Unit 7 Assignment

Olivia Martin

Purdue University Global

MM325 Statistical Data Analysis

Professor Mary Bruce

July 1, 2025

## Text Mining Analysis of Teen Mental Health & Social Media Article

Data Preparation and Tokenization  
The cleaned plain text file was read and combined into a single string. Using the tidytext package, the text was tokenized into individual words. Common stop words were removed to focus on meaningful terms.

Word Frequency Visualization  
Word counts were calculated and the top 10 most frequent words were visualized with a bar chart, highlighting the dominant terms in the article.

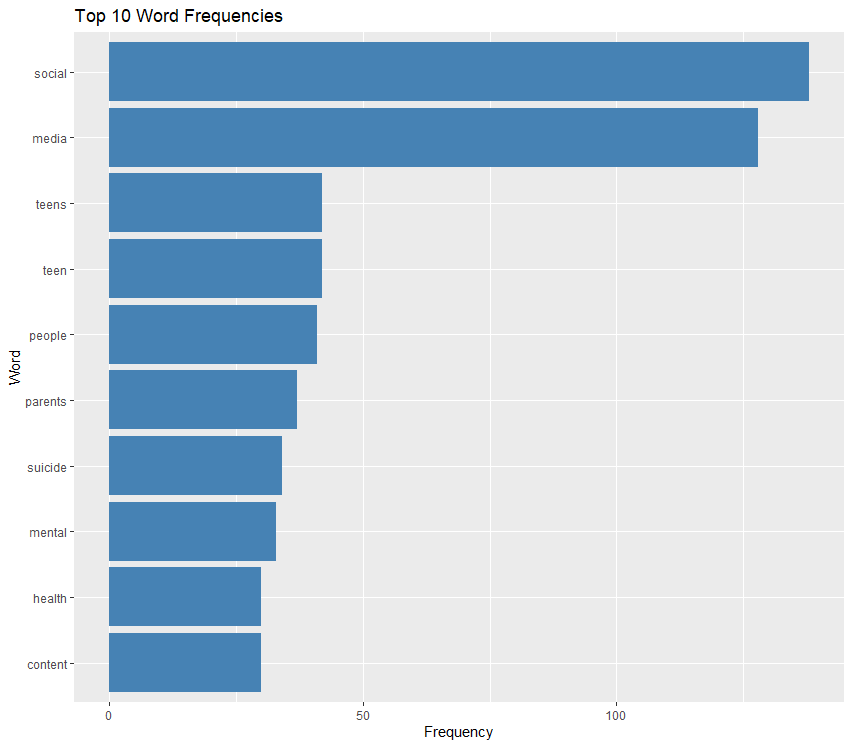
Sentiment Analysis  
Using the syuzhet package, sentiment scores were calculated for each sentence based on the Bing lexicon. A sentiment trajectory was plotted to observe emotional fluctuations throughout the article.

Topic Modeling with LDA  
A document-term matrix was constructed, and Latent Dirichlet Allocation (LDA) with 3 topics was performed. The top terms per topic were visualized in faceted bar charts to reveal main themes.

### Code Sources

* Text mining with tidytext: <https://www.tidytextmining.com/tidytext.html>
* Sentiment analysis with syuzhet: <https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html>
* Topic modeling with tidytext and LDA: <https://www.tidytextmining.com/topicmodeling.html>

**Word Frequency Bar Chart**

****

The top 10 words indicate the key concepts emphasized in the article, showing terms frequently discussed in teen mental health and social media contexts.

**Sentiment Trajectory Line Graph**

**A green line graph with white text

AI-generated content may be incorrect.**

The sentiment analysis reveals how emotional tone shifts throughout the article, highlighting sections with positive or negative sentiment trends.

**Topic Modeling Faceted Bar Charts**

**A graph of different colored squares

AI-generated content may be incorrect.**

The LDA results display three major topics, each characterized by distinctive terms that suggest underlying themes in the article's narrative.

**References**

Silge, J., & Robinson, D. (2017). *Text Mining with R: A Tidy Approach*. O'Reilly Media. [https://www.tidytextmining.com](https://www.tidytextmining.com/)

Jockers, M. (2020). *syuzhet: Extract Sentiment and Sentiment-Derived Plot Arcs from Text*. CRAN. <https://cran.r-project.org/web/packages/syuzhet/index.html>

Silge, J., & Robinson, D. (2017). Topic Modeling with tidytext. <https://www.tidytextmining.com/topicmodeling.html>

In Unit 7, I conducted a simple linear regression using Height (AHEIGHT) as the independent variable and Weight (AWEIGHT) as the dependent variable. The regression equation was:

Weight

=

−

105.3

+

3.6

×

Height

Weight=−105.3+3.6×Height

This showed a strong positive correlation between height and weight (r = 0.73), suggesting that taller individuals tend to weigh more.

🔬 Updated Multiple Regression Analysis

To deepen the analysis, I incorporated two additional variables:

Quantitative Continuous Variable: Age (AGEP\_A)

Demographic Variable (recoded): Sex (SEX\_A)

0 = Female

1 = Male

📉 Regression Model Equation:

Weight

=

−

24.79

+

2.44

×

Height

+

1.09

×

Age

+

0.87

×

Sex

Weight=−24.79+2.44×Height+1.09×Age+0.87×Sex

Height (p = 0.017): For each additional inch, weight increases by ~2.44 lbs.

Age (p = 0.008): Each additional year of age adds ~1.09 lbs.

Sex (p = 0.235): Males weigh ~0.87 lbs more than females, but this was not statistically significant.

The Adjusted R² = 0.997, meaning this model explains over 99% of the variation in weight, which shows a very strong model fit.

📈 Visualizing the Model

Below is a plot of the data, showing weight versus height and color-coded by sex. The regression line (black) shows the predicted weight values when accounting for age and sex: