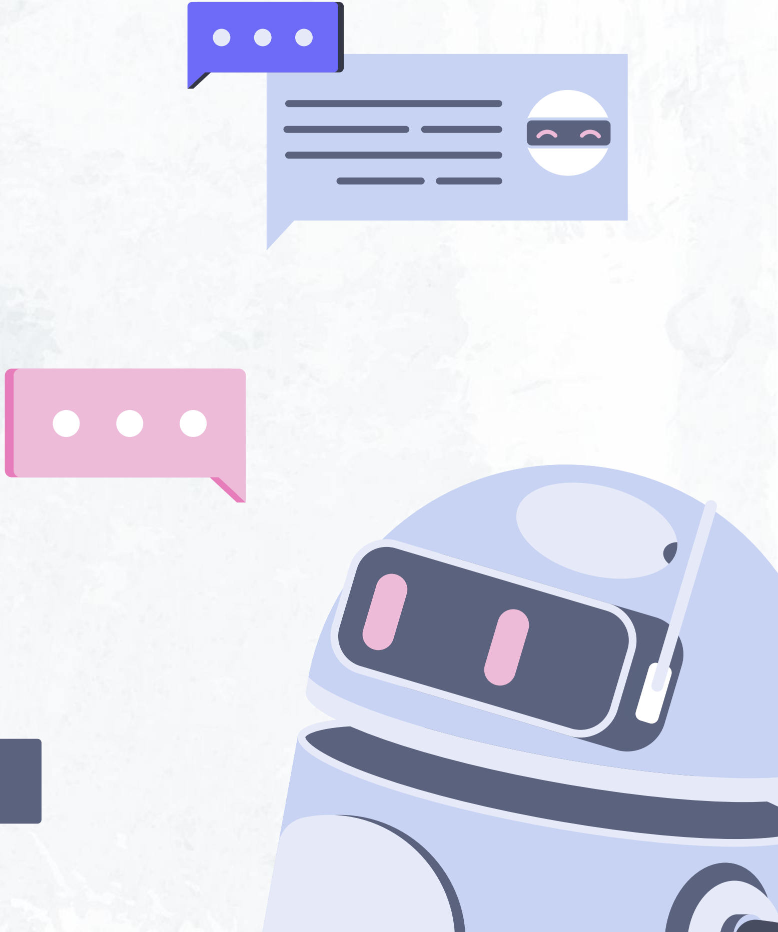
Project 3:
Web APIs & NLP



Group 2 - Brendan, Rebecca, Sherlyn, Sunisa

Table of Contents

01 Background

02 Data Acquisition & Cleaning

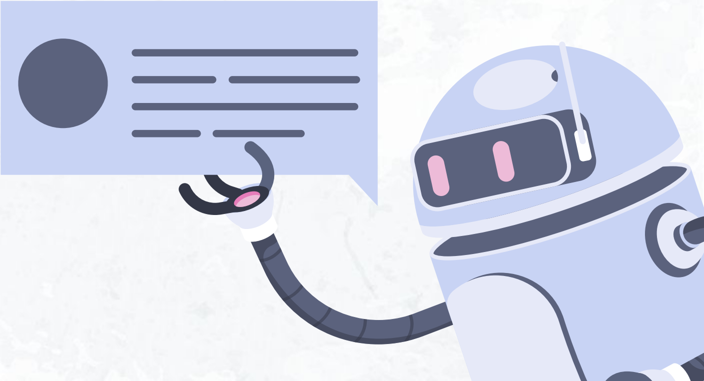
03 Exploratory Data Analysis & Visualizations

04 Text Classification

05 Conclusion & Next Steps

01

Background



Machine Learning & Statistics

Both are closely related and intertwined fields, with ML being heavily based on statistical theories.

Statistics

Contributes key concepts such as probability theory, sampling distributions, statistical inference, and experimental design, which are essential for understanding the theoretical underpinnings and evaluating the efficacy of machine learning models.

Machine learning

Complements by offering computational approaches that can handle large-scale, high-dimensional datasets and complex models and provide powerful tools for automated feature extraction, pattern recognition, and predictive modeling.

Excels at discovering intricate relationships, non-linear dependencies, and intricate structures within data that may be challenging to capture with traditional statistics.

