May 2024

University of Texas at Arlington

PROFESSOR

Dr. Angela Liegey-Dougall

STUDENT

Zehra Erden

Psych Students' Data Dive

Table of Contents

Survey Question

"When you think of data science, what does it mean to you (please be honest and write what first comes to mind)?" was asked to psychology students.

Text answers are saved in csv format.

> 255 rows (students)

> 7 features

Sentiment scores manually coded in the following format:

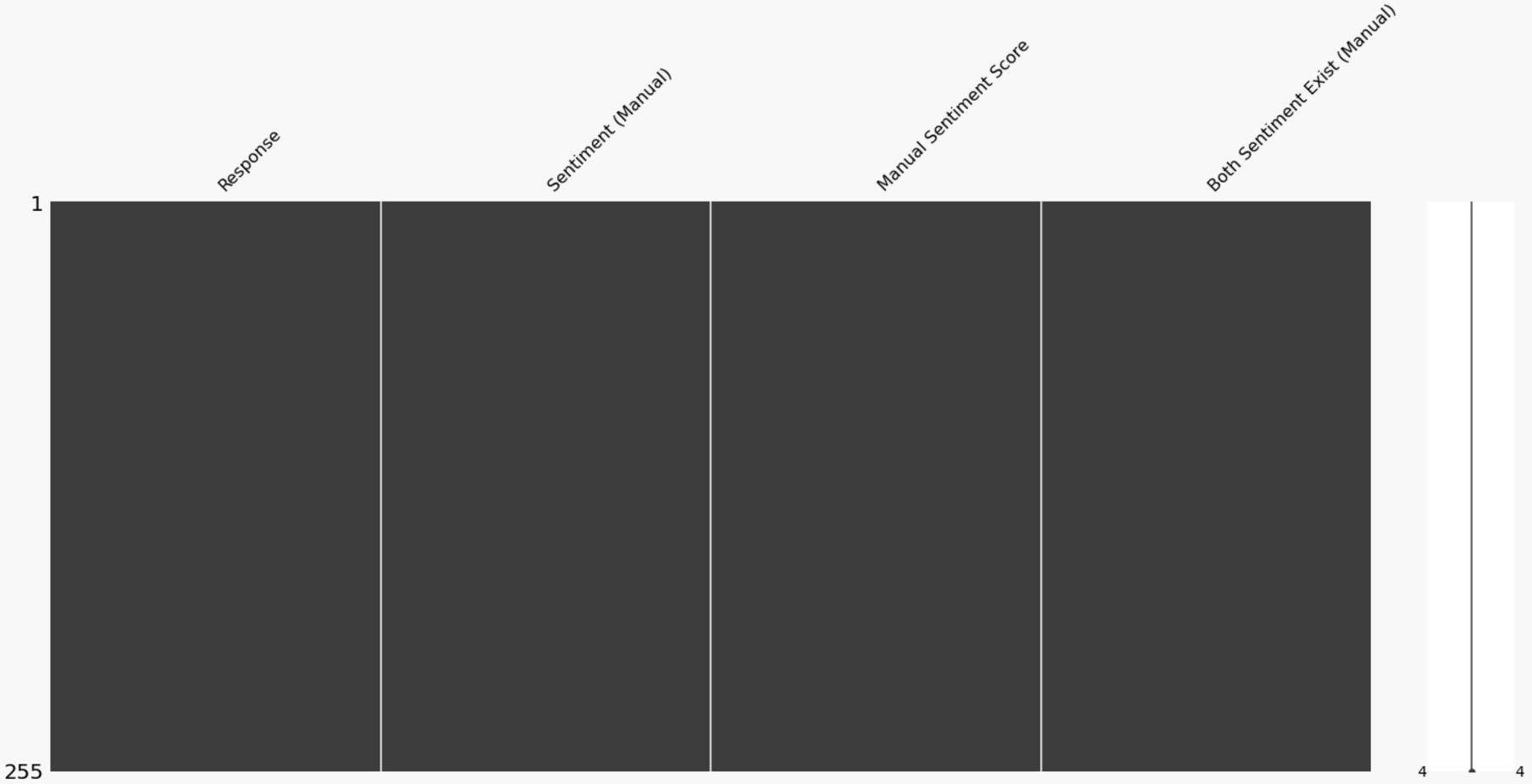
> 'Sentiment (Manual)': Very Positive, Positive, Neutral, Negative, Very Negative

> 'Manual Sentiment Score': Scale of 1 to 5

> 'Both Sentiment Exist': Binary score to represent text contains both sentiments.

