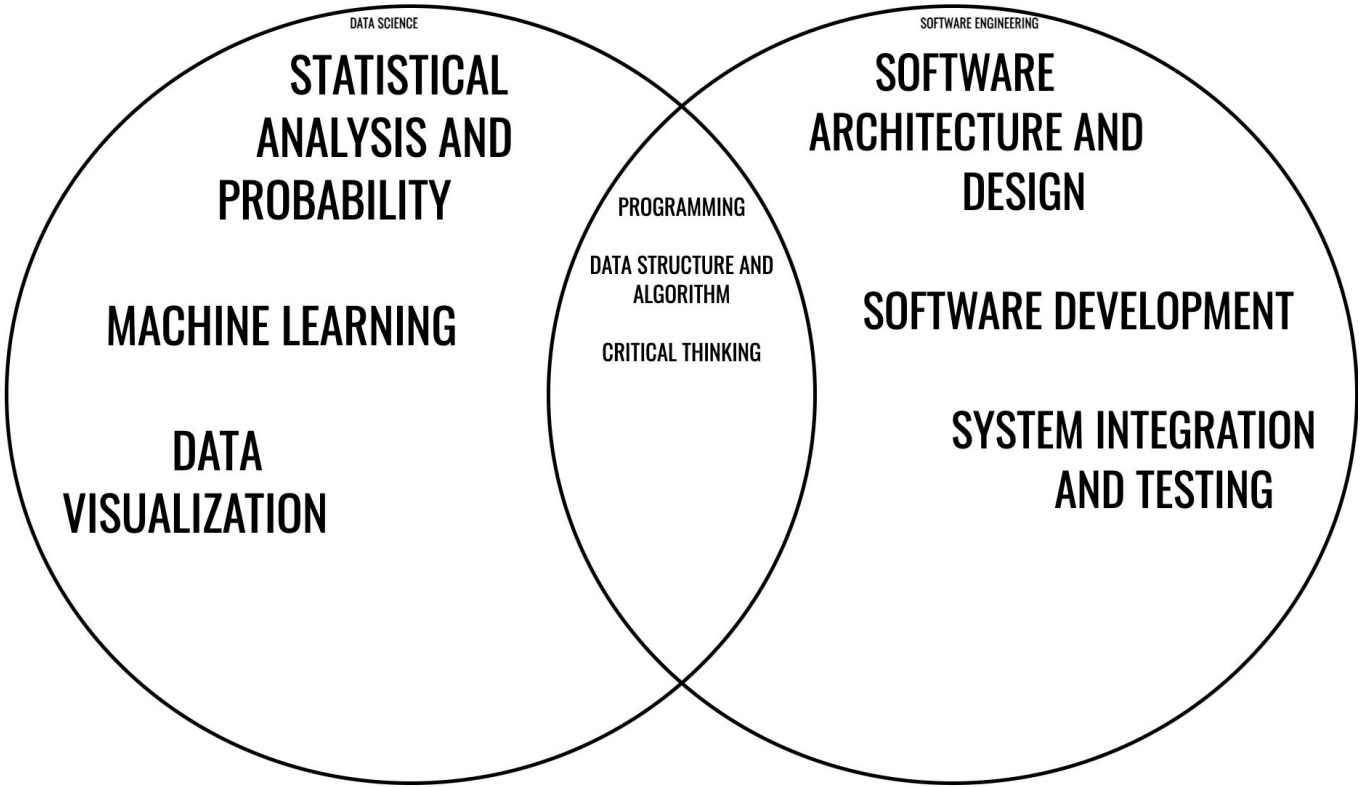
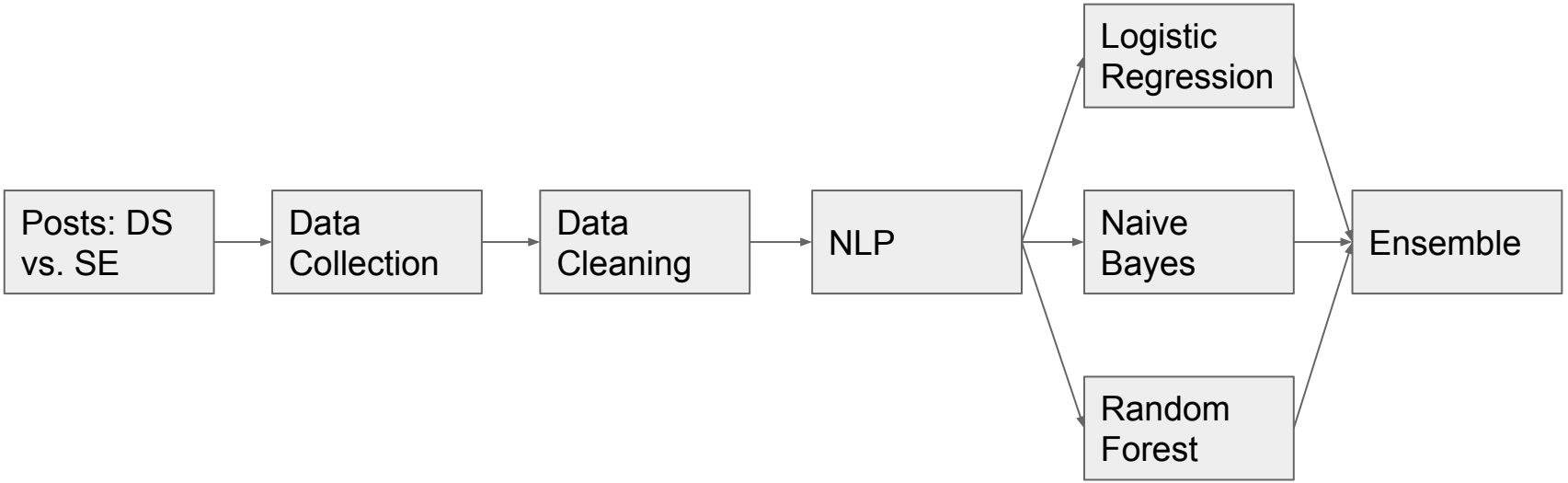
Data Science vs. Software engineering: Analyzing subreddit content with NLP and classification models