Goal of Project

- Use Natural Language Processing (NLP) on the corpus of data scraped from:
    - r/statistics
    - r/machinelearning
- Explore the intersection of statistics and machine learning and gain a deeper understanding of how these two fields intertwine
- Build a text classifier to classify whether the post belongs to r/machinelearning or r/statistics
- Success of this project will be evaluated:
    - Whether we can identify any distinct topics/communities from the posts scrapped
    - whether we can build a classifier that can accurately (above 90% accuracy) classify a post into r/statistics or r/machinelearning

Techniques

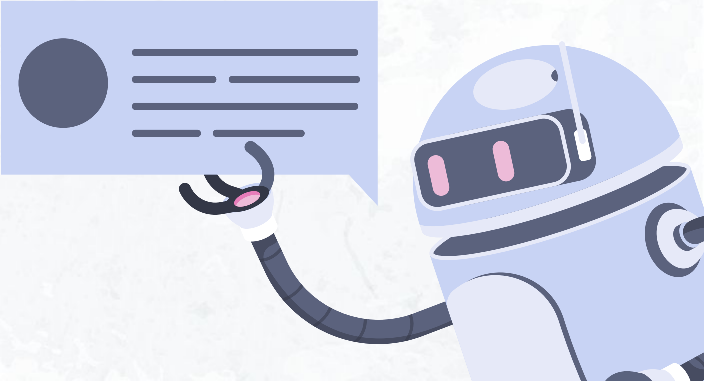
1. Topic Modeling

2. Community Detection

3. Text Classification

02

Data Acquisition & Cleaning



Data Acquisition

The datasets used in this project are obtain from scraping 2 subreddits:

- r/statistics
- r/machinelearning

Data range of posted scrapped from 2023-04-03 to 2023-05-27

Number of unique posts

- r/machinelearning: 975
- r/statistics: 999

Tagging

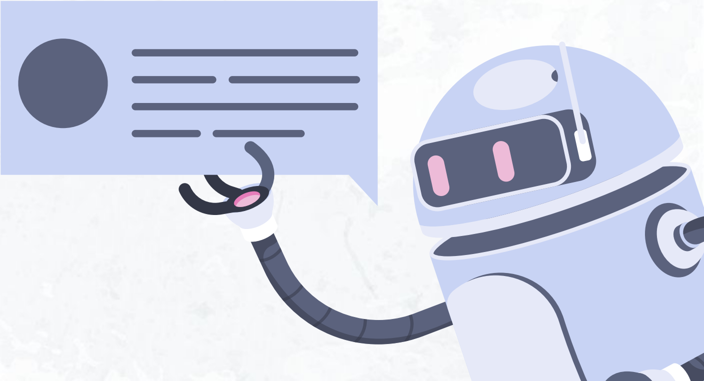
Tag	Abbreviation (r/statistics)	Abbreviation (r/machinelearning)
[Research]	[R]	[R]
[Software]	[S]	
[Question]	[Q]	
[Discussion]	[D]	[D]
[Education]	[E]	
[Career]	[C]	
[Meta]	[M]	
[News]		[N]
[Project]		[P]

Data Cleaning

- Separate the tagging
- Impute the tag for untagged post
- Explore missing content - impute the missing content with nil
- Combine the dataset and create target labels
- Split multiple comments and rejoin to one string
- Tokenize the data

03

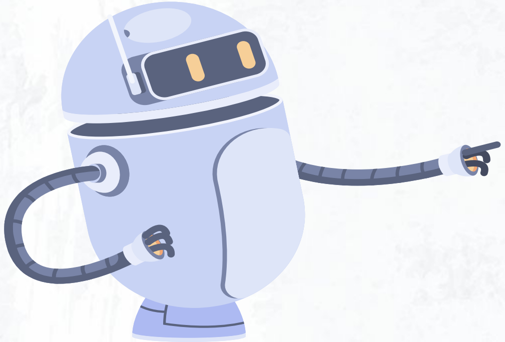
Exploratory Data Analysis & Visualizations



