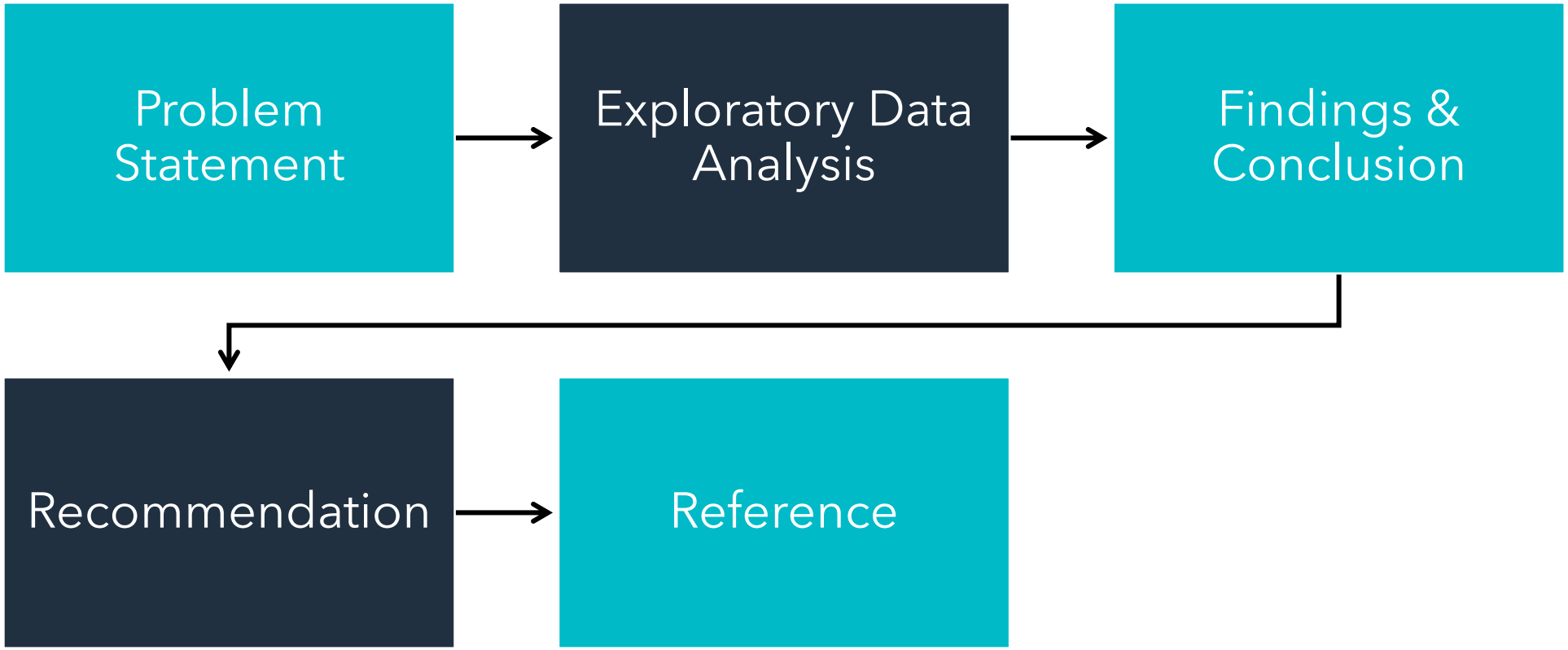
Survey on Subreddit Post Using NLP approach

By: Faiz Puad

Agenda


Problem Statement
Exploratory Data Analysis
Findings & Conclusion
Recommendation
Reference

Problem Statement

Findings the **latest update** of **career pathway** is a **must.** However, industry such as digital keeps growing at **fast pace** for the past few years.