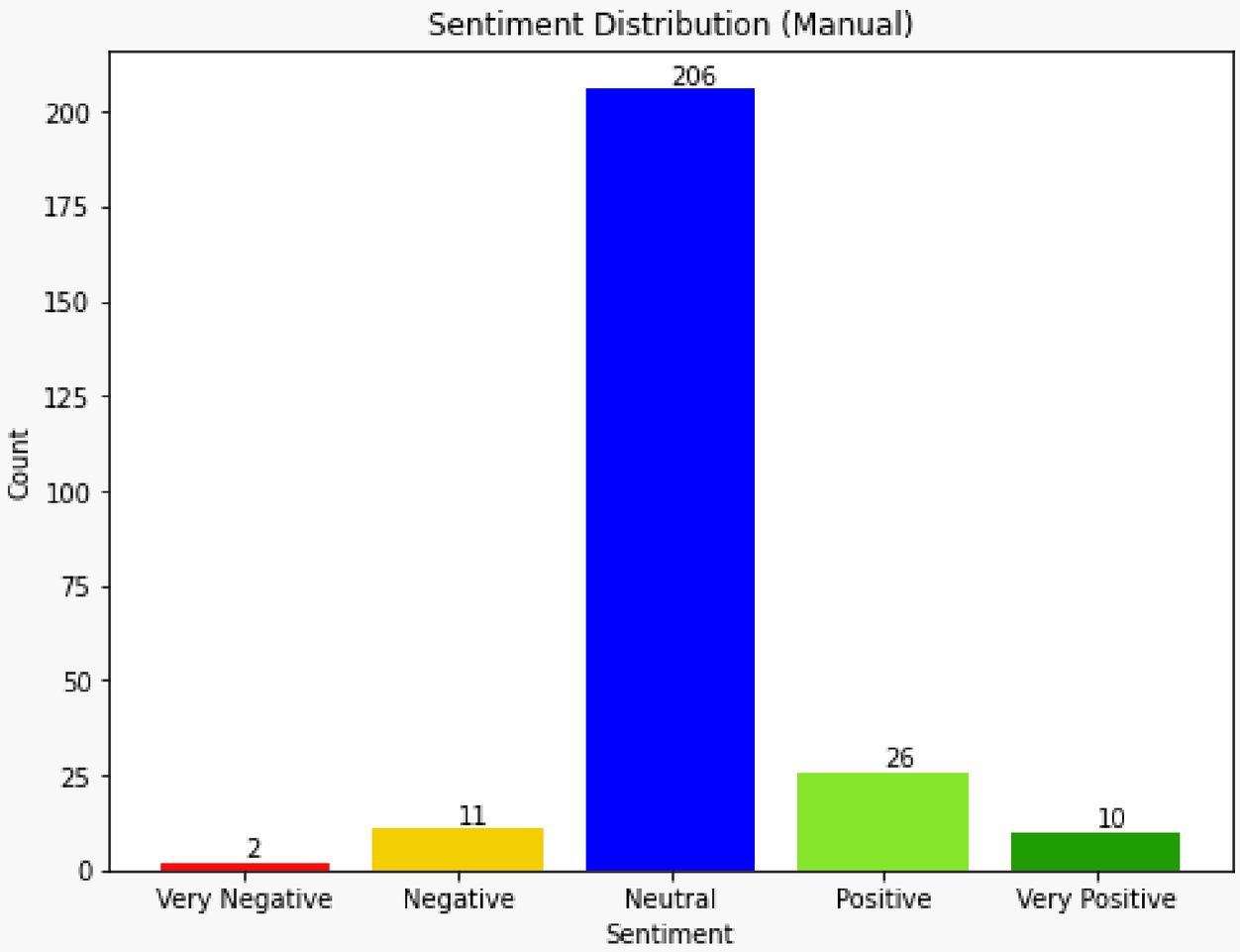
Missing Values

Dataset does not have any missing (null) value.



Manually Coded Sentiment Distribution

Manual sentiment codes display that the majority of the responses are natural, while there are more positive responses than negative ones.