(a) Distribution of Subreddit

(b) Popularity Patterns

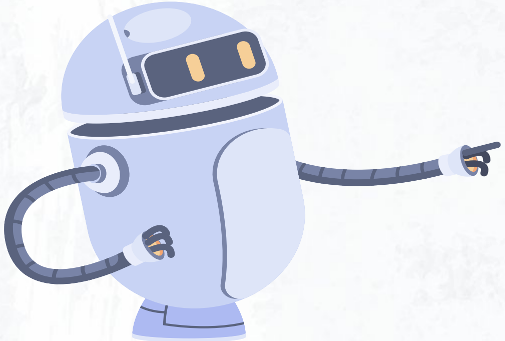
(c) Community Detection



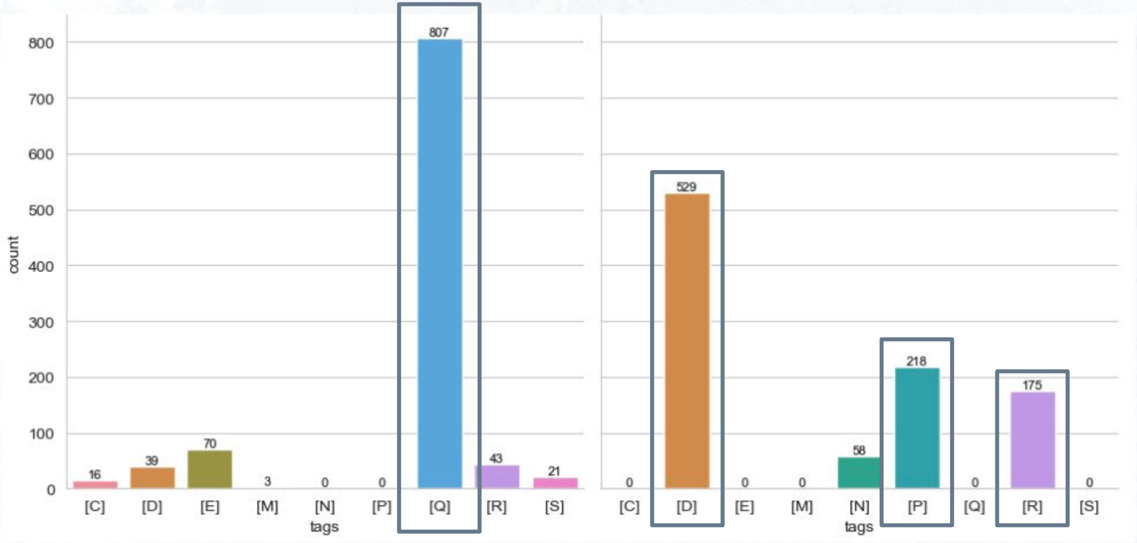
(a) Distribution of Subreddit

(b) Popularity Patterns

(c) Community Detection



Distribution of Subreddit (Post Type)



r/statistics

r/machinelearning

Legend

[C]: Career

[D]: Discussion

[E]: Education

[M]: Meta

[N]: News

[P]: Project

[Q]: Question

[R]: Research

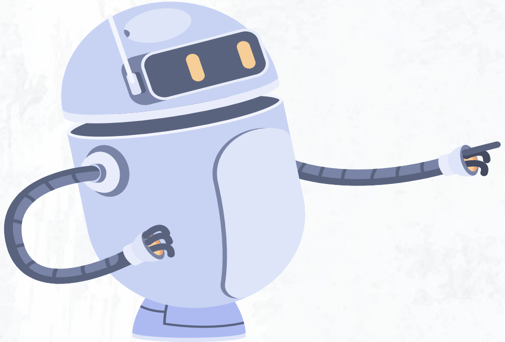
[S]: Software

Posts in r/statistics are predominantly questions, whereas most posts in r/machinelearning are discussions, projects and research.

