE"],
      "text_clean_stopword":["finding job much easier job treat job search like dont job right"],
      "id":["0eb1e522-299a-4dd5-ba41-10ea5a5cf6c2"],
      "_version_":1762076364452659200},
    {
      "dataset":["train"],
      "category":["comment"],
      "subreddit":["r/cscareerquestions"],
      "author":["Countrydeepfrypie"],
      "created_date":["12/1/2022 2:58"],
      "score":[3],
      "text":["200 job applications in one night, 5 second interviews, 3 job offers, one job successful. 4 weeks looking. 3 years experience."],
      "text_clean":["job applications in one night second interviews job offers one job successful weeks looking years experience"],
      "label_1":["POSITIVE"],
      "label_2":["POSITIVE"],
      "label_3":["NEUTRAL"],
      .
```

Frontend UI

Indexing & Classification

Live Demonstration





Classification

Let's begin.



Stacked Ensemble

The use of multiple models, base estimators, in the prediction process. A stacked classifier combines the outputs of the baseline models and lets a final meta classifier learn and make a concluding decision. The ensemble utilizes the strengths of each individual model and seeks to improve model performance.

Subjectivity & Polarity Classification			
Model	Wall Time	F1 Score (%)	Accuracy (%)
StackedClassifier	21min 21s	66.7	68.2

NER

Starting on a new project begins with a curation of ideas. This digital brainstorm session will allow us to lay down and organize all our thoughts, ideas, and inspiration. Next, we will vote on ideas that are promising and add comments.

