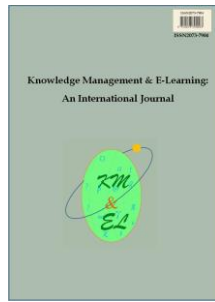

From data to insights: Machine learning in thematic analysis of complex health conditions on social media

**Anila Virani
Ahsan Mollani
Piper Jackson**

Thompson Rivers University, Canada



Knowledge Management & E-Learning: An International Journal (KM&EL)
ISSN 2073-7904

Recommended citation:

Virani, A., Mollani, A., & Jackson, P. (2025). From data to insights: Machine learning in thematic analysis of complex health conditions on social media. *Knowledge Management & E-Learning*, 17(4), 666–682.
<https://doi.org/10.34105/j.kmel.2025.17.031>

From data to insights: Machine learning in thematic analysis of complex health conditions on social media

Anila Virani* 

School of Nursing
Thompson Rivers University, Canada
E-mail: avirani@tru.ca

Ahsan Mollani 

Faculty of Science
Thompson Rivers University, Canada
E-mail: mollania22@mytru.ca

Piper Jackson 

Faculty of Science
Thompson Rivers University, Canada
E-mail: pjackson@tru.ca

*Corresponding author

Abstract: Social media data has the potential to enable the exploration of public perspectives on health conditions, interventions and policies. However, the resource-intensive nature of qualitative analysis creates a barrier to the timely utilization of social media data. Artificial intelligence can provide innovative ways to reduce the burden by augmenting the data analysis process for researchers. Therefore, the purpose of this research is to explore the feasibility of using Machine Learning (ML) in augmenting thematic analysis of complex health issues data available on social media. First, we performed a human-determined deductive thematic analysis of the 7,177 comments posted by YouTube video viewers on postpartum depression. Then we used the same data to perform machine-assisted analysis using five Natural Language Processing (NLP) classification models with and without class balancing techniques (class weight balance and SMOTE) to balance the unequal number of comments across themes. Our analysis suggested that supervised machine learning may not be optimal for analyzing complex health datasets. However, integrating ML-NLP techniques with thematic analysis can effectively filter out ineligible data, thereby enhancing the efficiency of traditional thematic analysis processes. This integration allows researchers to focus on valuable data, producing meaningful insights and comprehensive analysis while saving time otherwise spent sifting through a large amount of ineligible data. Social media data can offer significant public health insights, and to enhance the use of such data, future research should focus on developing artificial intelligence tools to improve the efficiency of thematic analysis using larger datasets and unsupervised machine learning models for analyzing complex health issues.

Keywords: Machine learning; Natural language processing; Thematic analysis; Social media data; Complex health issues

Biographical notes: Dr. Anila Virani serves as an Assistant Professor at the School of Nursing, Thompson Rivers University, Canada. With over 25 years of extensive experience in healthcare, academia, and research, Dr. Virani's scholarly journey encompasses significant national and international contributions. Her doctoral research at the University of Calgary and subsequent postdoctoral project at the University of British Columbia focused on the integration of informatics solutions within healthcare systems. Dr. Virani's research program is dedicated to investigating the intricate dynamics between individuals and technology in healthcare and educational settings, with a particular emphasis on technological innovation.

Ahsan Mollani holds a Master's degree in Data Science from Thompson Rivers University. His research interests include technology-enhanced learning and the application of machine learning and natural language processing (NLP) techniques to analyze and interpret complex datasets.

Dr. Piper Jackson is an Associate Professor of Computing Science at Thompson Rivers University in Kamloops, Canada. He completed his doctorate at Simon Fraser University and later became the director of the Modelling of Complex Social Systems Program at SFU's IRMACS Centre. Grounded in cross-disciplinary collaboration and community engagement, Dr. Jackson's research has spanned several fields, including criminology, political science, gerontology, and public health. Starting with computational simulations developed from domain expert knowledge, his work has shifted to applications of artificial intelligence and machine learning to health and social issues.