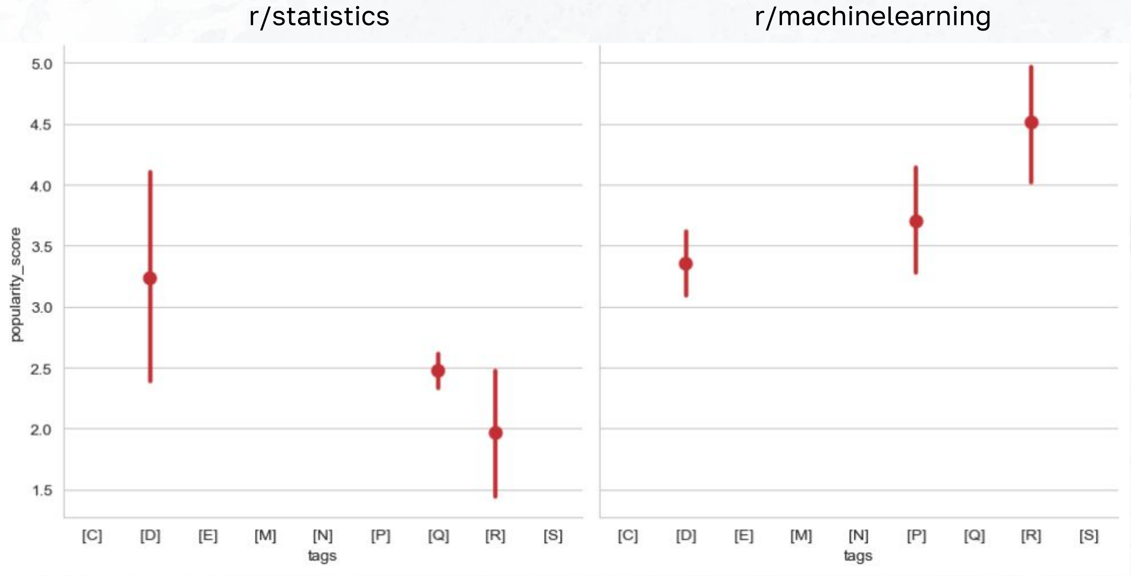
(a) Distribution of Subreddit

(b) Popularity Patterns

(c) Community Detection



Popularity Patterns (Popularity Score)



r/statistics

r/machinelearning

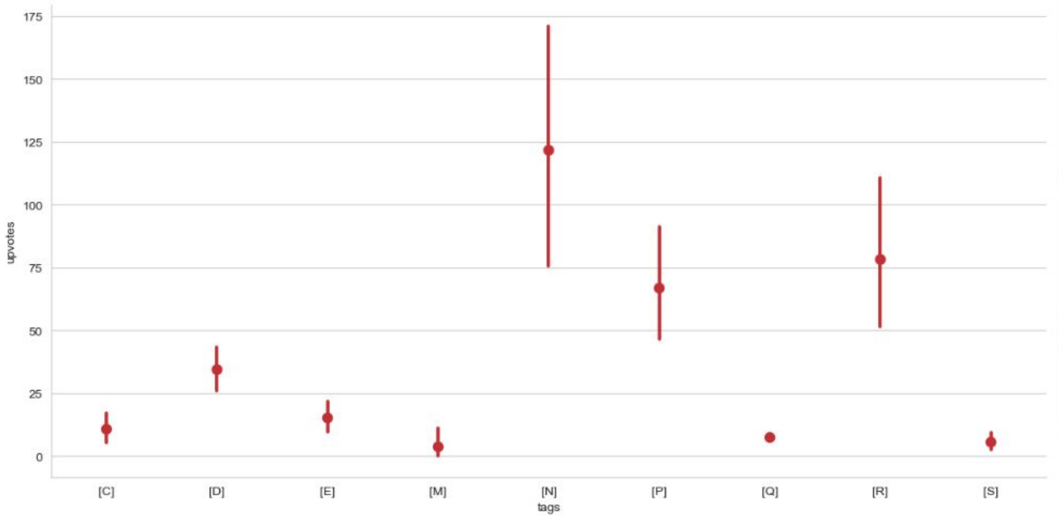
r/statistics

- Questions posted are generally not as popular, as compared to posts related to projects and research

r/machinelearning

- Posts related to discussion, project and research are relatively more popular than those in r/statistics

Popularity Patterns (Upvotes)

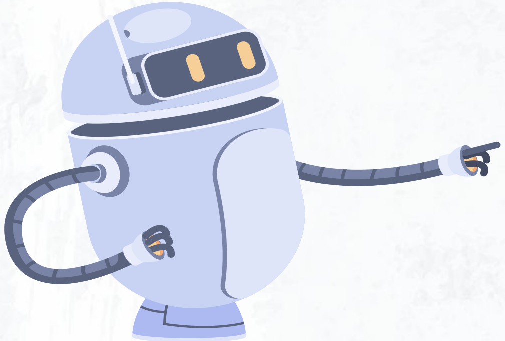


- News has the highest number of overall upvotes, followed by research and projects
- Other post types do not seem to be garnering much attention

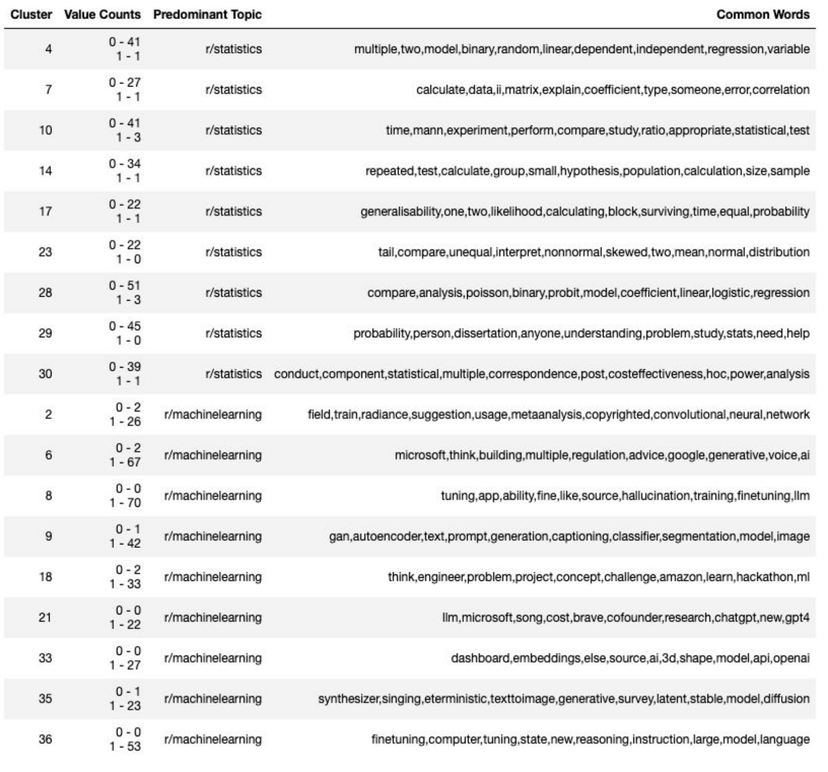
(a) Distribution of Subreddit

(b) Popularity Patterns

(c) Community Detection



Community Detection (Post Title)



Cluster	Value Counts	Predominant Topic	Common Words
4	0 - 41 1 - 1	r/statistics	multiple,two,model,binary,random,linear,dependent,independent,regression,variable
7	0 - 27 1 - 1	r/statistics	calculate,data,ii,matrix,explain,coefficient,type,someone,error,correlation
10	0 - 41 1 - 3	r/statistics	time,mann,experiment,perform,compare,study,ratio,appropriate,statistical,test
14	0 - 34 1 - 1	r/statistics	repeated,test,calculate,group,small,hypothesis,population,calculation,size,sample
17	0 - 22 1 - 1	r/statistics	generalisability,one,two,likelihood,calculating,block,surviving,time,equal,probability
23	0 - 22 1 - 0	r/statistics	tail,compare,unequal,interpret,nonnormal,skewed,two,mean,normal,distribution
28	0 - 51 1 - 3	r/statistics	compare,analysis,poisson,binary,probit,model,coefficient,linear,logistic,regression
29	0 - 45 1 - 0	r/statistics	probability,person,dissertation,anyone,understanding,problem,study,stats,need,help
30	0 - 39 1 - 1	r/statistics	conduct,component,statistical,multiple,correspondence,post,costeffectiveness,hoc,power,analysis
2	0 - 2 1 - 26	r/machinelearning	field,train,radiance,suggestion,usage,metaanalysis,copyrighted,convolutional,neural,network
6	0 - 2 1 - 67	r/machinelearning	microsoft,think,building,multiple,regulation,advice,google,generative,voice,ai
8	0 - 0 1 - 70	r/machinelearning	tuning,app,ability,fine,like,source,hallucination,training,finetuning,llm
9	0 - 1 1 - 42	r/machinelearning	gan,autoencoder,text,prompt,generation,captioning,classifier,segmentation,model,image
18	0 - 2 1 - 33	r/machinelearning	think,engineer,problem,project,concept,challenge,amazon,learn,hackathon,ml
21	0 - 0 1 - 22	r/machinelearning	llm,microsoft,song,cost,brave,cofounder,research,chatgpt,new,gpt4
33	0 - 0 1 - 27	r/machinelearning	dashboard,embeddings,else,source,ai,3d,shape,model,api,openai
35	0 - 1 1 - 23	r/machinelearning	synthesizer,singing,eterministic,texttoimage,generative,survey,latent,stable,model,diffusion
36	0 - 0 1 - 53	r/machinelearning	finetuning,computer,tuning,state,new,reasoning,instruction,large,model,language