01

Copy a sticky note, and then we'll write our thoughts, ideas, and inspiration.



02

Use the stars to vote which ones we like to pursue.



03

Circle or comment on any promising ideas.



GoEmotions

GoEmotions is a comprehensive and diverse emotional analysis dataset that contains 58k English comments from Reddit. It allows us to explore and analyze complex human emotions.

We used the bert-base-uncased model to tokenize the text data and arpanghoshal's EmoRoBERTa model to classify the tokenized text into 28 emotions.

Positive		Negative		Neutral
admiration 🙌	joy 😄	anger 😡	grief 😞	confusion 😕
amusement 😂	love ❤️	annoyance 😠	nervousness 😰	curiosity 🤔
approval 👍	optimism 🙌	disappointment 😞	remorse 😞	realization 💡
caring 😊	pride 😊	disapproval 🗨️	sadness 😞	surprise 😲
desire 😍	relief 😌	disgust 🤢		
excitement 😄		embarrassment 😳		
gratitude 🙏		fear 😨		

arpanghoshal's EmoRoBERTa model

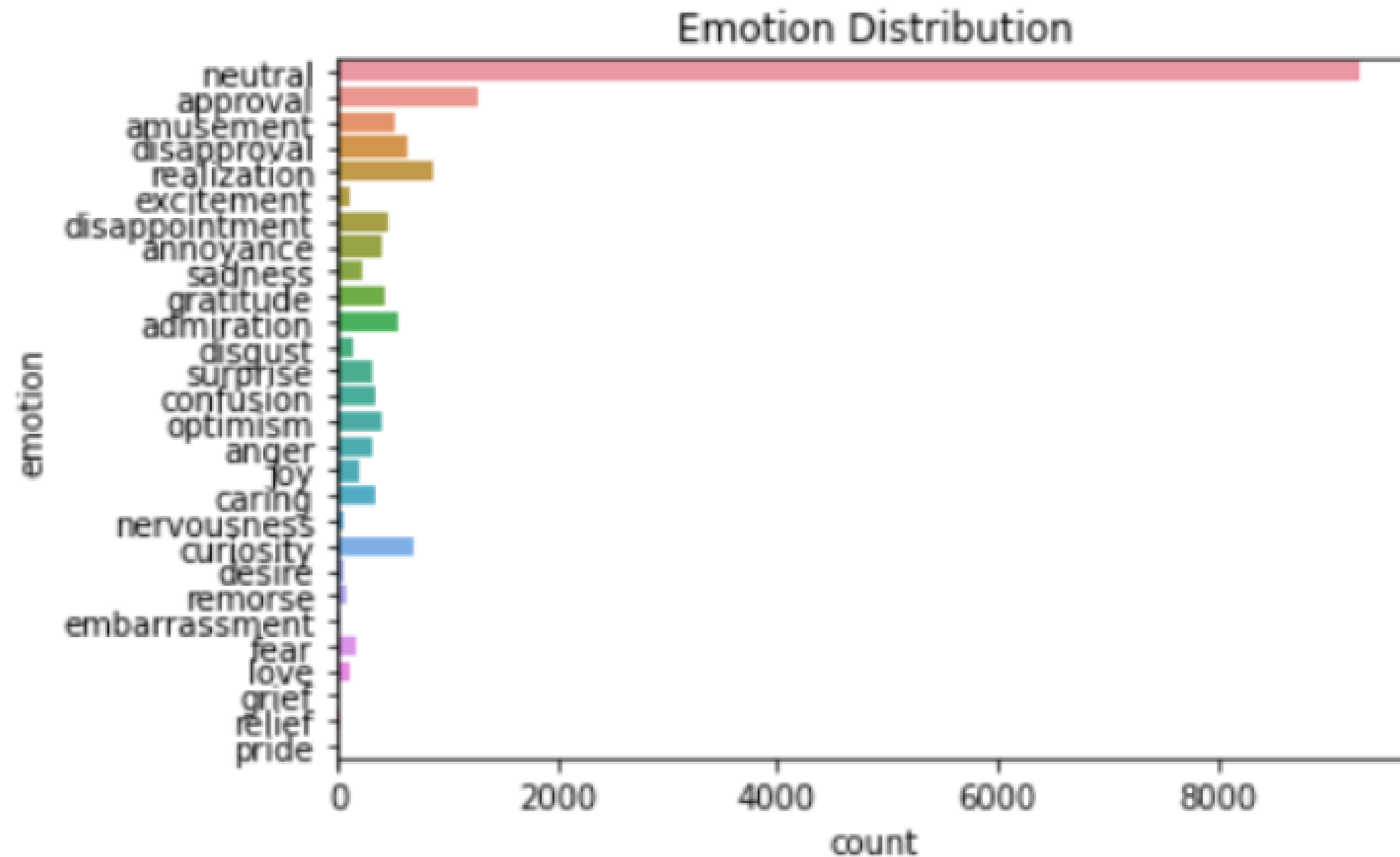
Limitation

The bert-base-uncased tokenization model could only accommodate 512 characters.

Hence, we had to truncate the data, which led to a loss in the amount of data sampled and may have also resulted in misinterpretations of emotion.

GoEmotions

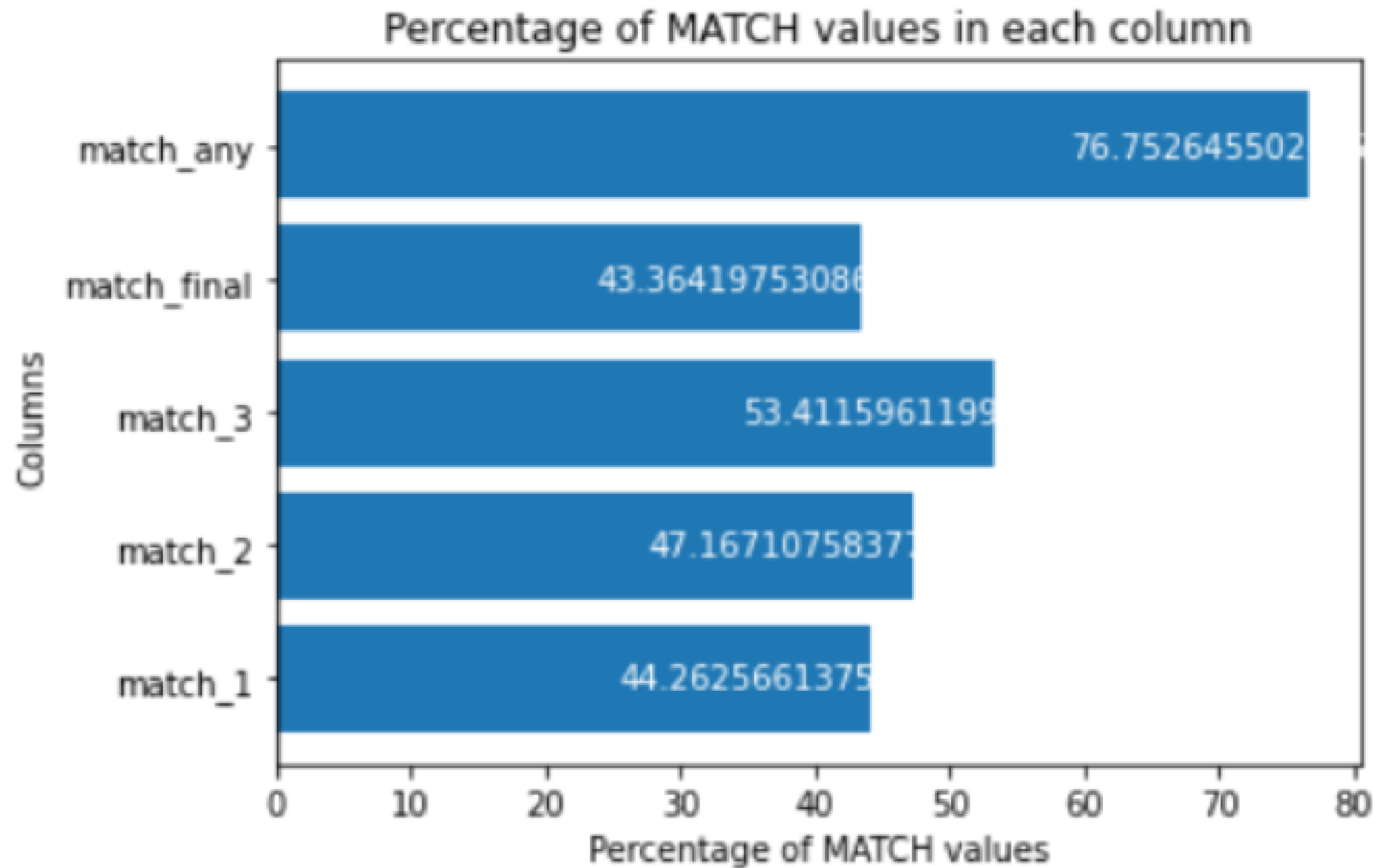
`Text(0.5, 1.0, 'Emotion Distribution')`



Emotion Distribution

- Vast majority of the emotions are neutral
- Hypothesis: Prevalence of un-opinionated articles such as news articles, journal excerpts, etc.

