INSY 669- Text Analytics

Winter 2025

Prof. Taha Havakhor

Student Well-Being: Analyzing McGill Subreddits

February 18, 2025

Prepared By:

Ayda Elzohbi (260899679)
Berly Biju (261208805)
Helia Mahmood Zadeh (261224416)
Rajiha Mehdi (261203537)
Maisy Mofakhami (261178077)

1 Introduction

In today's digital era, students increasingly turn to online platforms to share their struggles, but many remain unsupported. With 1 in 5 postsecondary students facing mental health challenges and 11% reporting thoughts of suicide (Wiens et al., 2020), the need for intervention is urgent. Social media and discussion forums provide a window into these struggles, so how can we leverage these digital spaces to provide effective mental health support?

Studies have demonstrated that Reddit is an effective source for mental health research due to its large, diverse, and active user base (Gkotsis et al., 2017). By tapping into McGill's subreddit data, university administrators and mental health professionals can identify areas of concern, and behavior patterns and be prepared for stressful times in the academic calendar to promote a healthier educational environment.

1.1 Objective

The goal of this analysis is to detect both the presence of mental health challenges and the frequency at which they occur by examining social media posts for relevant patterns or signals.

2 Methodology

The analysis pipeline integrated two text vectorization techniques—Word2Vec and TF-IDF—with four machine learning models: Logistic Regression, Support Vector Machines (SVM), Random Forest, and AdaBoost utilizing cross-validation and hyperparameter tuning.
These models were trained on two labeled datasets (both sourced from Reddit posts) focusing on depression and stress detection. The trained models were then used to classify posts from the McGill subreddit. Additionally, we incorporated sentiment analysis using VADER to establish a baseline, allowing us to compare different model predictions across multiple labels.

2.1 Data Collection

The data collection process was split into training data and testing data:

Training Data:

- Depression Dataset from Kaggle (7,732 rows).
- Stress Detection Dataset from Social Media Articles (5,557 rows).

Testing Data:

- McGill Subreddit posts were scraped using the BeautifulSoup library (2,800 rows).

2.2 Text Preprocessing

We prepared the data using a text preprocessing pipeline that included tokenization, stopword removal, and lemmatization. Noise such as non-alphabetic characters, URLs, symbols, and emojis was removed, yielding 2,592 clean, unique subreddit posts for analysis.

2.3 Tools and Techniques

The analysis applied NLP techniques and models to extract insights. Word embedding methods like Word2Vec and TF-IDF represented textual data, while models such as Logistic Regression, SVM, Random Forest, and AdaBoost classified and analyzed text. Feature engineering refined text representations by limiting vocabulary to the top 3,000 words, excluding those in fewer than two documents, and ignoring terms appearing in over 95% of the dataset. Cross-validation using StratifiedKFold with five folds ensured robust evaluation.

3 Results and Findings

3.1 Classification Model Performance

- SVM with TF-IDF classifier: 96.27% test accuracy for depression detection
- Random Forest with Word2Vec classifier: 82.99% test accuracy for stress detection

3.2 Applying Classifier on McGill Data Results

- Word2Vec had a higher correlation with posts marked as negative posts, while TDIF marked 22% of positive posts as stress-related and 29% as depression related
- Upon removal of positive posts, depression incidence was 12%, and stress 14% (Word2Vec)
- High Incidence Period: Apr-May (32%) and Sept-Oct (34%), correlating with exams

3.3 Topic Modeling Analysis

The LDA analysis identified distinct themes in depression- and stress-related posts. Depression-related discussions covered academic burdens (coursework, exams, finals), emotional distress (feelings of struggle, motivation issues), and social concerns (seeking help, friendships). The coherence score of 0.3527 suggests an overlap between academic and emotional stress. Stress-related discussions were more focused, primarily on academic pressure (courses, finals, grades, performance expectations). The higher coherence score of 0.5124 indicates a more defined and singular topic, reinforcing that stress discussions are centered around academic workload, whereas depression discussions are broader and more varied.

4 Conclusion

This analysis highlights how academic stress, support-seeking, and community connection appear in social media discussions, revealing key mental health patterns among students. It demonstrates the potential of social media analysis for early detection of mental health concerns, helping universities refine support services and interventions during high-stress periods. Future work could expand the study to other universities, improve model accuracy, and explore depression posts through psychological theories, reinforcing the value of social media data in mental health research.