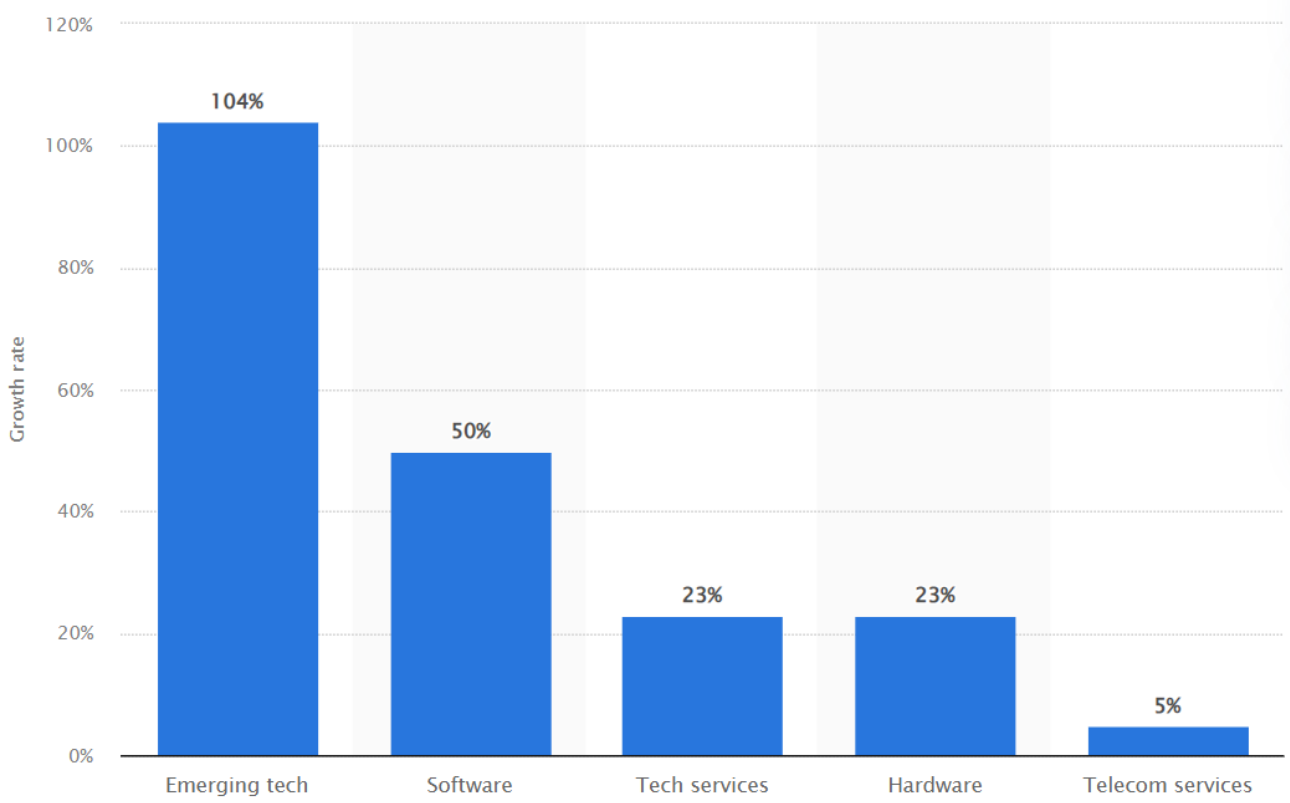


Image from:
Statistica.com
- IT industry growth forecast by segment

Rely On **Traditional** Approach**:**

Refer to online **articles**

Pamphlet online

Depend on available **survey**

Might not feasible as it is **time consuming** and might even be **costly**

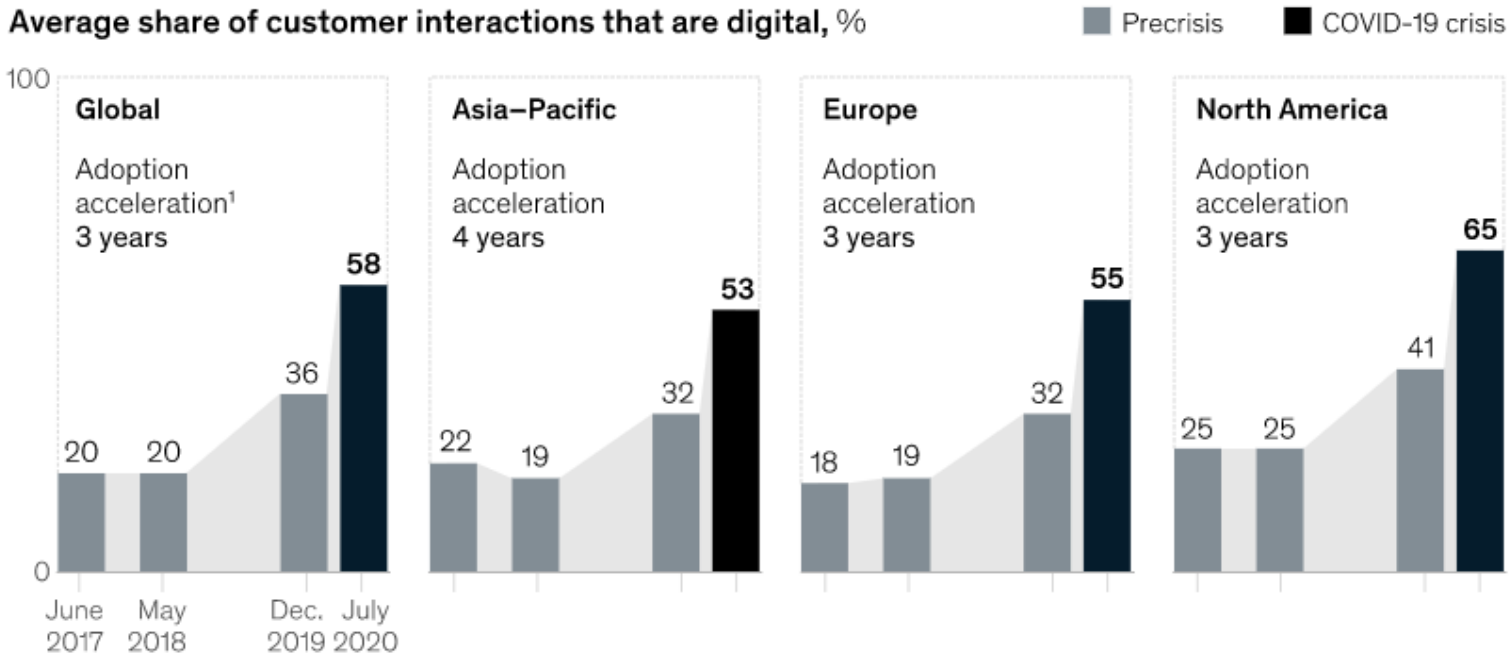


Image from:
mckinsey.com
- Covid 19 impact on technology acceleration