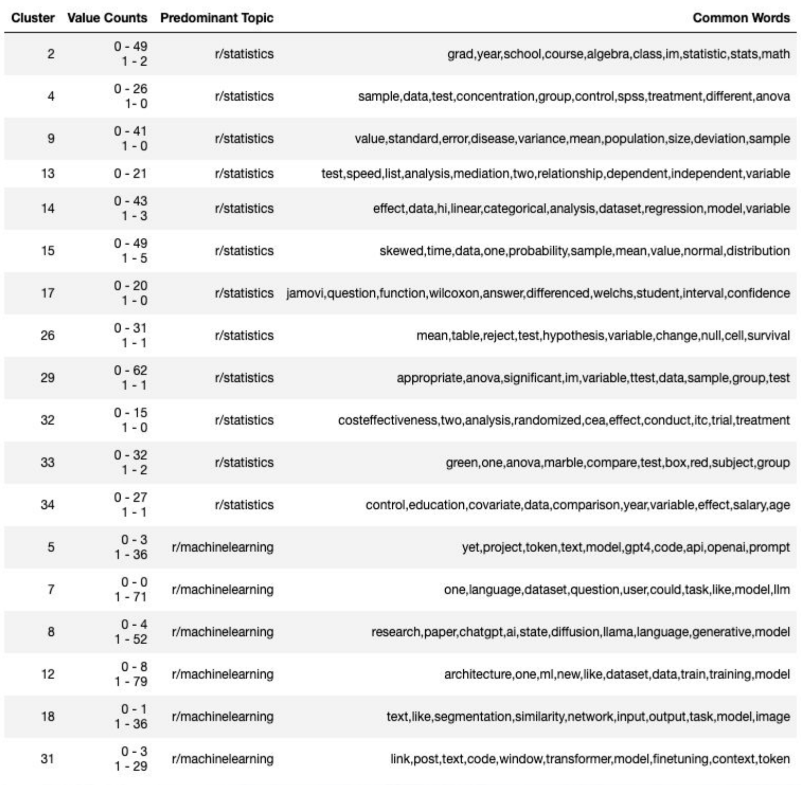
r/statistics:

- Fundamental statistics concepts (keywords such as distribution, mean, poisson, regression, logistic/linear in clusters 4, 23, 28)
- Statistical tests (keywords such as experiment, study, ratio, and test in cluster 10)
- Data interpretation (keywords such as power, correspondence etc in cluster 30)

r/machinelearning:

- Advanced machine learning concepts and applications (keywords such as convolutional, neural, network, llm, embeddings, diffusion in clusters 2, 8, 9, 33 and 35)
- Industry-related topics (by mentions of tech companies such as Microsoft, Amazon, Google)
- Implementation aspects (building, project, challenge, hackathon)

Community Detection (Post Content)



