From Data to Careers: A Subreddit Journey through r/datascience and r/jobs

Anthony Amadasun

TABLE OF CONTENTS

01

Overview

Introduction, problem statement, project objective, deeper goals

Overview

Extract valuable insights from two diverse subreddits, r/datascience and r/job:

1. What linguistic nuances and thematic elements differentiate posts from the r/datascience and r/jobs subreddits, and how can these distinctions be harnessed to construct a reliable classification model?
2. What distinctive characteristics, as reflected in these posts, define an ideal data scientist for our startup, as perceived by the hiring team?
3. Apply Natural Language Processing (NLP) techniques like Latent Dirichlet Allocation, stop word removal, stemming and lemmatization to understand and classify posts.

Overview

01

Uncover insights into the job market for data scientists

02

Understand challenges faced by data scientists and job seekers

03

Inform HR about desired skills and qualifications

04

Build classifiers to determine the origin of a post

05

Gain understanding of language and concerns in online communities

06

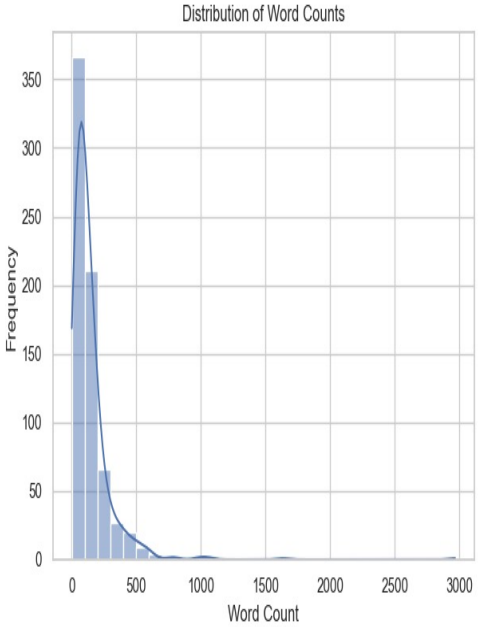
Provide a practical tool for our HR seeking tailored information