Exploratory Data Analysis

DATA
OVERVIEW

DATA
CLEANING

FEATURES
ENGINEERING

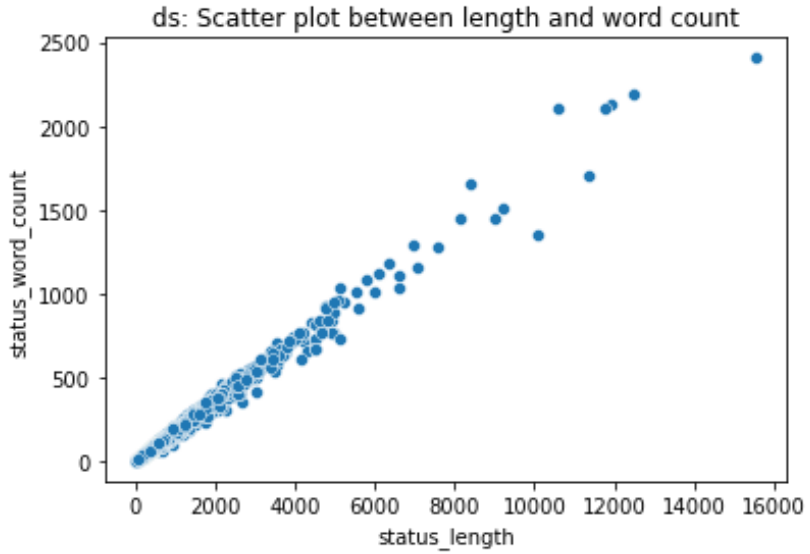
MODEL
EVALUATION

Data Overview

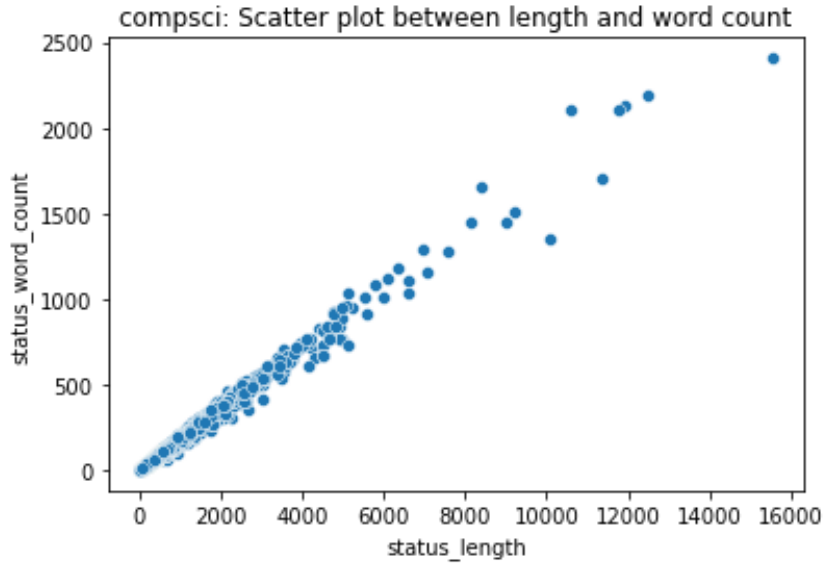
- **Data science subreddit**
- 10000 rows, 16 columns
- Unique author: 6968
- Top author post: 33
- Average comment /post: 2
- Highest comment/post: 346
- Lowest subscribers: 608,935
- Highest subscribers: 783,616

- **Computer science subreddit**
- 15000 rows, 16 columns
- Unique author: 15080
- Top author post is 471
- Average comment /post: 5
- Highest comment/post: 341
- Lowest subscribers: 308,929
- Highest subscribers: 2 million

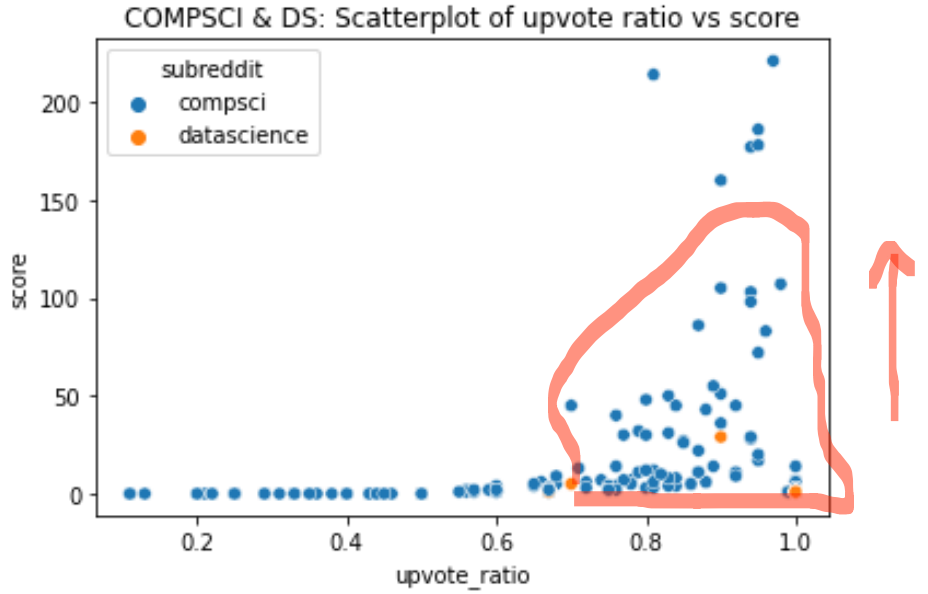
A


ds: Scatter plot between length and word count
status_word_count
status_length
0
500
1000
1500
2000
2500
0
2000
4000
6000
8000
10000
12000
14000
16000