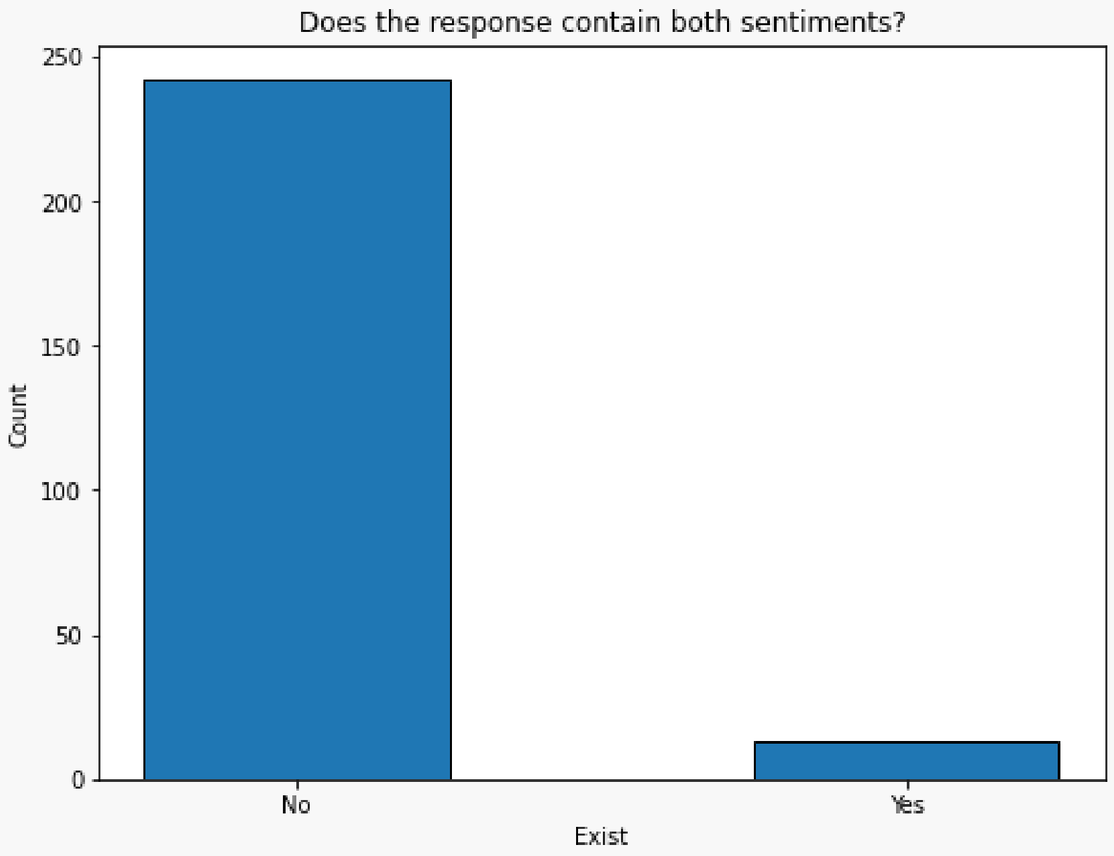


Manually Coded Sentiment Distribution

Mean Sentiment Score (Manual): 3.12

Most Frequent Sentiment Category (Manual):

Neutral



Text Preprocessing

Techniques used:

> Converting the input text to lowercase

> Creating a set of English stop words

> Tokenization

> Removing non-alphanumeric characters from each token

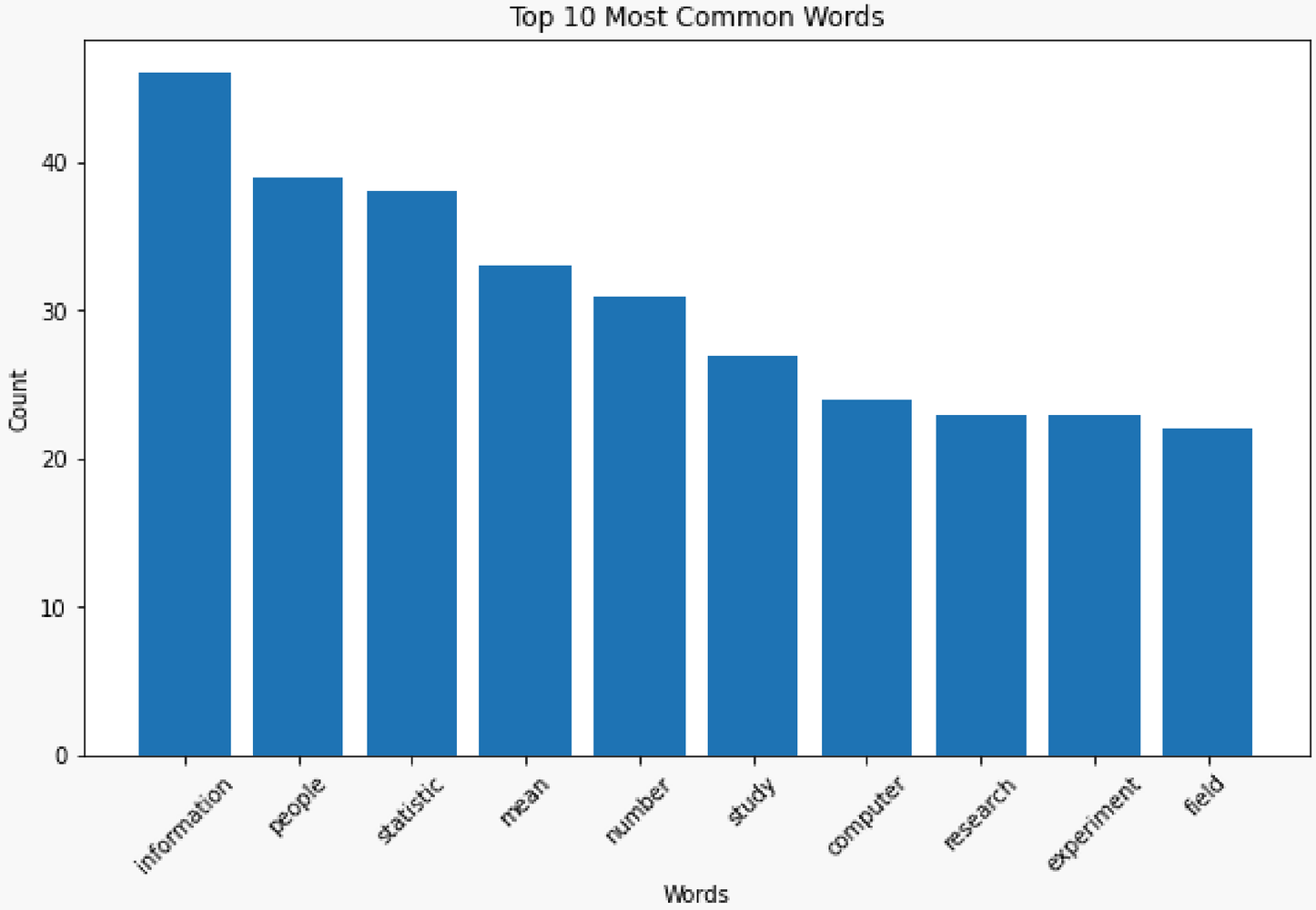
> Lemmatization

> Additionally included stop words based on frequency on text:

- 'include', 'involve', 'data', 'science', 'think',
- 'know', 'may', 'way', 'like', 'something', 'using', 'lot', 'really'

