Analyzing Student Sentiments and Challenges in Learning Data Science: A Reddit-Based Study

By Sagnik Chand

M.Sc, Data Analytics and Computational Social Science

Introduction

In recent years, data science has emerged as a highly sought-after field, offering students promising career prospects and presenting unique challenges. As more students delve into this growing discipline, they encounter excitement and frustration due to its steep learning curve and vast scope. This project aims to explore the sentiments and perceptions of students learning data science, specifically through their discussions on Reddit. By analyzing posts from popular subreddits such as r/datascience, r/datasciencestudents, r/learndatascience, and r/learnmachinelearning, the study seeks to uncover recurring themes, challenges, and emotions expressed by students.

The research focuses on understanding the common hurdles and frustrations, alongside the positive aspects and motivations students experience while navigating the complexities of data science. Employing text analysis techniques such as sentiment analysis, topic modeling, and word frequency analysis, the study will identify both the enthusiasm for the subject matter and the frustrations stemming from its challenging nature. The ultimate goal is to provide a deeper understanding of the student learning journey in data science, offering valuable insights that could inform educational strategies and course development, making data science education more accessible, effective, and supportive for future learners.

Research Question

How do students' sentiments and challenges in learning data science, expressed on Reddit, differ across subreddits like r/datascience, r/datasciencestudents, r/learndatascience, and r/learnmachinelearning?

Hypotheses:

- H1-1: Significant correlation between the themes and emotional sentiments of students on Reddit.
- H1-2: Frustration levels correlate with challenges like the steep learning curve and complexity of data science.
- H1-3: Positive sentiments are linked to excitement about career opportunities in data science.
- Themes: Key topics in student posts (e.g., Python, machine learning, statistical techniques).
- Sentiments: Emotional tone of posts (positive, negative, neutral) derived through sentiment analysis.
- Context: Control for sentence structure and wording to accurately interpret sentiment and themes.

Contribution:

This research fills a gap by analyzing student discussions about data science on Reddit using text analysis techniques, such as sentiment analysis and topic modeling. Unlike previous studies that focus on structured academic surveys or interviews, this approach provides insights into real-time student sentiments and challenges expressed in an online forum. By understanding these discussions' emotional and thematic underpinnings, the study aims to inform strategies for improving data science education and support learners navigating its complexities.

Dataset

This section explains the origin of the data, the collection process, and key preprocessing steps undertaken to prepare it for analysis.

The data was collected from Reddit, specifically from four relevant subreddits:

- r/datascience: General discussions about data science.
- r/learndatascience: Focused on resources and tips for learning data science.
- r/learnmachinelearning: Discussions related to learning machine learning concepts.
- r/datasciencestudents: A community dedicated to students pursuing data science.

• Reddit's API was used to scrape data, including posts, titles, comments, and relevant metadata (e.g., post scores and timestamps).

• The scraped dataset combines both original posts and user interactions (comments), allowing for a deeper understanding of the discussions and sentiments within these subreddits.

Text data underwent cleaning and transformation to ensure it was suitable for analysis:

- Stop Words Removal: Common words (e.g., "the", "is", "and") that add little analytical value were removed.
- Tokenization: Text was broken down into individual words or phrases The cleaned dataset was then combined into a unified corpus, enabling further analysis like word frequency analysis, sentiment analysis, and topic modeling.

This structured and preprocessed dataset forms the foundation for the subsequent analyses, ensuring that the text data is clean, meaningful, and ready for exploration.

Methodology

This panel outlines the methods and tools used to analyze the collected Reddit data.

1. Wordcloud Generation:
Word clouds were generated to visually represent the most frequent terms and keywords within Reddit posts and comments. This provided an initial, intuitive understanding of prominent themes and discussions across the subreddits.

2. Sentiment Analysis:
A dictionary-based sentiment analysis approach was applied to classify the text data into positive, neutral, or negative sentiments. This helped quantify the emotional tone of student discussions and identify key areas of enthusiasm or frustration.

3. Topic Modeling:
Latent Dirichlet Allocation (LDA): Used to identify hidden topics and themes by analyzing word distributions across posts.
Structural Topic Modeling (STM): Extended analysis by incorporating metadata (e.g., post scores, subreddit categories) as covariates to understand how external factors influence topic distributions and student discussions.

4. Tools Used:
R: Quantitative text analysis libraries like quanteda and data manipulation tools in tidyverse.

Additional text analysis libraries and visualization tools were employed to preprocess, analyze, and interpret the data effectively.