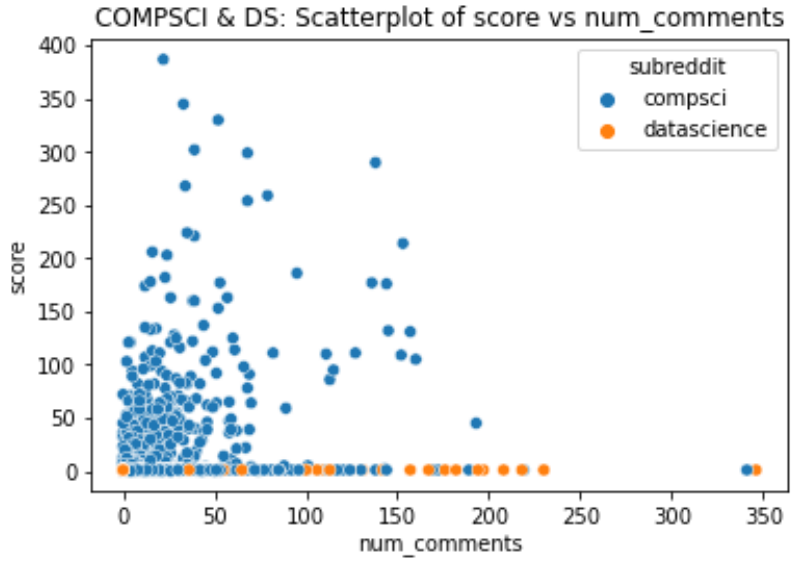
B

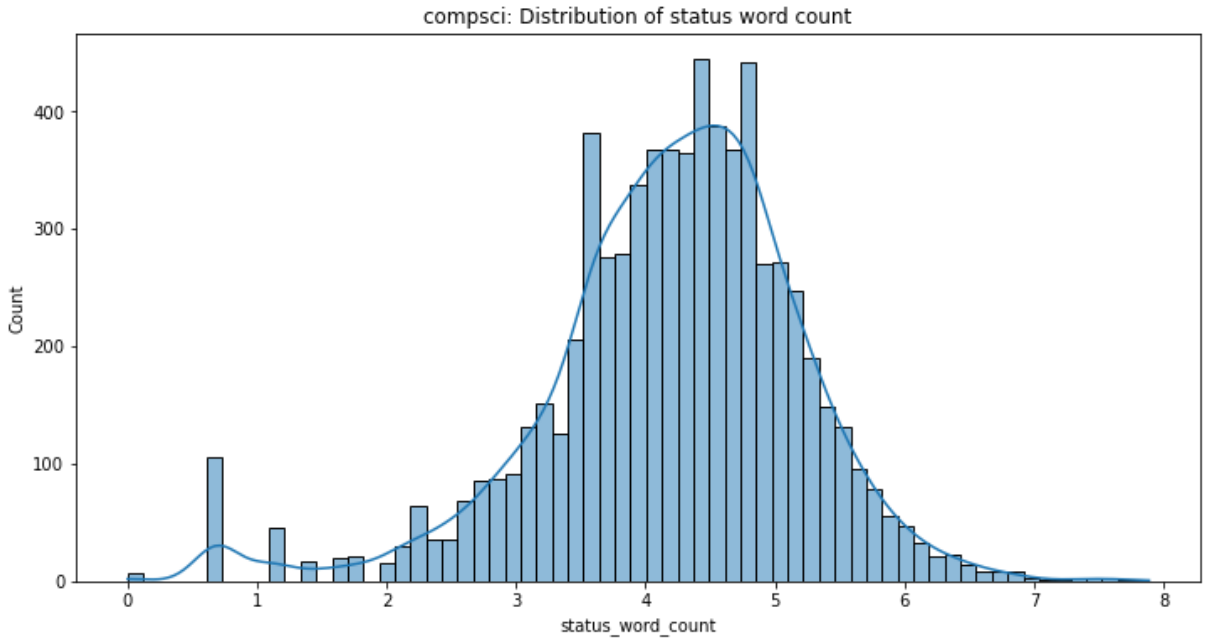

compsci: Scatter plot between length and word count
status_word_count
status_length
0
500
1000
1500
2000
2500
0
2000
4000
6000
8000
10000
12000
14000
16000

C


COMPSCI & DS: Scatterplot of upvote ratio vs score
subreddit
compsci
datascience
score
upvote_ratio
0
50
100
150
200
0.2
0.4
0.6
0.8
1.0

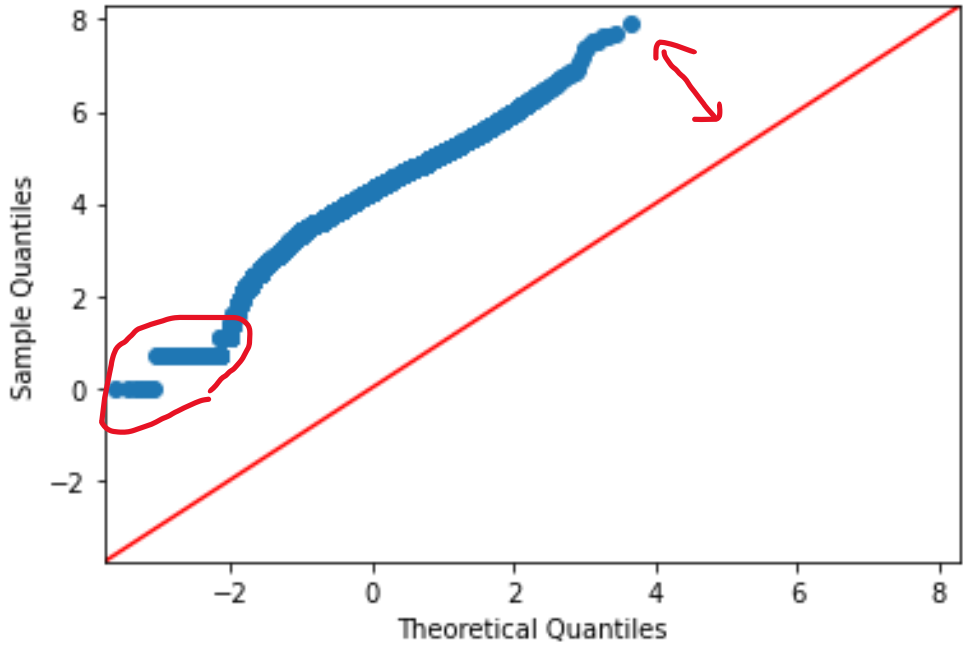
D


COMPSCI & DS: Scatterplot of score vs num_comments
subreddit
compsci
datascience
score
num_comments
0
50
100
150
200
250
300
350
400
0
50
100
150
200
250
300
350