1. Introduction

Every day, a large amount of content is posted on social media platforms in the form of audio, visual and text that are available to researchers to study various phenomena occurring in people's daily lives. Social media content can provide valuable insight into several health-related topics. It can help researchers identify issues or gain deeper insight related to the phenomenon of interest or public responses to interventions such as awareness campaigns and policy changes (Bacsu et al., 2024; Han et al., 2018; Hemphill et al., 2022). Social media research is generally conducted on freely available public data. Since it does not require direct interaction with humans, researchers can quickly access the data without the Human Research Ethics Board's approval. Further, this data is free from response bias, the Bradley effect and the observer's paradox/Hawthorne effect that comes with planned research projects where participants are aware that they are being researched (Tatman, 2018).

Many researchers have studied the use of social media platforms such as Facebook, TikTok, Twitter and YouTube for providing and accessing social support in health-related conditions (Gage-Bouchard et al., 2017; Haslam et al., 2017; Johnson et al., 2020). Comments posted by viewers of the original post can provide a substantial understanding of the topic and the types of support people seek and receive. Thematic analysis of such posts is a widely used qualitative method employed by social media researchers. Thematic analysis is a process by which researchers identify common themes that run across the data. It can take many forms, such as inductive, where themes emerge from the data or deductive, where themes are derived from an existing theory, model or literature (Braun & Clarke,

2006). Thematic analysis is a time-consuming and labour-intensive process (Castleberry & Nolen, 2018). Therefore, researchers look for methods that could provide similar results but are less resource-intensive than thematic analysis. For example, Taylor et al. (2018) compared thematic analysis with rapid analysis and found that rapid analysis took 43 hours, whereas thematic analysis required 117 hours. However, they also reported that for the interpretation and the report writing, rapid analysis took six times longer than thematic analysis. Social media data is based on public opinion and not specifically collected for a specific research topic from the participants in a research setting. Therefore, it may contain a lot of irrelevant and ineligible data, such as off-topic comments, language slang, emojis, incomplete comments or comments that are difficult to comprehend when read without the context or original post (Azizi, 2022). These unique data characteristics pose an added burden on social media researchers trying to find relevant data while sieving through the ineligible data.

Artificial intelligence (AI) has great potential to support social media researchers in their thematic analysis. AI tools such as Machine Learning (ML) and Natural Language Processing (NLP) can support researchers in lessening the burden of thematic analysis of social media data. ML is a part of AI, which provides algorithms to support humans in completing time-consuming and labour-intensive tasks efficiently and accurately. ML is broadly divided into two types: supervised and unsupervised. Supervised ML approaches require training data to complete the task. It is commonly used to classify data and make predictions. Unsupervised ML can process raw data and is used for understanding the relationships and patterns within the unlabeled dataset. NLP is a form of ML that processes human language by using computational modelling to recognize, understand and interpret linguistic data. It can be used with either supervised or unsupervised ML approaches (Khurana et al., 2023).

The importance of social media data in exploring health conditions, combined with the labor-intensive nature of thematic analysis, has prompted many researchers to seek methods to alleviate the burden of thematic analysis. Various researchers have adopted ML-NLP techniques to accelerate qualitative data analysis. For instance, Towler et al. (2023) demonstrated that human-led analysis required 148 hours, whereas machine-assisted analysis took only 40 hours. Similarly, Li et al. (2021) reported that the NLP approach was significantly less time-consuming, taking 15 hours compared to 150 hours for human coding. However, many ML studies focus on predicting the risk of a condition, determining the severity of issues, and evaluating responses to interventions (Iyortsuun et al., 2023; Li et al., 2021; Sajno et al., 2023; Towler et al., 2023). A review on the use of ML-NLP in mental health conditions suggested that NLP studies are primarily used for symptom identification, treatment comparisons, and questioning the current classification of diseases to support existing ideas rather than generating new information (Le Glaz et al., 2021). Similarly, research on Postpartum Depression (PPD), a mental health condition, in relation to social media and ML focuses on developing predictive models to identify PPD risk and severity (Gopalakrishnan et al., 2023; Safa et al., 2022). There is a notable lack of NLP research utilizing social media data to analyze the types of support individuals with PPD receive from these platforms. Therefore, this research study aims to explore the feasibility of augmenting thematic analysis of complex health issues, such as PPD, using supervised ML-NLP models.