GoEmotions



Comparing GoEmotions with the VADER, TextBlob, and TweetNLP Models

- GoEmotions label shows the highest correspondence with TweetNLP Model at 0.53412.

Conclusion

Raw Data

Reddit + PRAW + keywords

Preparing Data Sets

Pre-processing noisy text
Manual + VADER, TextBlob and
TweetNLP Hugging Face models

Text Normalization

Tokenization → Removal of
Stopwords → Stemming →
Lemmatization

Classification

BERT, Naïve Bayes, Support Vector
Machines, XGBoost

Innovations

Ensemble Classifier, Named Entity
Recognition, GoEmotions

Indexing

Apache Sol-r
HTML, CSS, Javascript,
Bootstrap5, amCharts, AnyChart

Conclusion

- BERT: outperforms traditional ML models in terms of F1 score and accuracy
- Ensemble Classifier: slight decrease in accuracy and F1 score when compared to XGBoost
- NER: did not improve training results and did not incorporate it
- GoEmotions: majority of the texts have been classified as neutral



Thank you!



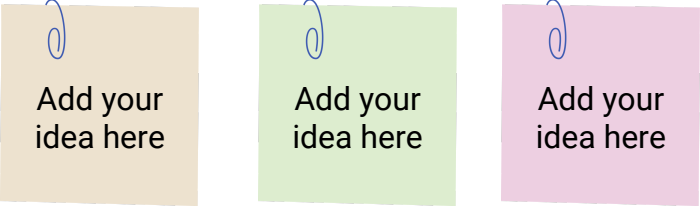
What's your brainstorming topic?

Brainstorm Area

Our Favorite Ideas

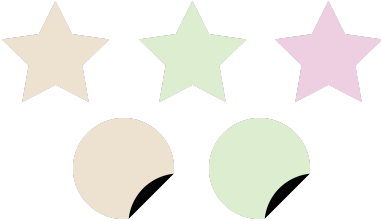
01

Copy a sticky note, and then we'll write our thoughts, ideas, and inspiration.



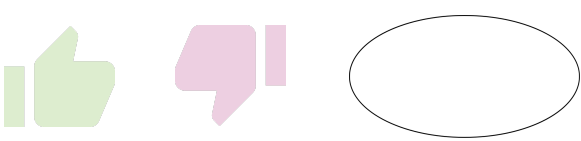
02

Use the stars to vote which ones we like to pursue.