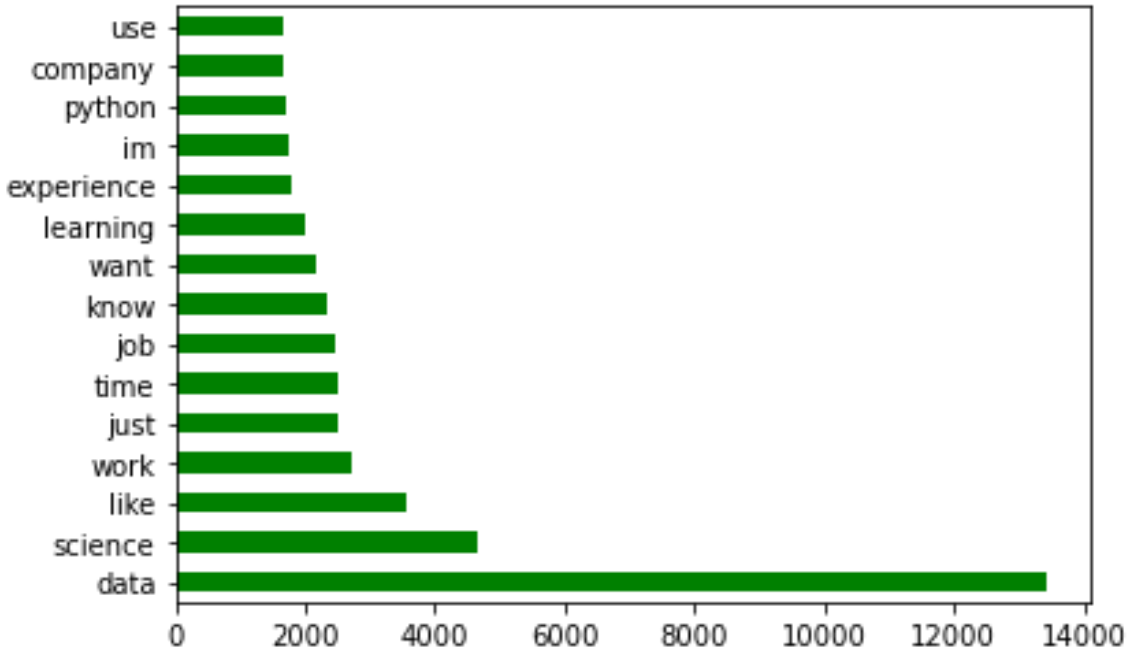


Original:
Skewness: 6.788906574552096
Kurtosis: 86.11276298774011

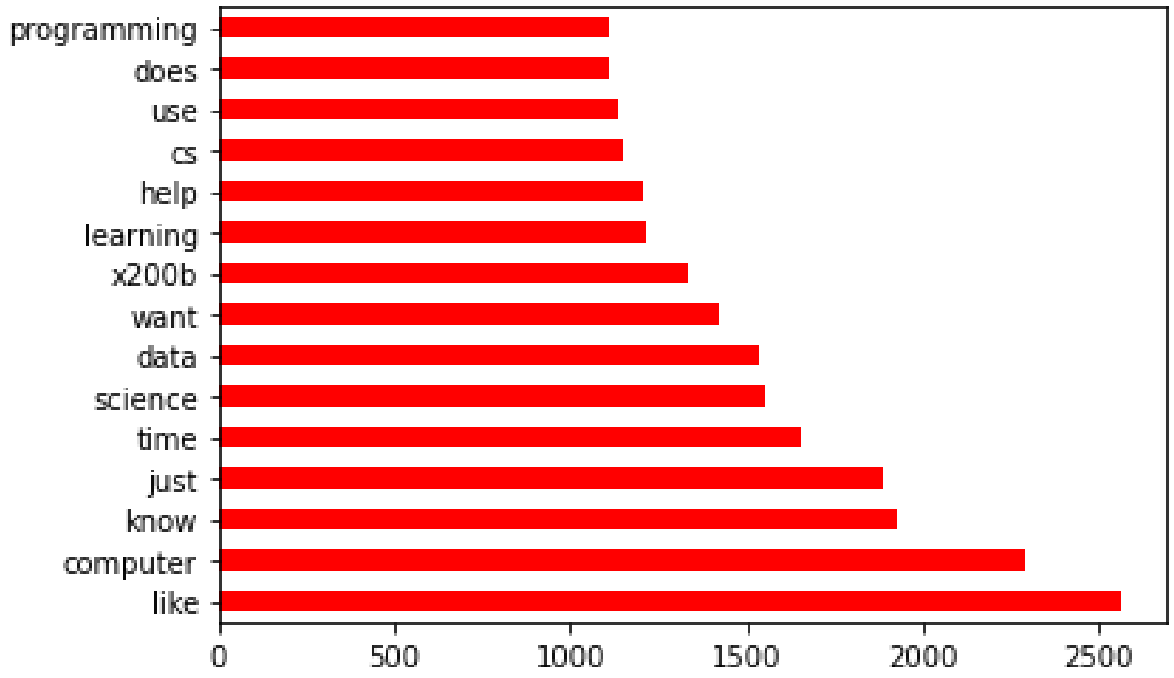
After log:
Skewness: -0.8105170015250748
Kurtosis: 1.8747987098655061