2. Data description

PPD is one of the most common mental health concerns globally for women after childbirth. A systematic review of 58 studies on PPD suggested a 25.8% prevalence among women (Dadi et al., 2020). The 2018/2019 maternal mental health survey reported that 23% of Canadian mothers experienced PPD (Statistics Canada, 2019). COVID-19 pandemic-related restrictions posed additional challenges to the mental health of parents. Several studies reported an increased prevalence of PPD during the pandemic (Chen et al., 2022). A well-known factor in minimizing the negative effect of depression is access to and provision of social support (Dadi et al., 2020; Dol et al., 2021; Izadirad et al., 2017).

Social media provides a common public ground to express feelings and connect with others with similar interests and issues. The freedom to be anonymous creates an opportunity to share feelings openly and honestly. The increased use of social media to provide and receive support among parents is well documented (Archer & Kao, 2018; Auxier et al., 2020). We retrieved 18 months of YouTube data on PPD to analyze the types of support provided on social media. We extracted comments that were posted by YouTube viewers in response to the YouTube videos selected for analysis. We used Patricia Leahy-Warren's (2014) social support theory to deductively analyze the data into three major types of support: 1) appraisal, 2) emotional and 3) informational. Appraisal support refers to supporting actions or behaviours of others in the form of affirmation, feedback, and social comparison. Emotional support is defined as expressing some form of caring attributes to an individual, such as expressing empathy, concern, caring and love. Informational support refers to providing relevant information, advice, and directives that the individual can use in coping with issues. Although the theory provides information on other types of support, our manual thematic analysis data indicated the dominance of these types of support in the context of social media.

In this paper, we focus on the supervised ML approach and NLP classification models that were utilized to process the manually thematically analyzed data on PPD and social support. We compared the data for accuracy and suggested a model that incorporated the intricacies of data. We also discuss the limitations of using ML-NLP approaches in the context of thematic analysis and make suggestions for future ML-NLP applications in this area.

3. Method

Our team, consisting of content experts and research assistants, conducted a manual deductive thematic analysis using Leahy-Warren's (2014) social support theory, which outlines types of social support in relation to PPD, and Braun and Clarke's (2006) steps of thematic analysis. We refer to this as a human-determined analysis. The human-determined data revealed common types of support, such as appraisal, informational, and emotional support, that people obtained from social media platforms. We then performed ML-NLP modeling using five classifiers on the human-determined analysis to assess its effectiveness in analyzing complex health data related to PPD. Since the human-determined analysis identified unequal amounts of data under each type of support, we used class balancing techniques. This section describes the ML-NLP methodological approach in detail as illustrated in Fig. 1.