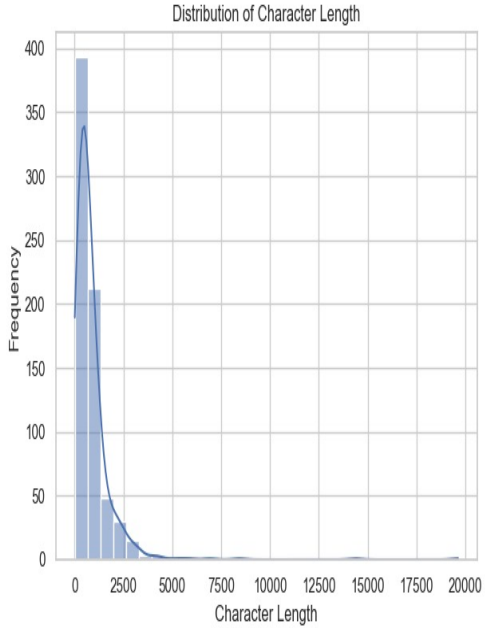
02

Methodology

Data wrangling, gathering, and acquisition

Data Wrangling/Gathering/Acquisition

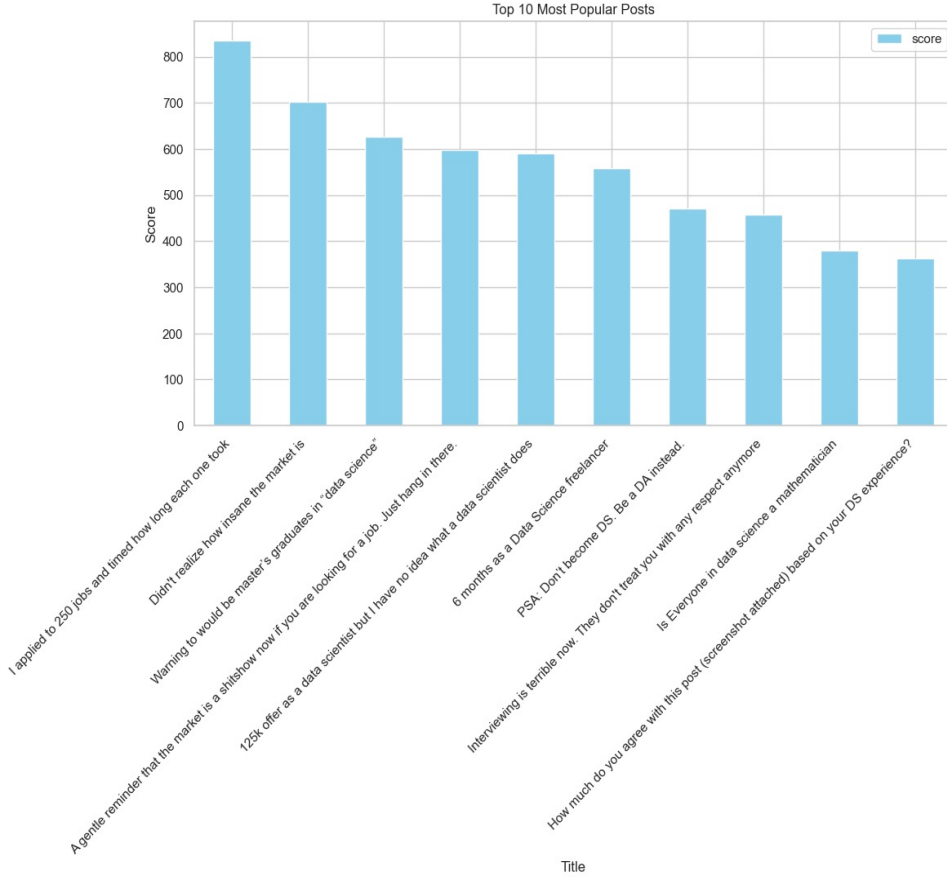




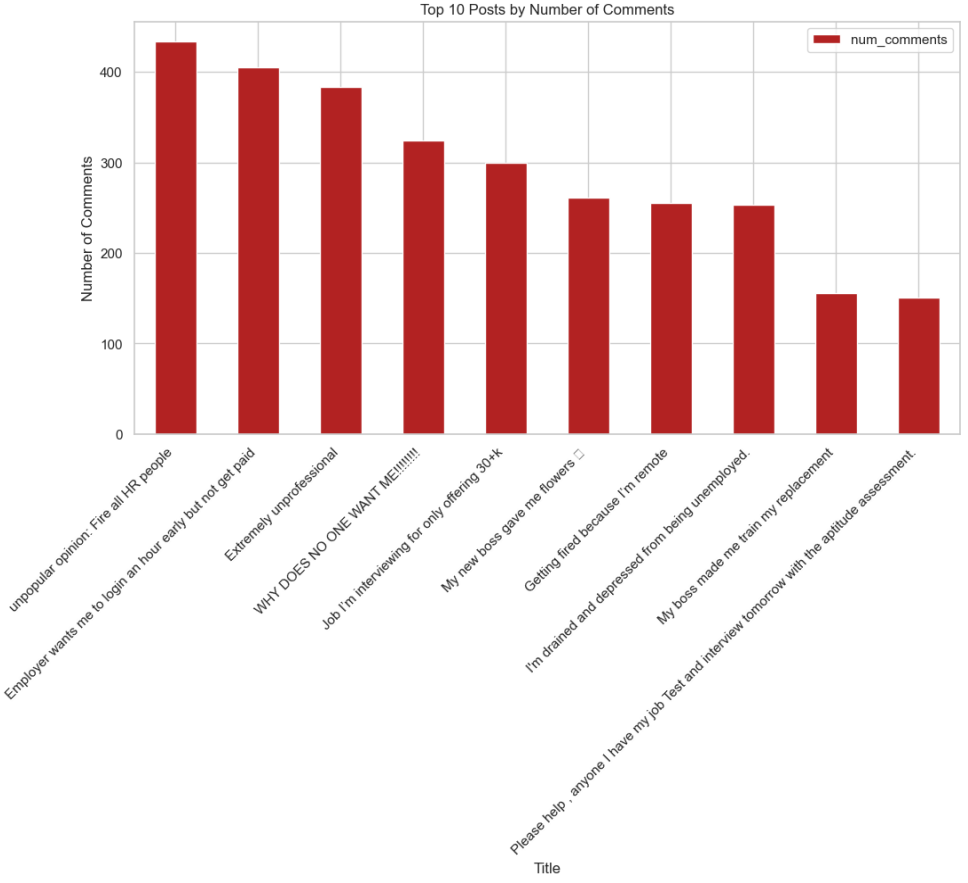
- Extracted data had 3000 rows and 116 columns for r/datascience and 3000 rows and 96 columns for r/jobs, before cleaning and eda.
- After cleaning and eda, the rows dropped to 711 rows and 98 for r/datascience and 868 rows and 96 columns for r/jobs
- Main reason was due to removing duplicated post, dropping columns with missing values, and post where word count was 0.