Word Frequency Analysis

10 Most Frequent Words



Word Frequency Analysis

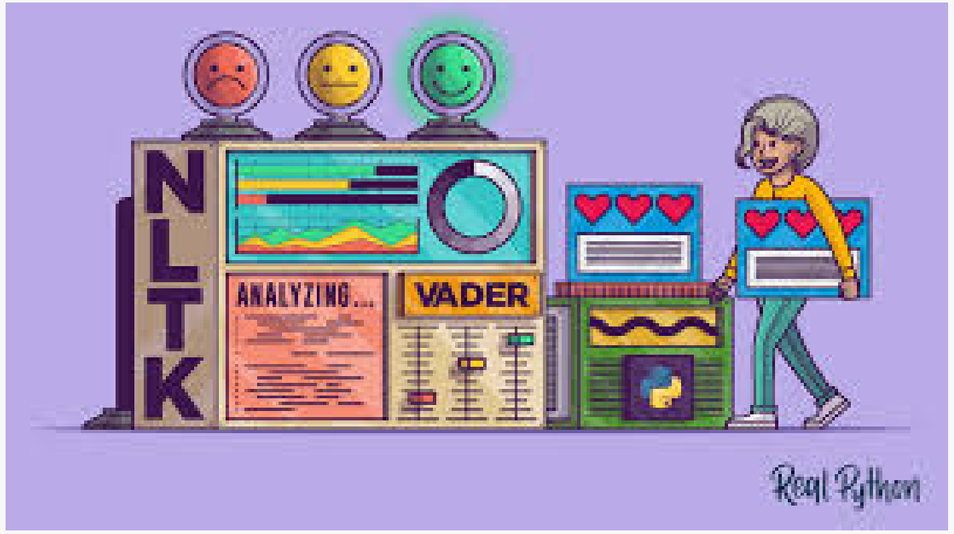
Word Cloud of Responses

What is VADER?

> VADER (Valence Aware Dictionary and sEntiment Reasoner)

> A lexicon and rule-based sentiment analysis tool

> Assigns a sentiment score to a piece of text based on the words it contains, taking into account both the polarity (positive, negative, or neutral) and the intensity of the sentiment



Advantages and Disadtavgesof VADER

Open Source	Binary Classification
Fast Processing	Dependency on Lexicon
Domain Adaptability	Limited Context Understanding
Handles Negations and Intensifiers	Difficulty with Sarcasm and Irony

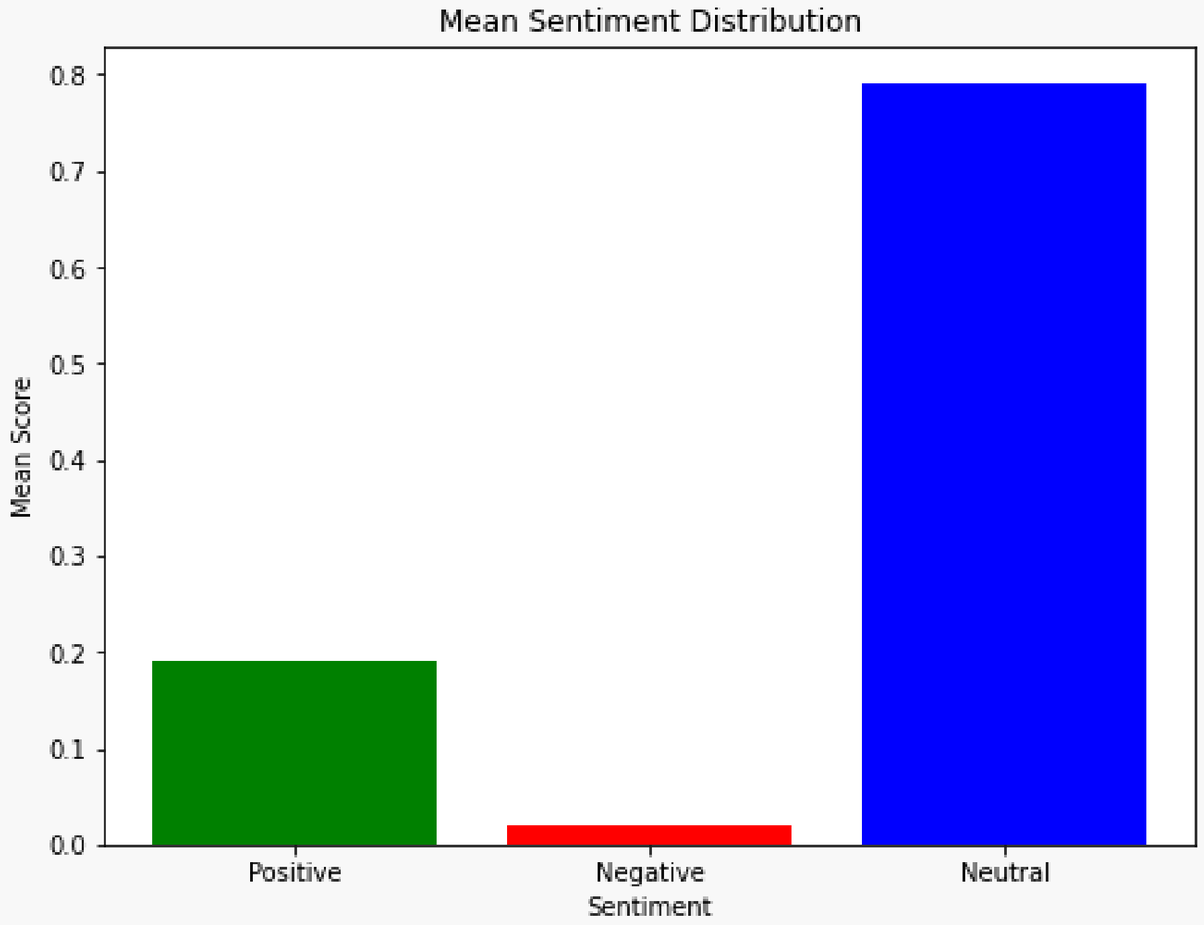
Tasks Done

Using the VADER SentimentIntensityAnalyzer, sentiment scores are calculated and assigned to each response/text.