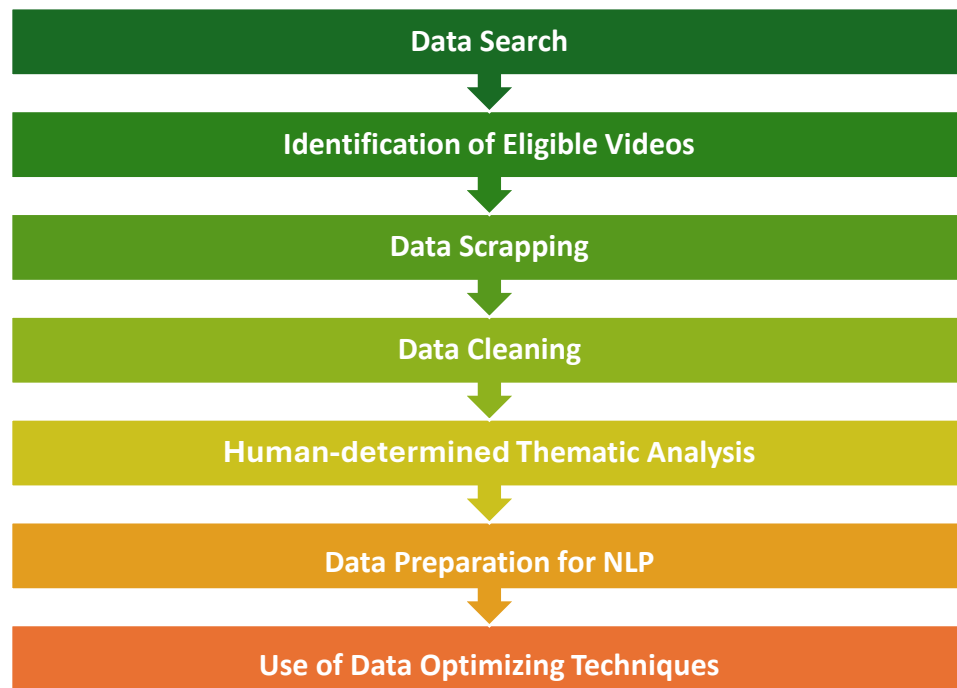


Fig. 1. Methodological workflow

3.1. Data search

We used Google search filters to identify YouTube videos related to social support in PPD. We collected data from January 2022, when pandemic-related restrictions began easing off, to June 2023, when the restrictions officially ended (WHO, 2022, 2023) to understand the types of social support obtained by parents dealing with PPD through social media during the pandemic. We used the search terms: Depression AND Postnatal OR Perinatal OR Postpartum OR PPD AND COVID OR pandemic. We obtained the top 10 videos of each month to cover the social media use for social support during the COVID-19 pandemic. In total, we retrieved 180 videos.

3.2. Identification of eligible videos

Out of 180 videos, 34 videos were deemed eligible using the following eligibility criteria: a) in English; b) has one or more comments; c) relevant to the topic; and d) focused on people's personal experiences mentioned by individuals experiencing PPD and/or their families or reported by newscasters, podcasters or YouTubers through interviews or compiling the news available to the public such as a celebrity struggle with PPD.

Videos were considered ineligible if there were no comments ($N = 52$); in a language other than English ($N = 34$); informational videos by healthcare professionals, bloggers and others ($N = 12$); and PPD-related movies, movie trailers and business advertisements ($N = 3$). Videos were considered irrelevant if they shared normal postpartum and parenting struggles ($N = 4$); neither focused on depression or COVID-19 ($N = 11$) nor focused on overall mental health affected by COVID-19 ($N = 30$); these videos

were not specific to PPD, or PPD was briefly mentioned as one of the many topics in mental health.

3.3. Data scraping

We extracted the comments data using Google Sheets. We linked the Google Sheet to App Script to activate Google Data API v3 and ran the scraping code that populated the data of that particular video in the Google Sheet. We repeated this process for each eligible YouTube video and compiled the data into an Excel sheet for cleaning and analysis. We retrieved a total of 8,476 comments from 34 eligible videos. We extracted the following data for each video: a) video name and URL; b) username; c) date and time when the comment was posted; d) comments; d) reply count; and e) likes count.

3.4. Data cleaning

After scraping, we cleaned the data using Python for the comments that were not in English, duplicates, and empty. We also removed emojis, URLs, HTML, hashtags, and special entities from the comments. We removed a total of 1,299 comments during the cleaning process. After the cleaning process, the remaining 7,177 comments from 33 videos were used for data analysis.