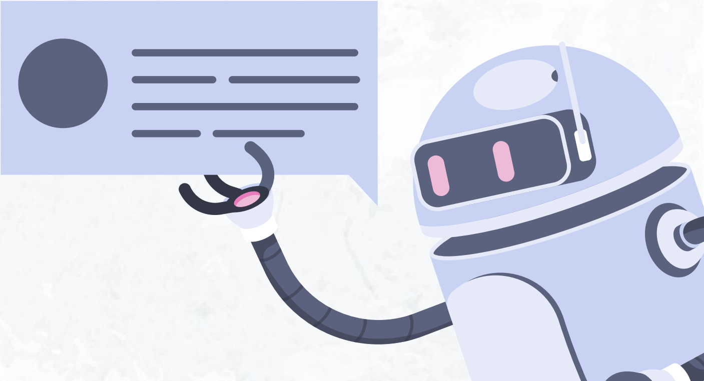
Cluster	Value Counts	Predominant Topic	Common Words
2	0 - 49 1 - 2	r/statistics	grad,year,school,course,algebra,class,im,statistic,stats,math
4	0 - 26 1- 0	r/statistics	sample,data,test,concentration,group,control,spss,treatment,different,anova
9	0 - 41 1 - 0	r/statistics	value,standard,error,disease,variance,mean,population,size,deviation,sample
13	0 - 21	r/statistics	test,speed,list,analysis,mediation,two,relationship,dependent,independent,variable
14	0 - 43 1 - 3	r/statistics	effect,data,hi,linear,categorical,analysis,dataset,regression,model,variable
15	0 - 49 1 - 5	r/statistics	skewed,time,data,one,probability,sample,mean,value,normal,distribution
17	0 - 20 1 - 0	r/statistics	jamovi,question,function,wilcoxon,answer,differenced,welchs,student,interval,confidence
26	0 - 31 1 - 1	r/statistics	mean,table,reject,test,hypothesis,variable,change,null,cell,survival
29	0 - 62 1 - 1	r/statistics	appropriate,anova,significant,im,variable,ttest,data,sample,group,test
32	0 - 15 1 - 0	r/statistics	costeffectiveness,two,analysis,randomized,cea,effect,conduct,itc,trial,treatment
33	0 - 32 1 - 2	r/statistics	green,one,anova,marble,compare,test,box,red,subject,group
34	0 - 27 1 - 1	r/statistics	control,education,covariate,data,comparison,year,variable,effect,salary,age
5	0 - 3 1 - 36	r/machinelearning	yet,project,token,text,model,gpt4,code,api,openai,prompt
7	0 - 0 1 - 71	r/machinelearning	one,language,dataset,question,user,could,task,like,model,llm
8	0 - 4 1 - 52	r/machinelearning	research,paper,chatgpt,ai,state,diffusion,llama,language,generative,model
12	0 - 8 1 - 79	r/machinelearning	architecture,one,ml,new,like,dataset,data,train,training,model
18	0 - 1 1 - 36	r/machinelearning	text,like,segmentation,similarity,network,input,output,task,model,image
31	0 - 3 1 - 29	r/machinelearning	link,post,text,code,window,transformer,model,finetuning,context,token

- Common words across both subreddits (keywords such as feedback, thread, feature, observation, data, linear, categorical, analysis, regression, model, variable) indicate similarities in terms of interaction patterns, focus on data and feature analysis, and discussions around modeling and regression

- Post contents across both subreddits differs in a similar way revealed in the earlier analysis for post title

04

Text Classification



Phased Training Approach

(P1) Text data

Trained using tokenized title, content and top (5) comments concatenated

(P2) Text data + tags (category)

Trained using (a), plus tags One Hot Encoded

(P3) Text data + tags + number of comments + upvotes

Trained using (b), plus number of comments and upvotes of post

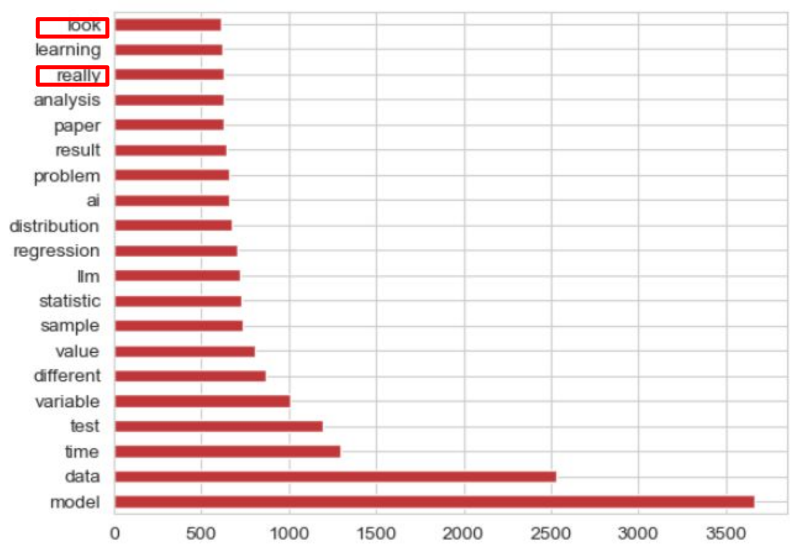
Preprocessing

(All) train-test-split

(P1) We iteratively visualize the top 20 keywords present in the text data, identify stop words, and add them to our custom list of stop words and re-visualize the top 20 words, until there are no stop words left.