We calculated the mean sentiment scores for each sentiment category.

	Positive	Negative	Neutral	Compound
0	0.237	0.198	0.565	0.1779
1	0.412	0.000	0.588	0.3612
2	0.324	0.000	0.676	0.5859
3	0.000	0.000	1.000	0.0000
4	0.000	0.000	1.000	0.0000
...	...	...	...	...
250	0.000	0.000	1.000	0.0000
251	0.333	0.000	0.667	0.4588
252	0.394	0.000	0.606	0.0772
253	1.000	0.000	0.000	0.3182
254	0.592	0.000	0.408	0.7003

Overall Mean Sentiment Distribution of Responses



Lexicon-based sentiment analysis using VADER yielded similar results to our overall manually coded sentiments.

Most of the responses are neutral, with positive responses outnumbering negative ones.

Why Naive Bayes Classifier?

- Naive Bayes classifiers perform well with text data, which is often sparse and high-dimensional.
- It performs reasonably well even with small training datasets.
- The probabilistic nature of Naive Bayes classifiers allows for easy interpretation of results.

Likelihood

Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)\,P(c)}{P(x)}$$

Posterior Probability

Predictor Prior Probability

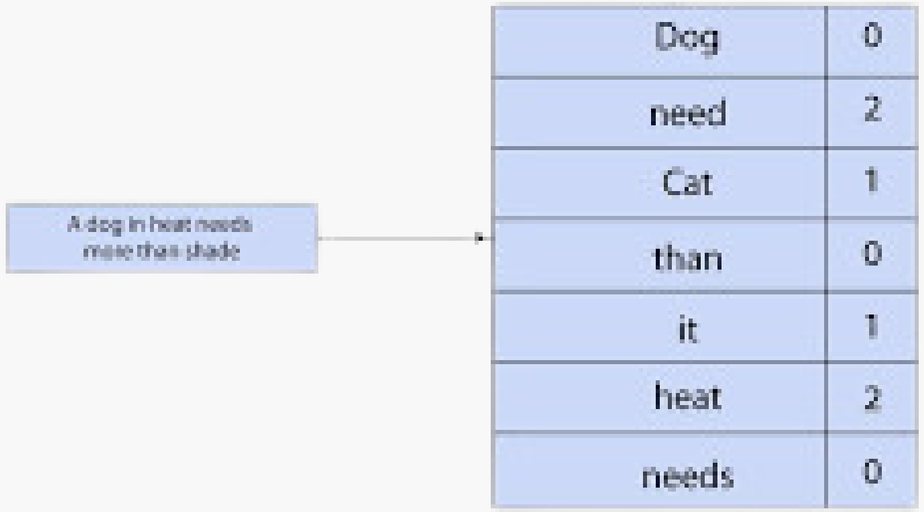
$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Vectorization

> Vectorization transforms textual data into numerical vectors that algorithms can process.

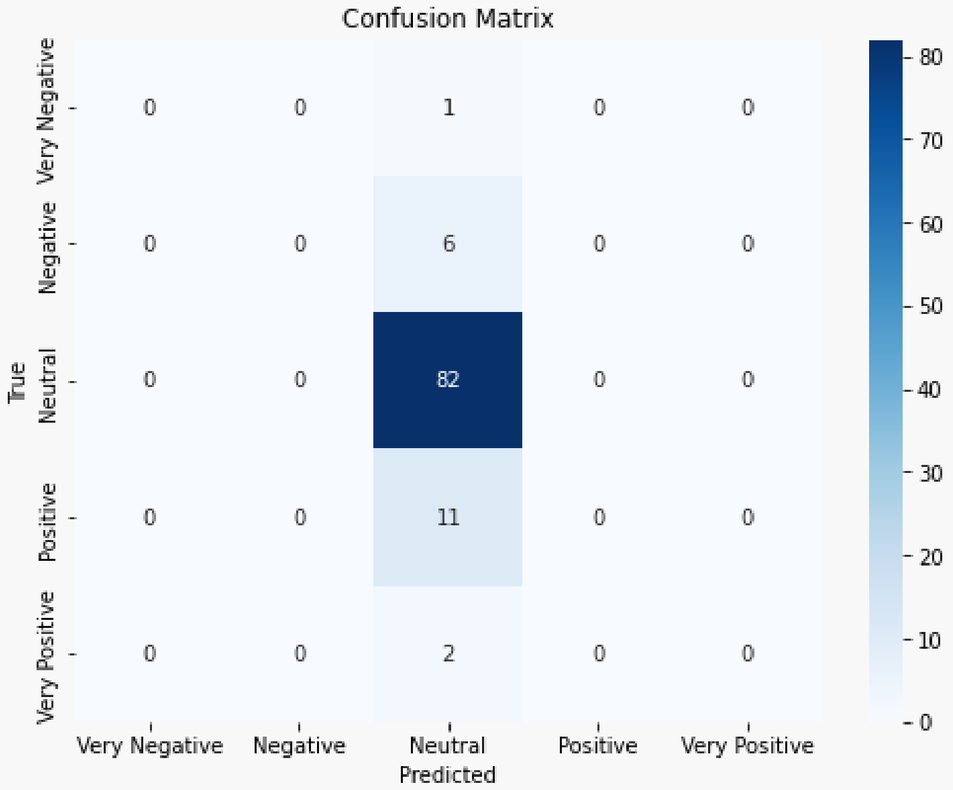
Bag of Words (BoW)

>BoW represents text data by counting the frequency of each word in a document. Each document is then represented as a vector, where each element corresponds to the count of a particular word in the document.