3.5. Human-determined thematic analysis

We did a human-determined deductive thematic analysis of the 7,177 comments using Patricia Leahy-Warren's (2014) Social Support theory. We hired a research assistant who was supervised by the team to determine the themes. The research assistant spent 134 hours coding the data. The RA received periodic supervision from the team (50 hours) to ensure that coding was performed correctly. We coded the data in three themes:

- Appraisal Support: $N = 3,375$
- Informational Support: $N = 940$
- Emotional Support: $N = 600$

Comments were also coded as ineligible if they were not relevant, incomplete or difficult to understand ($N = 2,262$).

3.6. Data preparation for NLP

To use data in NLP, the data must be prepared to allow a machine to process the data (Jakhotiya et al., 2022). This section details the techniques we used to prepare our data.

Converting comments to lowercase comments. NLP requires consistency in data for analysis, and therefore, all the comments need to be converted to lowercase. For example:

- Original: *I so relate to your testimony. I too stayed silent for many years with worry and anxiety, but praise God he finds strength in my weakness!*
- Converted for NLP analysis: *i so relate to your testimony. i too stayed silent for many years with worry and anxiety, but praise god he finds strength in my weakness!*

Tokenization is a process of breaking down the text into units called tokens. Each token is a piece of text, usually a word, that can be used for analysis. During this process,

the sentence is split up into individual words. This helps the machine process each word individually.

- Original sentence: *I so relate to your testimony. I too stayed silent for many years with worry and anxiety, but praise God he finds strength in my weakness!*
- Converted for NLP analysis: ['i', 'so', 'relate', 'to', 'your', 'testimony', 'i', 'too', 'stayed', 'silent', 'for', 'many', 'years', 'with', 'worry', 'and', 'anxiety', 'but', 'praise', 'god', 'he', 'finds', 'strength', 'in', 'my', 'weakness']

Stopword removal is a process of cleaning up the text by removing common words such as like, and, the, that do not add meaning in the context of analysis.

- Original: *I so relate to your testimony. I too stayed silent for many years with worry and anxiety, but praise God he finds strength in my weakness!*
- Converted for NLP analysis: ['so', 'relate', 'testimony', 'stayed', 'silent', 'many', 'years', 'worry', 'anxiety', 'praise', 'god', 'finds', 'strength', 'weakness']

Stemming is a process of trimming words down to their root form to simplify the text and make it easier for the machine to analyze. For example, in this sentence, the word “stayed” was reduced to “stay”

- Original: *I so relate to your testimony. I too stayed silent for many years with worry and anxiety, but praise God he finds strength in my weakness!*
- Final text converted for NLP analysis: ['so', 'relate', 'testimony', 'stay', 'silent', 'many', 'year', 'worry', 'anxiety', 'praise', 'god', 'find', 'strength', 'weakness']

Term Frequency-Inverse Document Frequency (TF-IDF) Vectorization: Textual data needs to be converted into numbers for the machine to process the data. TF-IDF uses statistical methods to transform the text into numerical form (vectors) by calculating its weight in the given text. (Singh & Shashi, 2019). The TF-IDF algorithm calculates the importance of a word in a document by comparing its relative frequency to its frequency throughout the entire text. Words frequently used in a specific dataset are considered important and given higher TF-IDF weights. On the other hand, words that are common in many datasets, such as articles and prepositions, are considered less important and given lower TF-IDF weights (Zhang et al., 2020) For example, In the sentence “What a strong and brave man,” the words “strong” and “brave” are given more importance than “what,” “a,” and “man” because they are more specific and less common, making them key terms that identify the main topic.