(P1) Tf-idf Vectorizer to assign weights to different words

(P2) One Hot Encoding of Categorical Data

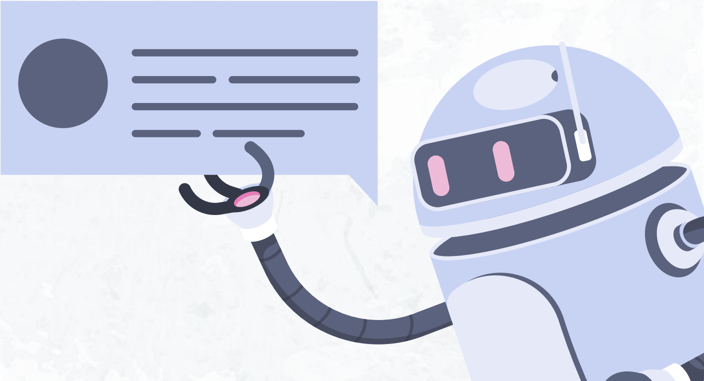


Modeling Results

Model	**Text (P1)**	**P1 + tags (P2)**	**P2 + upvotes + comments**
Logistic Regression	0.960	0.982 ▲	0.985 ▲
Support Vector Classifier	0.958	0.980 ▲	0.979 ▼
Decision Tree Classifier	0.871	0.964 ▲	0.962 ▼
Random Forest Classifier	0.958	0.977 ▲	0.977
AdaBoost Classifier	0.950	0.980 ▲	0.980
GradientBoost Classifier	0.948	0.968 ▲	0.969 ▲

05

Conclusion & Next Steps



Conclusion

- Scraped ~1000 reddit post from r/machinelearning and r/statistics
- Analyzed the posts to better identify the commonalities and differences in topics between the 2 subreddit threads
- Identify common topics and communities to better understand the distinctions and intersections between the 2 threads

Conclusion

Topic Modelling

r/statistics

Topic	Words
Statistical methods and tests	"regression", "correlation", "test", "hypothesis", "probability"
Data analysis techniques	"model", "analysis", "calculate", "coefficient"
Study design and experiment management	"training", "project", "hackathon", "app", "dashboard"
Statistical education and help	"explain", "help", "understanding", "need"

r/machinelearning

Topic	Words
Machine Learning Models & Techniques	"neural", "network", "classifier", "segmentation", "model", "fine tuning", "gan", "autoencoder", "embeddings", “neural networks”, “GANs (Generative Adversarial Networks), autoencoders, and model fine-tuning”
Specific AI technologies, platforms and companies	"Microsoft", "Google", "OpenAI", "GPT4", "LLM"
Applications and projects	"training", "project", "hackathon", "app", "dashboard"
Machine Learning Research	"research", "generation", "latent", "diffusion", "field", "study"

Conclusion

Community Detection

r/statistics

Topic	Words
Students and Learners	"calculate", "explain", "need", "help", "understanding", "study"
Educators and Professionals	"model", "analysis", "hypothesis", "regression", "probability"
Practicing researchers	"experiment", "population", "sample", "study", "test"

r/machinelearning

Topic	Words
Machine Learning Practitioners	"training", "neural", "network", "model", "classifier", "fine tuning"
Researchers and Academics	"research", "latent", "field", "study", "generation", "diffusion"
Industry Professionals	"Microsoft", "Google", "Amazon", "project", "app", "dashboard"
Students and Learners	"learn", "problem", "concept", "challenge"

Conclusion

Text Classification

- Build accurate text classifier
    - Post contents (i.e., title, content and top comments, extracted tags) are sufficient to analyze posts and identify the commonalities and differences in topics between the 2 subreddit threads
- High model accuracy (~98%)
    - LogisticRegression, Support Vector Classifier
- KMeans clustering to identify distinct clusters

Next Steps

- Temporal analysis
  - Analyze trends in community interactions and topics overtime
- Deepen community analysis
  - Analyze user activity patterns within each subreddit, e.g., response times, active contributors
  - Understanding of community dynamics
- Visualization
  - Interactive dashboards that are more comprehensible and engaging for a wider audience

Thank You

CREDITS: This presentation template was created by **Slidesgo** and includes icons by **Flaticon**, infographics & images by **Freepik** and content by **Eliana Delacour**