Model Evaluation

- The model is trained on manually coded sentiment scores and text data.
- Accuracy: 0.80



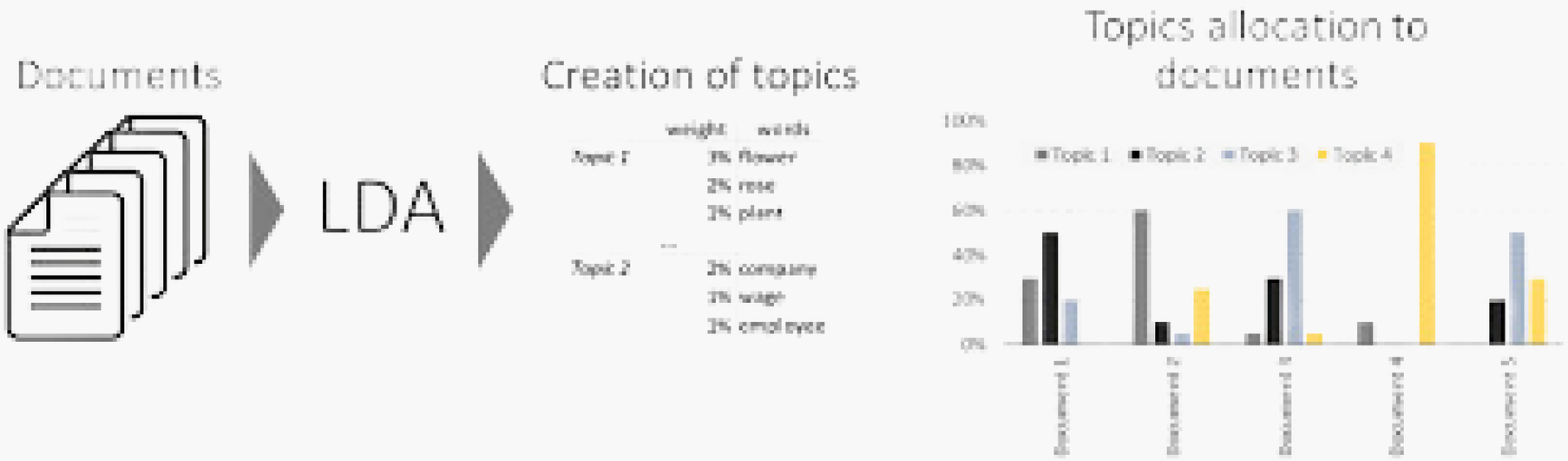
Model Evaluation

	precision	recall	f1-score	support
Very Negative	0.00	0.00	0.00	1
Negative	0.00	0.00	0.00	6
Neutral	0.80	1.00	0.89	82
Positive	0.00	0.00	0.00	11
Very Positive	0.00	0.00	0.00	2
accuracy			0.80	102
macro avg	0.16	0.20	0.18	102
weighted avg	0.65	0.80	0.72	102

- Apparently the model did not perform well despite the good accuracy score.
- All metric scores are very high for 'Neutral'.
- Looking at the confusion matrix too, if the model predicted all values only 'Neutral' it would get 80% accuracy.

Latent Dirichlet Allocation (LDA)

> Latent Dirichlet Allocation (LDA) is a probabilistic topic modeling technique used for discovering latent topics within a collection of documents.



Why Latent Dirichlet Allocation (LDA) ?

> Assumes that each document in a corpus is a mixture of various topics

> It does not require labeled data for training

> Reduces the dimensionality of the text data by representing documents as distributions over topics and words

Topic 1 - Computer Science and Math

Topic 1 - Word Cloud

› Psychology students seem to see data science as a field combining computer programming, math and statistics.