3.7. Use of data optimizing techniques

It is common for data sets to have an unequal number of comments across themes. For example, in this study, appraisal support has 3,375 comments, whereas emotional support has 600 comments. Therefore, balancing these themes or classes is crucial to incorporate the intricacies of data and achieve accurate results. We used two class balancing techniques to optimize our data: Class Weight Balancing Technique (class_weight = balance) and Synthetic Minority Over-sampling Technique (SMOTE). The class weight balancing technique balances the classes or themes by increasing the weightage of themes with a low number of comments to represent all themes equally during the machine analysis process. SMOTE creates additional data points for the theme with fewer comments. It creates new, unique data that represents the theme to ensure that all the themes are represented equally during the machine analysis process (Abdelhamid & Desai, 2024; Tanha et al., 2020). In

ML, hyperparameter optimization is also used for the training data to improve the NLP model performance (Probst et al., 2019). We used GridSearchCV to choose a set of optimal parameters. We tested all 5 models for each potential hyperparameter combination ten times to get the most effective parameter for each model. Once the data was prepared and optimized, we processed the data.

4. NLP classification models

ML programs computers to perform specific tasks such as classification, regression, and clustering (Mukhamediev et al., 2022). This study focuses on predicting discrete values using labeled data, specifically through classification tasks with supervised ML algorithms. Therefore, we used five NLP classification models.

- Decision Tree (DT) for topic classification functions like a flowchart, asking yes-or-no questions about the numerical representation of words in a comment to categorize it into topics (Rahman & Akter, 2019).
- Random Forest (RF) uses multiple decision trees and averages their outputs for more accurate results compared to a single decision tree (Jalal et al., 2022).
- Logistic Regression classifies comments by using probabilities based on word patterns and frequencies (Saif et al., 2018).
- Support Vector Machine (SVM) classifies topics by recognizing complex patterns in the numerical representation of words (Jalal et al., 2022).
- Multinomial Naïve Bayes (NB) classifies topics by calculating word likelihoods and predicting the most likely category based on word frequency (Xu et al., 2017).

5. Results

We used the supervised ML approach and tested the five NLP classification models. We divided the human-determined data into an 80% (5,741 comments) and 20% (1,436 comments) split for training and testing sets, respectively.

Table 1 provides a classification report that depicts NLP models' overall performance based on accuracy, precision, recall and F1 score. Each model was evaluated on the class weight balancing technique, SMOTE and without both techniques except MNB. MNB cannot be evaluated using class weight balancing techniques due to inherent limitations.

Table 1 shows that LR performed better with class weight balancing techniques and without class balancing techniques. SVM performed better with all class weight balancing techniques, SMOTE and without any class balancing techniques compared to other models. However, when we look at the results across classes or themes (see Table 2), LR with class weight balancing techniques provides the most balanced results for minority themes (informational and emotional), considering the real-life data problem of unequal comments in each theme within a data set.

Table 2 provides a comparison of LR and SVM with and without class balancing techniques. LR and SVM without using the class balancing techniques have an accuracy of 0.67 and 0.68, respectively (see Table 1), depicting the best scores among all the results. However, across themes, results indicate a poor balance between precision and recall,

suggesting a misleading F1 score (the harmonic mean of precision and recall). LR in emotional support has 0.72 precision and a recall of 0.24, highlighting a significant difference between the two values. Similarly, SVM presents a significant difference between the precision and recall values of minority themes. It has a precision of 0.76 and a recall of 0.21 for the emotional support theme, and a precision of 0.73 and a recall of 0.19 for the informational support.

Table 1
Classification report

NLP Model	Optimizing techniques	Accuracy	F1 Score	Precision	Recall
LR	Class Weight Balance Technique	0.66	0.65	0.65	0.66
	SMOTE	0.60	0.61	0.61	0.60
	None	0.67	0.65	0.67	0.67
SVM	Class Weight Balance Technique	0.67	0.65	0.67	0.67
	SMOTE	0.68	0.65	0.67	0.68
	None	0.68	0.65	0.69	0.68
RF	Class Weight Balance Technique	0.60	0.59	0.62	0.60
	SMOTE	0.55	0.53	0.60	0.55
	None	0.62	0.55	0.56	0.62
DT	Class Weight Balance Technique	0.51	0.51	0.54	0.51
	SMOTE	0.51	0.51	0.53	0.51
	None	0.58	0.56	0.56	0.62
MNB	SMOTE	0.62	0.62	0.63	0.62
	None	0.65	0.62	0.64	0.65