Results

Sentiment Analysis Findings

- Sentiment analysis revealed a highly positive outlook toward data science, with dominant emotions like joy, trust, and anticipation, reflecting optimism about career opportunities and shared community support.
- Negative emotions such as frustration and self-doubt appeared less frequently, mainly associated with challenges like the steep learning curve and technical difficulties.
- Subreddits like r/learndatascience showed the highest positivity, indicating a supportive environment for beginners.

Custom dictionary analysis emphasized:

- Progress and Learning – Focus on skill-building and growth.
- Career and Job Prospects – Discussions on career advancement and opportunities.
- Learning Resources and Tools – Practical tools and mentorship.
- Community and Collaboration – Reliance on shared knowledge.
- Imposter Syndrome and Self-Doubt – Less prominent but reflective of challenges in the field.

LDA Topic Modeling Findings

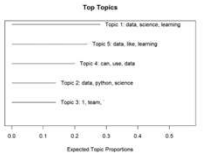
Key Themes Across Subreddits:

- Guidance for Beginners: Tips, learning pathways, and resources for newcomers (data, learning, guidance).
- Programming Tools and Collaboration: Focus on Python, R, and tools like Discord for team discussions (python, r, discord).
- Technical Problem-Solving: Troubleshooting errors, debugging, and teamwork during projects (team, error, Jupyter).
- Application of Technical Concepts: Discussions on model building, training, evaluation, and advanced learning (model, train, evaluation).
- Learning Platforms and Online Courses: Platforms like DataQuest, Microsoft, and course-based resources (Dataquest, course, Microsoft).
- Career Aspirations: Professional growth, preparation for roles like data scientist, and competitive pathways (job, company, salary, compensation).

CTM and STM Topic Modeling Findings

Common Topics Identified:

- Statistical and Foundational Methods: Discussions on PCA, p-values, and other fundamental concepts (PCA, p-value, eigenvalues).
- Career Discussions: Salary, compensation, and opportunities for professional advancement (job, company, salary, compensation).
- Programming Tools and Libraries: Dominant focus on Python, R, SQL, and efficient data manipulation libraries (python, r, SQL, pandas).
- AI and Technical Problem-Solving: Topics covering tools like GPT, coding, and AI integration (GPT, coding, SQL).
- Forecasting and Predictive Analytics: Focus on time series modeling and tools like SARIMA (time, forecasting, seasonality).



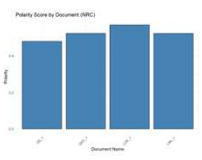
Model Comparison

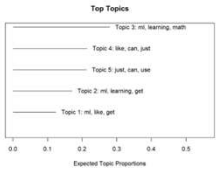
LDA vs STM:

Both LDA and STM uncovered similar topics, but STM provided more nuanced insights, particularly when accounting for covariates (like post metadata), showing differences in topic distribution across various Reddit threads. STM highlighted more subtle patterns tied to specific challenges students face, such as difficulty with complex algorithms and data preprocessing.

Sentiment Analysis vs Wordcloud:

While word clouds provided a broad view of the most frequent terms, sentiment analysis gave more context about student emotions. Combining both gave a clearer understanding of the enthusiasm and occasional frustration within the community.





Conclusion

Key Observations

- Positive Sentiment Dominates: Strong joy, trust, and anticipation highlight optimism and excitement about learning and career growth.
- Focus on Career Development: Career-oriented discussions around roles, salaries, and pathways are central themes across all platforms.
- Technical and Collaborative Engagement: Programming tools, AI techniques, and collaborative learning platforms like Python, SQL, and Discord dominate technical discussions.
- Challenges Exist but Don't Overpower Optimism: While frustration and self-doubt arise from technical complexity, these challenges are outweighed by overall positivity and mutual support within the community.

Summary: The study successfully identified the major themes and sentiments surrounding students' experiences in learning data science. As expected, while students expressed excitement and curiosity about data science and machine learning, they also faced significant challenges, particularly with the complexity of the subject matter and the learning curve.

Implications: These findings highlight areas where educational interventions could be focused, such as providing more resources on complex topics (like machine learning algorithms) and creating more structured learning paths to ease the transition for beginners.

Future Work: Further research could explore how these sentiments change over time or conduct deeper qualitative analysis to understand individual student struggles more comprehensively.