Data Science Subreddit



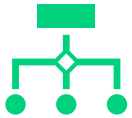
Computer Science Subreddit



Data Cleaning

Drop rows with null values & status ['deleted']

Compsc row removed: 73



New shape is:

Ds(rows, columns), Compsc(rows,columns)

(8160, 12), (7016, 12)

Drop 1144 rows in DS to balance dataset

Remove symbol, emoji, hyperlink using regex and custom function

Include: "https", "www.", "$", "@"

"Welcome to this week's entering & transitioning thread! This thread is for any questions about getting started, studying, or transitioning into the data science field. Topics include:\n\n* Learning resources (e.g. books, tutorials, videos)\n* Traditional education (e.g. schools, degrees, electives)\n* Alternative education (e.g. online courses, bootcamps)\n* Job search questions (e.g. resumes, applying, career prospects)\n* Elementary questions (e.g. where to start, what next)\n\nWhile you wait for answers from the community, check out the [FAQ](https://www.reddit.com/r/datascience/wiki/frequently-asked-questions) and Resources pages on our wiki. You can also search for answers in [past weekly threads](https://www.reddit.com/r/datascience/search?q=weekly%20thread&restrict_sr=1&sort=new)."

"welcome to this week's entering & transitioning thread! this thread is for any questions about getting started, studying, or transitioning into the data science field. topics include: * learning resources (e.g. books, tutorials, videos) * traditional education (e.g. schools, degrees, electives) * alternative education (e.g. online courses, bootcamps) * job search questions (e.g. resumes, applying, career prospects) * elementary questions (e.g. where to start, what next) while you wait for answers from the community, check out the (and (resources) pages on our wiki. you can also search for answers in ("

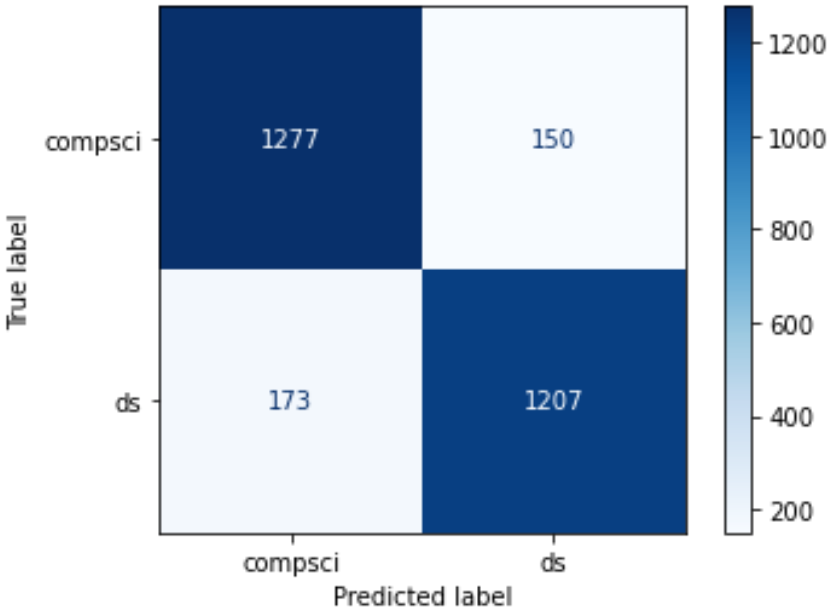
Feature Engineering

- Apply PorterStemmer on selftext column
  - Change word to its base form
  - Flies > fly
- Encode target variable into 1 (compsc) & 0 (datascience)
- Train & test split data with test size 20%
- Instantiate transformer and estimator inside pipeline
- Transformer use: CountVectorizer & TfidVectorizer
- Estimator use: Multinomial Naïve Bayes