Topic Modeling

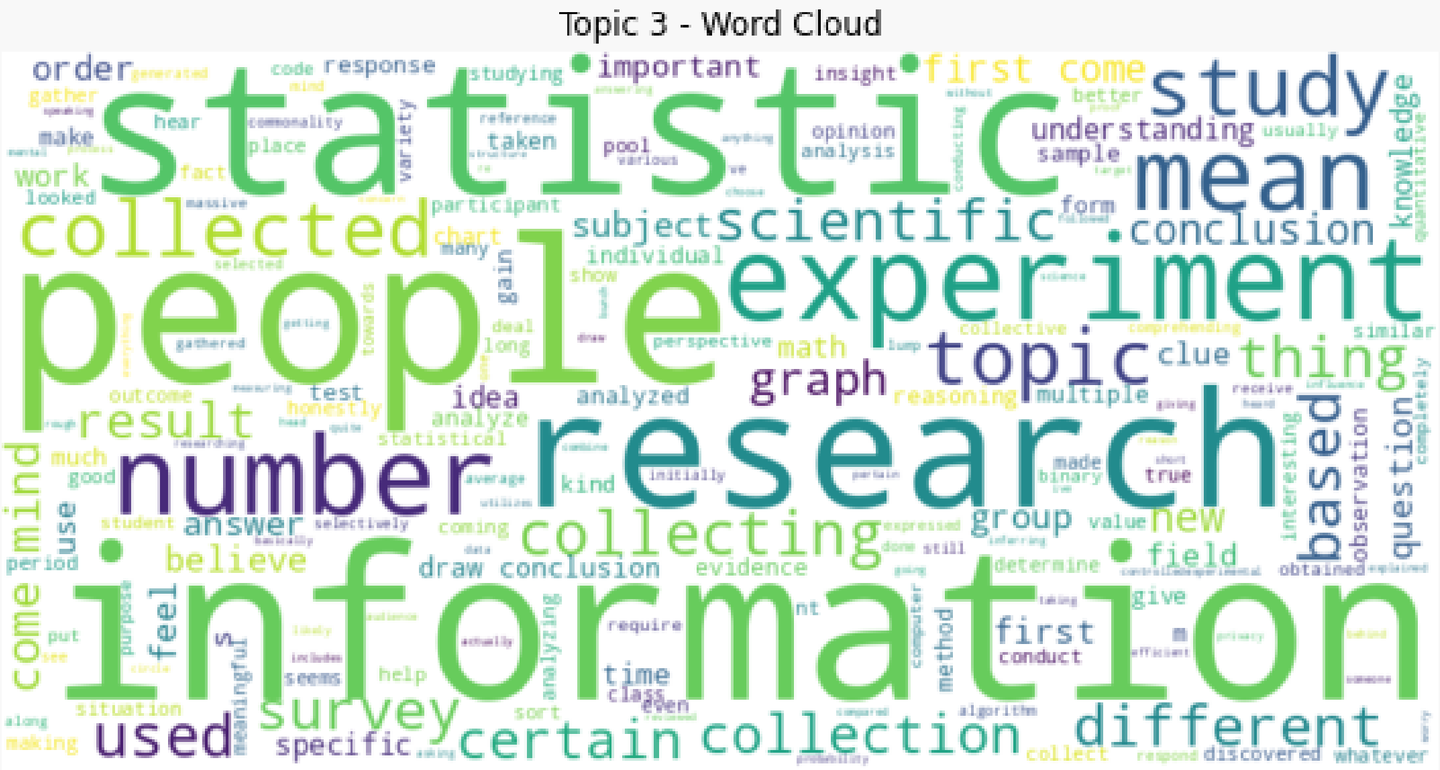
Topic 2 - Data Science as a Tool

Topic 2 - Word Cloud

- Psychology students do think that data science is a good tool to make various analysis to find trend, pattern, and meaning insight.

Topic Modeling

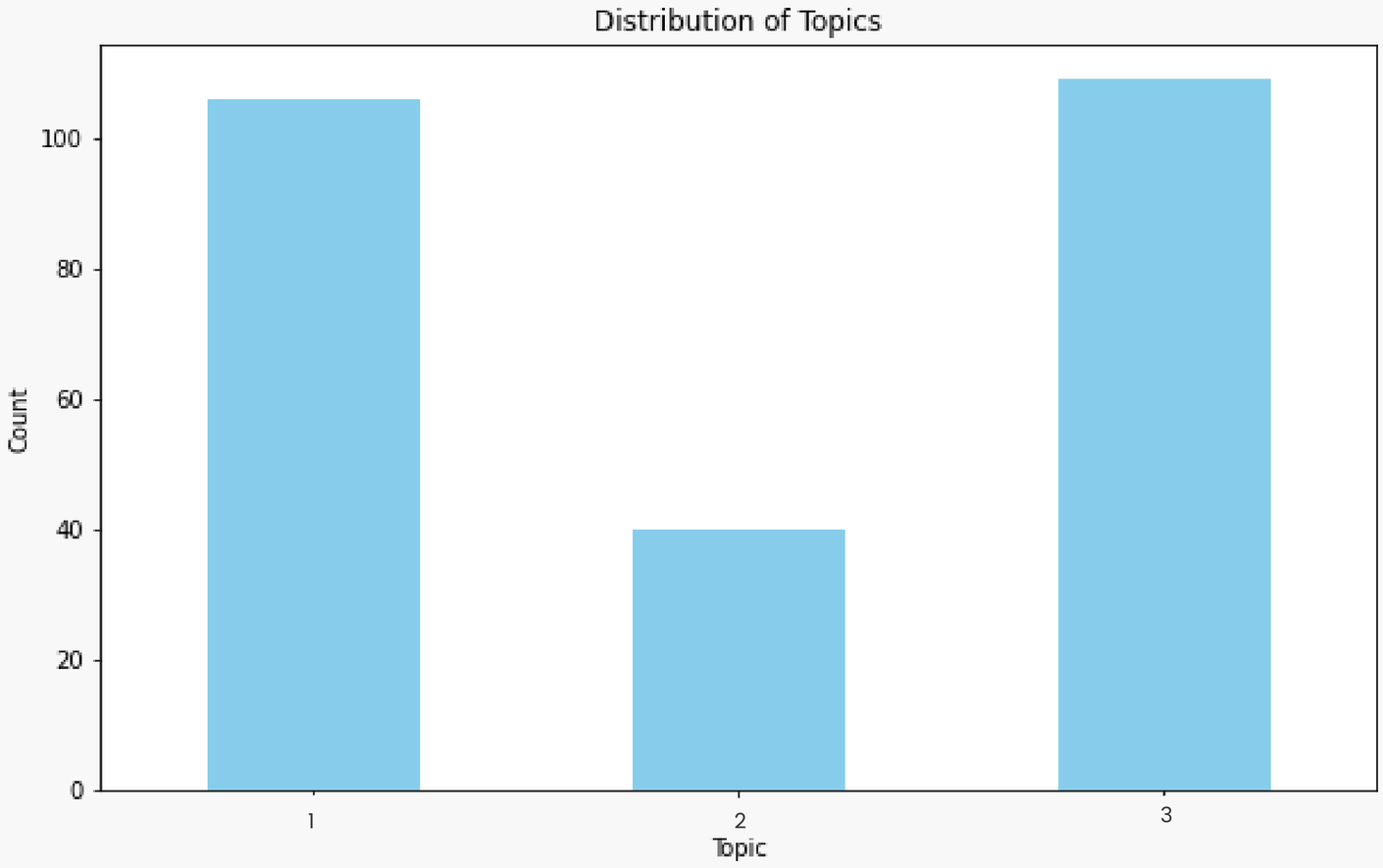
Topic 3 - Data Science in Psychological Research