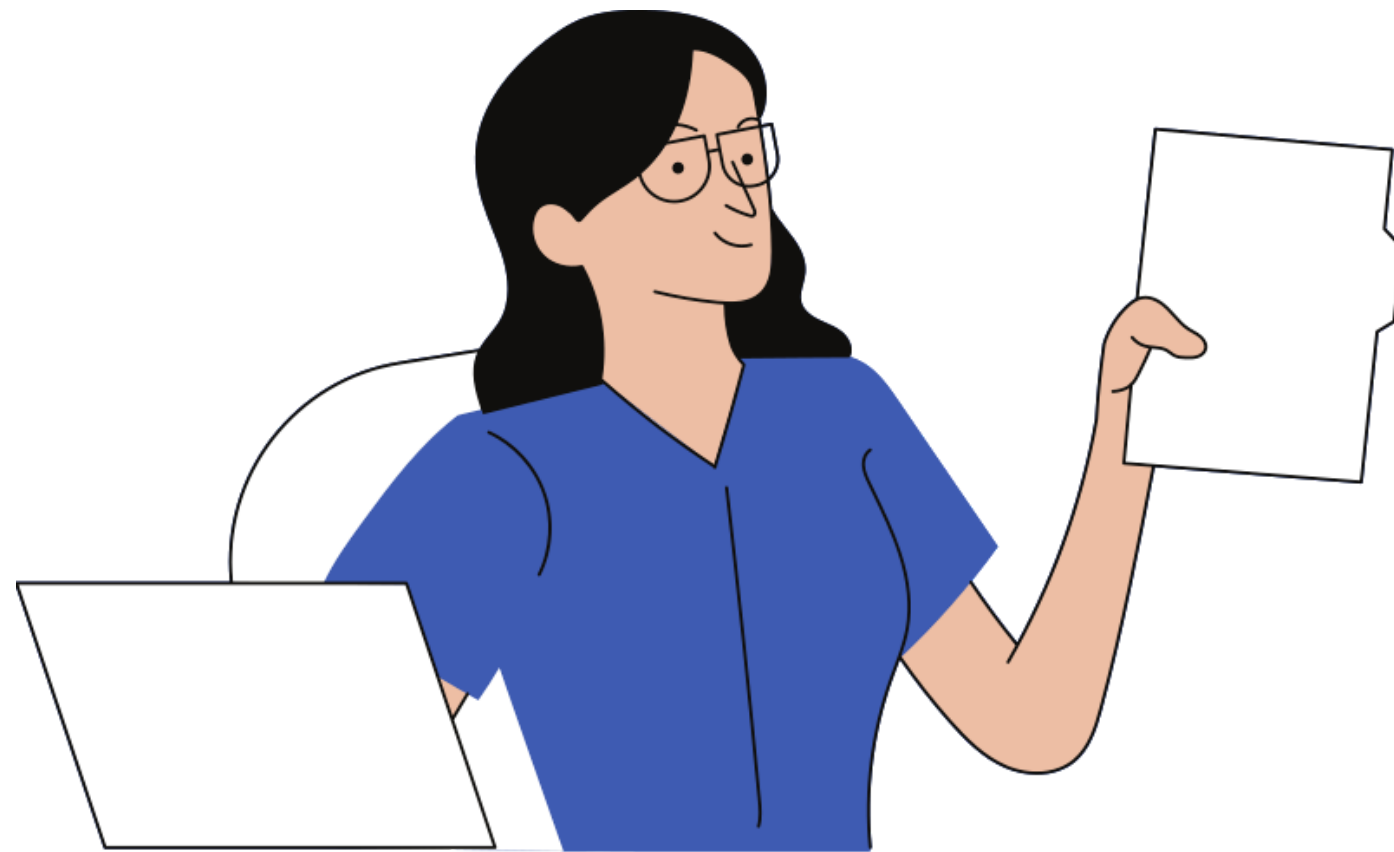
02

Circle or comment on any promising ideas.



Action Items

Let's go back to the previous pages and synthesize what actions are appropriate for moving forward as a group.



01

Write action items in the boxes.



Add your
idea here



Add your
idea here



Add your
idea here

02

Drag your photo under the action item you want to own.

Action Items

01 Write action items in the boxes.



Add your idea here



Add your idea here



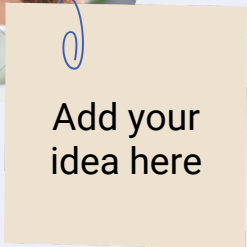

Add your idea here

02 Drag your photo under the action item you want to own.

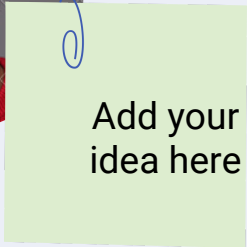



Action 1

Action 2



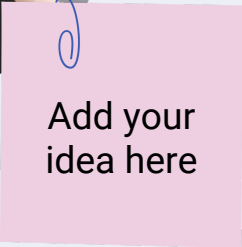
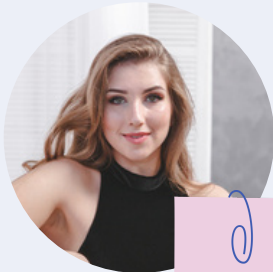
Add your idea here



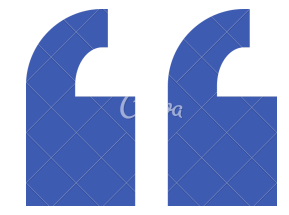
Add your idea here

Action 3

Action 4



Add your idea here



**Anything
worth having
takes time.**