Ran Ma

Background

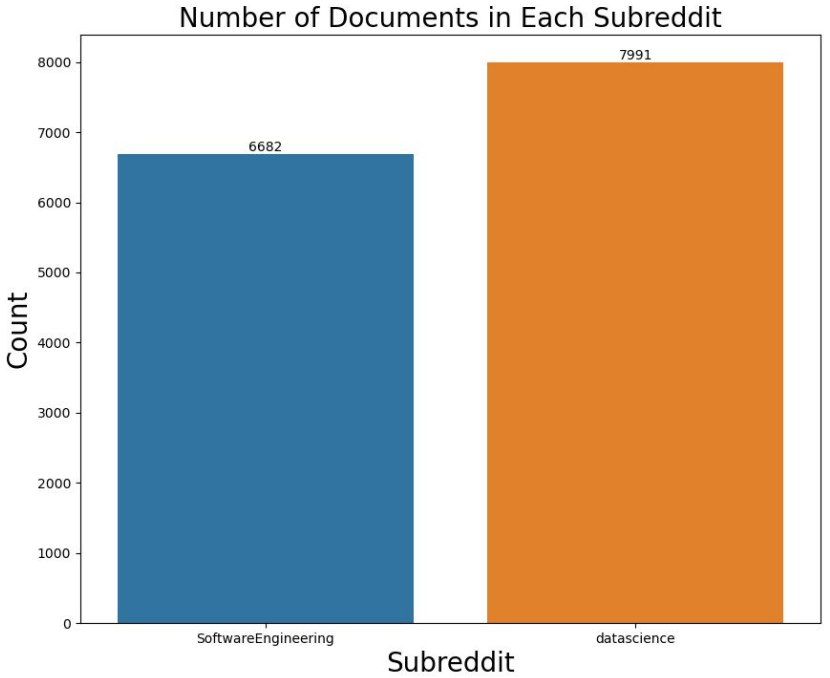

DATA SCIENCE
STATISTICAL ANALYSIS AND PROBABILITY
MACHINE LEARNING
DATA VISUALIZATION
PROGRAMMING
DATA STRUCTURE AND ALGORITHM
CRITICAL THINKING
SOFTWARE ENGINEERING
SOFTWARE ARCHITECTURE AND DESIGN
SOFTWARE DEVELOPMENT
SYSTEM INTEGRATION AND TESTING

Goal and workflow



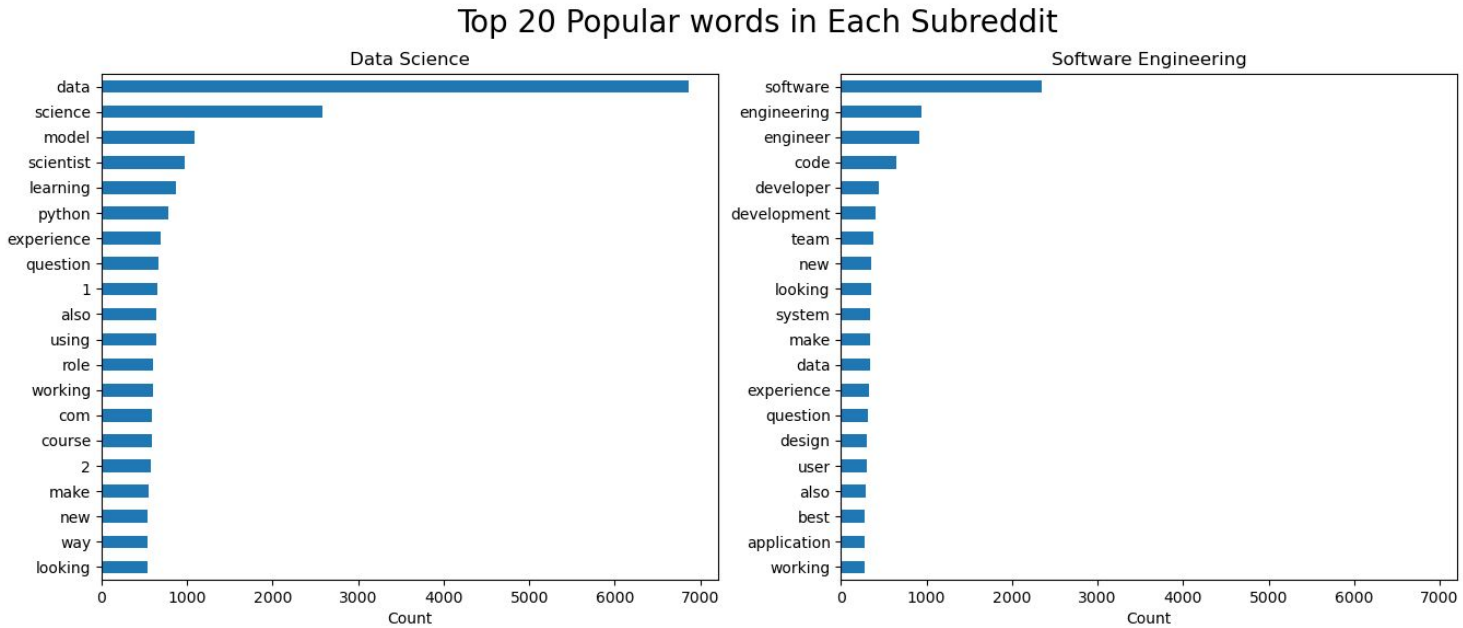
Data collection and cleaning

- Resources: **Pushshift API**
- Subreddit: Data Science vs. Software Engineering
- Documents: title or selftext
- Remove Missing values
- Remove Duplicate values
- Remove “[removed]” and “[deleted]”