Data Wrangling/Gathering/Acquisition

r/datascience



r/jobs



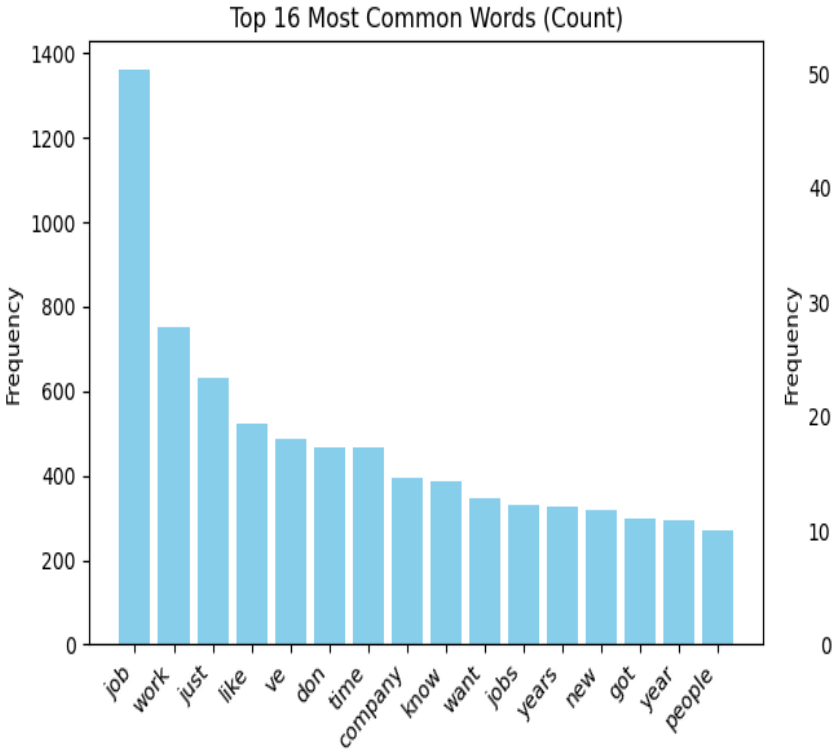
03

NLP Technique

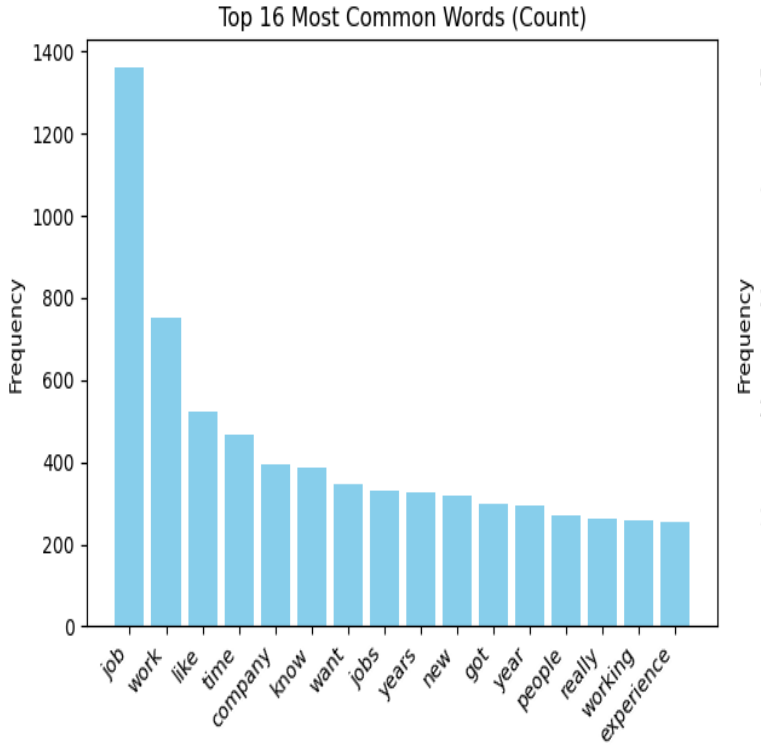
Topic Modeling through LDA, word frequency analysis, word cloud