Model Evaluation

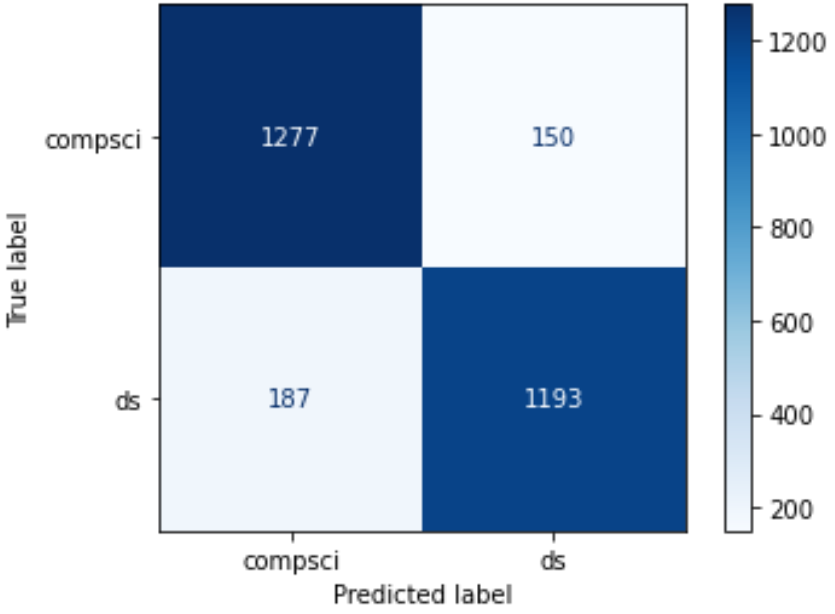
Test size	Row	Estimator	Transformer	Train Score	Test Score	ROC_AUC Score
0.3	2000	Multinomial NB	Count Vectorize	0.903	0.835	
0.3	2000	Multinomial NB	TfidVectorize	0.918	0.840	
0.3	4000	Multinomial NB	Count Vectorize	0.887	0.873	
0.3	4000	Multinomial NB	TfidVectorize	0.897	0.871	
0.3	6000	Multinomial NB	Count Vectorize	0.890	0.870	
0.3	6000	Multinomial NB	TfidVectorize	0.890	0.870	
0.2	6000	Multinomial NB	Count Vectorize	0.884	0.879	0.8797
0.2	6000	Multinomial NB	TfidVectorize	0.888	0.885	0.8848

Multinomial + CountVectorize



True Negatives: 1277
False Positives: 150
False Negatives: 173
True Positives: 1207
Specificity: 0.8949
Precision: 0.8895
Sensitivity: 0.8746
F1 score: 0.882
ROC AUC score: 0.8848

Multinomial + TfidVectorize



True Negatives: 1277
False Positives: 150
False Negatives: 187
True Positives: 1193
Specificity: 0.8949
Precision: 0.8883
Sensitivity: 0.8645
F1 score: 0.8762
ROC AUC score: 0.8797

Conclusion

- Multinomial Naïve Bayes with TfidVectorizer is chosed as the best model as it produce the highest accuracy mark for this study.
- Model is slightly overfit and can be further improve using other classification model
- Few possible reason are:
  - train data is not enough. This is proved as I increase from initial data of 2000 to 6000, train score drop from 95% to 88% but increase test score from 80% to 88%
  - data still contain unnecessary symbol including emoji which cause inaccurate prediction

Recommendation

- Model found to be useful for detecting sentiment and classifying a post into category. Thus, it can be fitted generally to all industry such as marketing to promote their product.
- Model results can be utilized as providing snapshot of a person to be filter for job application based on finding above.
- Introduce more data to fight overfit issue and there is possibility model perform higher than 90% using other model including decision tree, random forest and adaboost

Reference

- How COVID-19 has pushed companies over the technology tipping point—and transformed business forever. (2021, February 18). McKinsey & Company. Available at:
  - https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/how-covid-19-has-pushed-companies-over-the-technology-tipping-point-and-transformed-business-forever#
- Statista. (2022, February 21). IT industry growth rate forecast worldwide from 2018 to 2023, by segment. Available at:
  - https://www.statista.com/statistics/967095/worldwide-it-industry-growth-rate-forecast-segment/
- Varshney, P. (2021b, December 14). Q-Q Plots Explained - Towards Data Science. Medium. Available at:
  - https://towardsdatascience.com/q-q-plots-explained-5aa8495426c0
- Yadav, K. (2022, January 22). Cleaning & Preprocessing Text Data by Building NLP Pipeline. Medium. Available at:
  - https://towardsdatascience.com/cleaning-preprocessing-text-data-by-building-nlp-pipeline-853148add68a