Topic 3 - Word Cloud

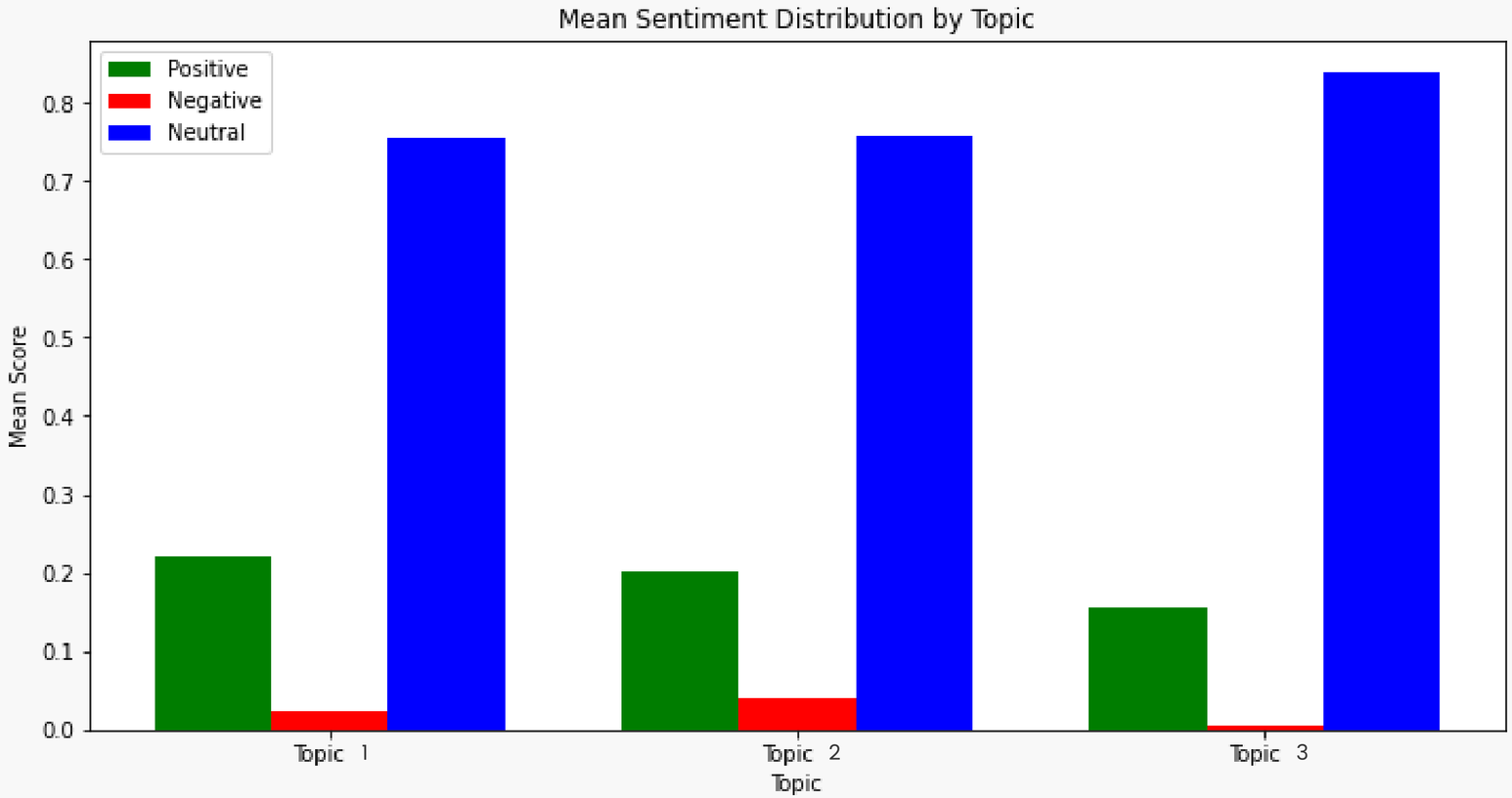
- Psychology students match data science with researches and experiments in their field.

Distribution of Topics

Computer science, mathematics, and psychological research have been mentioned most frequently. Students appear to grasp the concept of data science and its impact on their respective fields, but they often perceive it merely as a tool.



Sentiment Distribution by Topic



Conclusion

Overall students have positive approach through data science but are more negative against using data science as a tool.

This suggests a potential area of concern or misconception among students regarding the role and significance of data science as a practical tool within their academic disciplines.