Preprocessing

Before processing



After processing



Topic Modeling: Latent Dirichlet Allocation (LDA)

r/datascience wordcloud

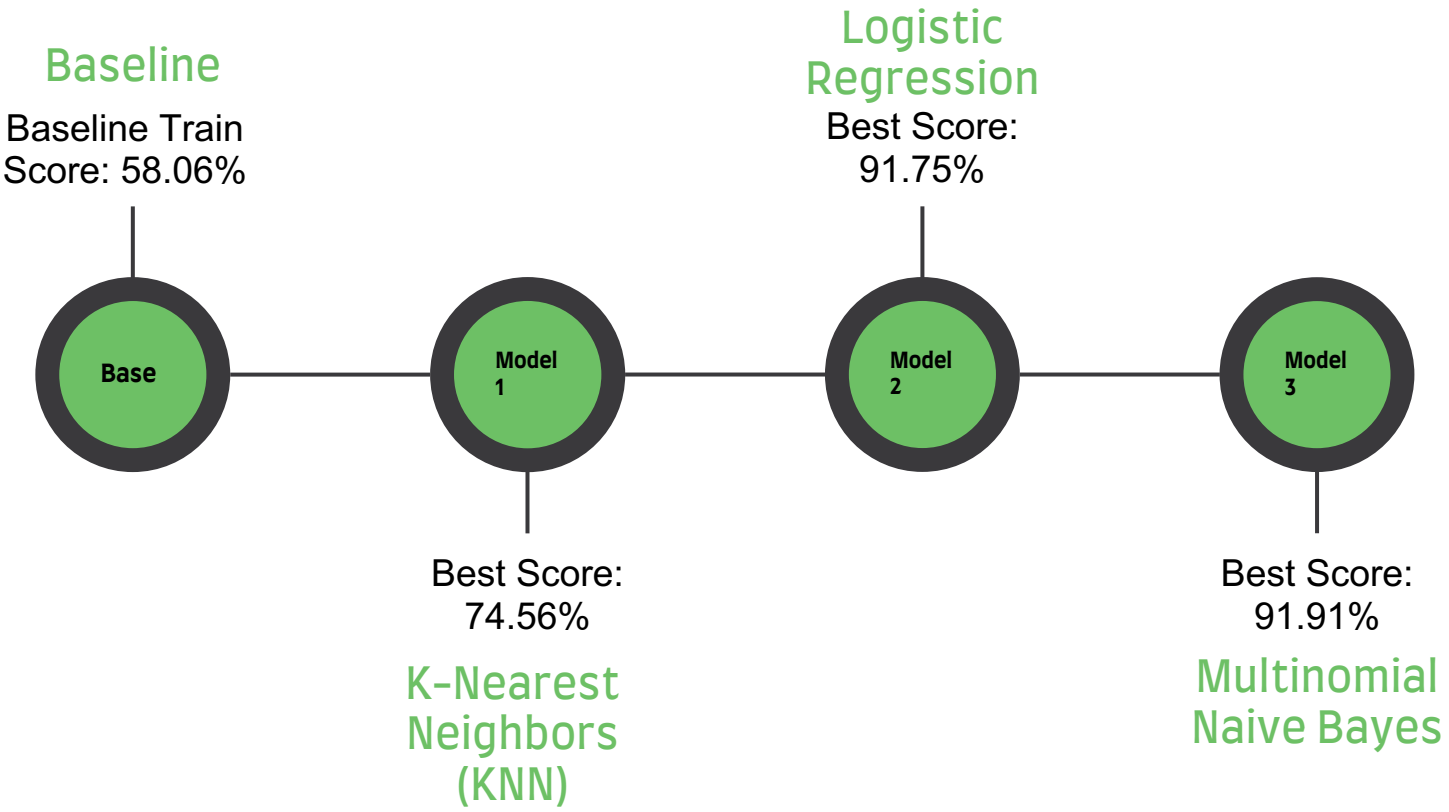
r/jobs wordcloud

04

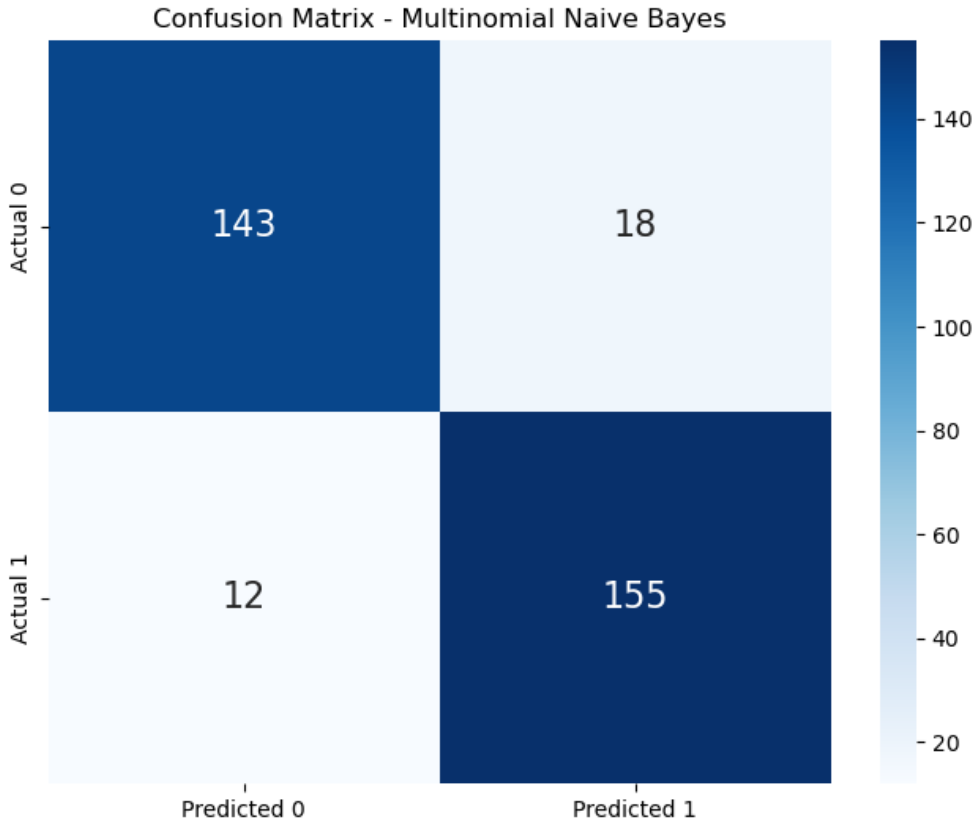
Classification Model

Multinomial Naive Bayes, Logistic Regression, KNN, model evaluation

Classification Models



Multinomial Naive Bayes model



Classification Report

- Best Cross-Validation Score: 91.91%
- Training Accuracy: 93.58%
- Testing Accuracy: 90.85%
- r/datascience (Class 0): Precision of 92% and Recall of 89%
- r/jobs (Class 1): Precision of 90% and Recall of 93% mean
- True Positives (155)
- True Negative (143)
- False Positives (18)
- False Negatives (12)

Conclusion

Multinomial Naive Bayes

robust and effective in distinguishing between the two subreddit

Ideal Candidates

- Technical proficient
- Familiarity with Academic Research and good with Time Management
- Adaptable with a desire for Growth
- Good with Collaboration and Team Skill
- Work experience and company knowledge

THANKS

Mr.employee@hrcompany.com
+91 620 421 838
HRcompany.com