NLP

- Vectorization: Count Vectorize/Tf-idf Vectorize
- Applying NLTK: Stop words Lemmatize
- Retains: Letters and Digits

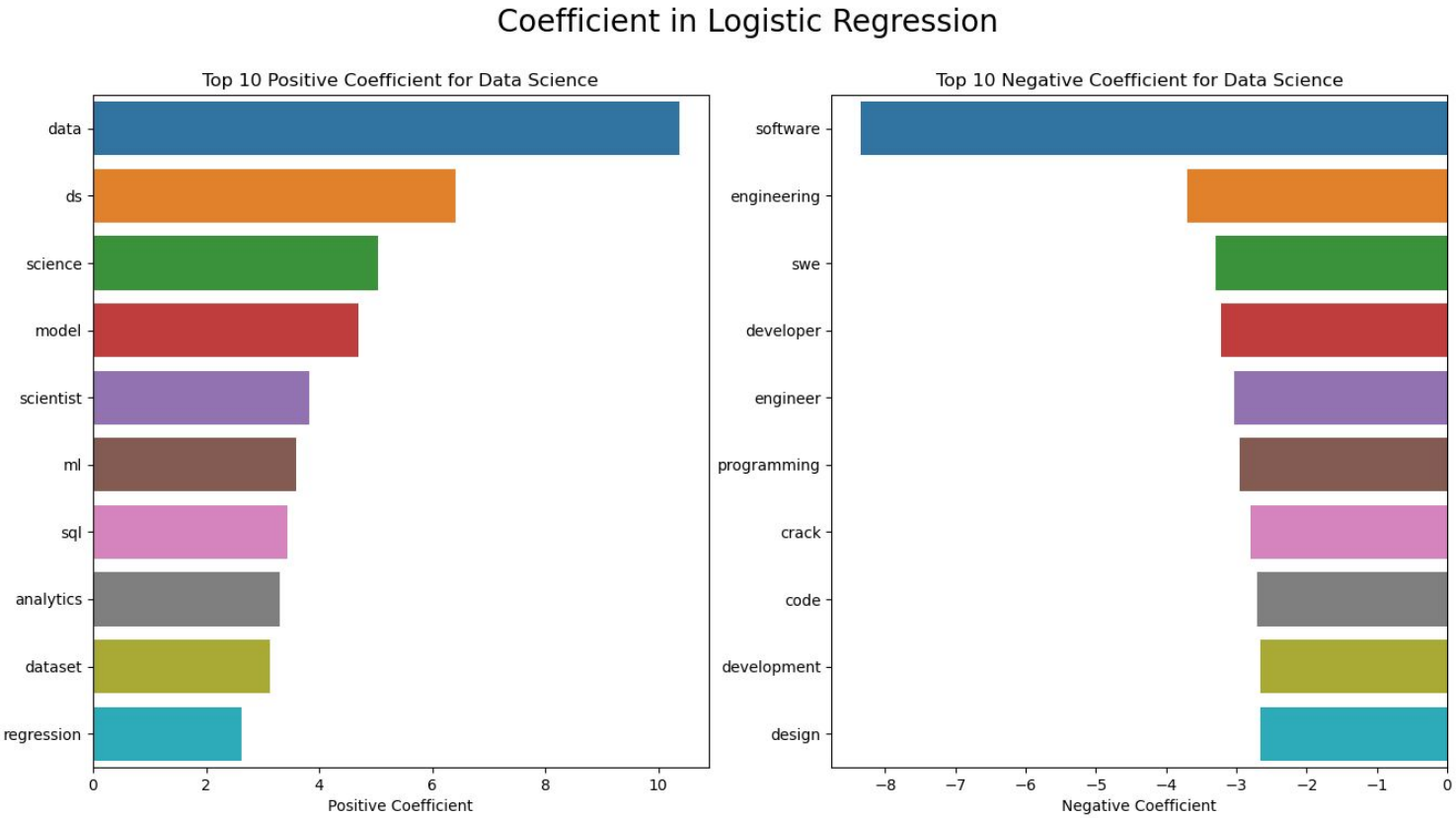


Model Building

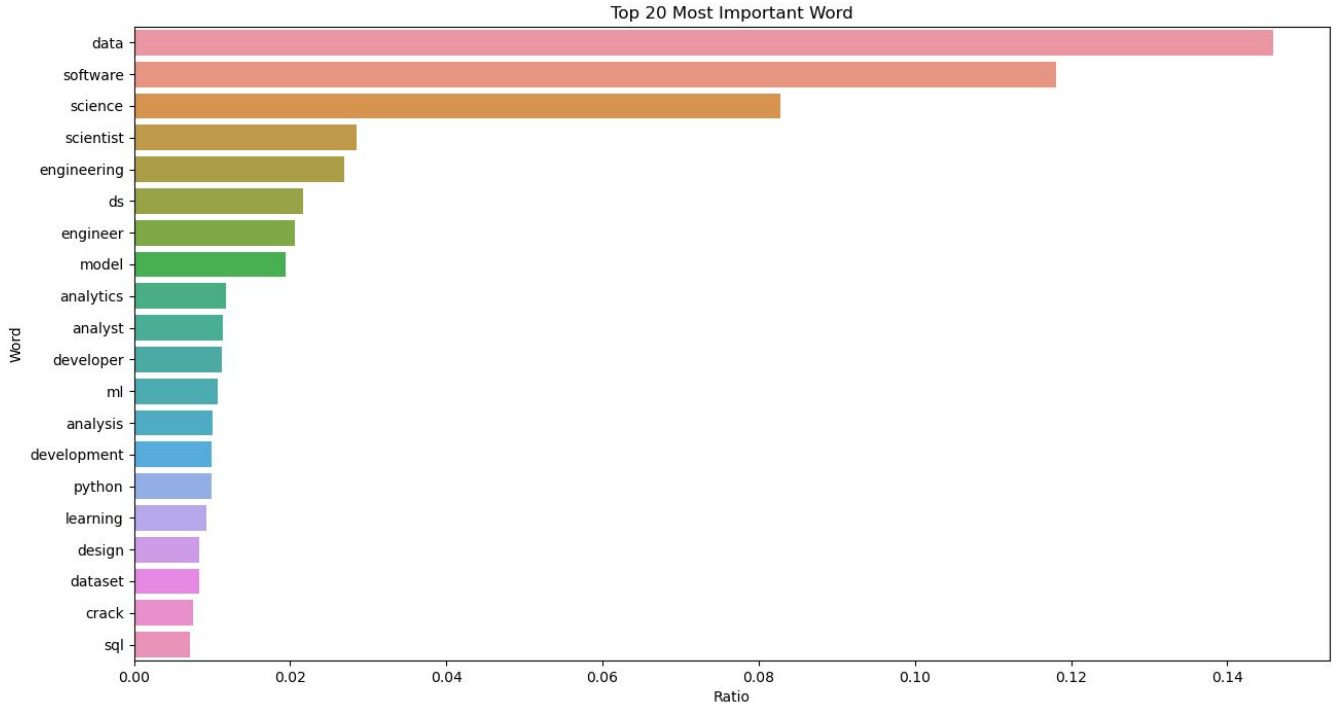
- Logistic Regression
- Multiple Naive Bayes
- Random Forest
- Weighted Averaging
- Stacking

	accuracy
Logistic Regression	0.843
Multiple Naive Bayes	0.834
Random Forest	0.822
Weighted Averaging	0.846
Stacking	0.847

Logistic Regression: Coefficient



RANDOM FOREST: IMPORTANT WORD



CONCLUSION

1. Data Science: Model, Machine Learning, SQL, Analytics
2. Software Engineering: Software, Code, Development, Design
3. Best Accuracy: Ensemble Method

Thank you