Note. MNB cannot be evaluated using class weight balancing techniques due to inherent limitations

LR and SVM both performed well with the class weight balancing technique. Both have high precision and recall, leading to an F1 score of 0.71 to 0.74 for the themes with a high number of comments, such as appraisal support and ineligible comments. For informational and emotional support that has a low number of comments and requires class balancing techniques to balance the data, SVM and LR performance are similar for informational support, with 0.36 and 0.35 F1 scores, respectively. However, for emotional support, LR performed better with a 0.49 F1 score compared to a 0.37 F1 score. Therefore, we conclude that LR with the class weight balancing technique is the best model for such data.

A confusion matrix displays the effectiveness of a classification model in terms of the actual and predicted classes. We used a multi-class confusion matrix due to three themes and the category of ineligible comments. It depicts the number of true positives (correctly identified comments within a theme or class); true negatives (correctly identified comments that do not belong to that particular theme or class); false positives (incorrectly identified comments within a theme or class); and false negatives (incorrectly identified comments that do not belong to the particular theme or class) (Müller & Guido, 2016). Figs 2 and 3 provide the confusion matrix for LR and SVM using the class weight balancing technique, respectively. Table 3 provides true positives, true negatives, false positives and false negatives data extracted from the confusion matrix shown in Fig. 2 and 3.

Table 2

Comparison of LR and SVM with and without class balancing techniques

NLP Models	Themes	Precision	Recall	F1-score
LR	Appraisal support	0.69	0.73	0.71
	Emotional support	0.64	0.40	0.49
	Informational support	0.50	0.27	0.35
	Ineligible	0.66	0.80	0.72
	Appraisal support	0.70	0.61	0.65
	Emotional support	0.33	0.39	0.36
	Informational support	0.39	0.40	0.39
	Ineligible	0.65	0.75	0.70
	Appraisal support	0.67	0.78	0.72
	Emotional support	0.72	0.24	0.36
	Informational support	0.59	0.25	0.35
	Ineligible	0.68	0.79	0.73
SVM	Appraisal support	0.67	0.79	0.73
	Emotional support	0.68	0.25	0.37
	Informational support	0.58	0.26	0.36
	Ineligible	0.69	0.79	0.74
	Appraisal support	0.68	0.84	0.75
	Emotional support	0.67	0.28	0.39
	Informational support	0.58	0.19	0.29
	Ineligible	0.72	0.77	0.74
	Appraisal support	0.67	0.83	0.74
	Emotional support	0.76	0.21	0.33
	Informational support	0.73	0.19	0.30
	Ineligible	0.70	0.80	0.74

Table 3

LR and SVM comparison with the class weight balancing technique

Types of support and total number of comments used for testing data	True positives		True negative		False positives		False negative	
	LR	SVM	LR	SVM	LR	SVM	LR	SVM
Appraisal support (N = 704)	513	554	501	462	231	270	191	150
Emotional support (N = 119)	48	30	1290	1303	27	14	71	89
Ineligible comments (N = 424)	339	334	837	863	175	149	85	90
Informational support (N = 189)	51	49	195	1211	52	36	138	140

Table 3 indicates LR and SVM correctly identified 531 (72%) and 554 (78%) comments out of 704 comments used in the testing data set for appraisal support, a theme containing a high number of comments. For the Ineligible comments, another category with a high number of comments, LR identified 339 (80%) and 334 (78%) out of 424 comments. These results suggest that both models performed well at predicting the correct number of comments in a high comments category. For information support, a theme with a low number of comments, LR and SVM showed similar performance for correctly identified comments 51 (27%) and 49 (26%) respectively, out of 189 comments. For emotional support, another theme with low comments, LR correctly identified 48 (40%) comments, and SVM identified 30 (25%) comments out of 119 comments. This theme had overall poor results; however, in comparison, LR performed better than the SVM.

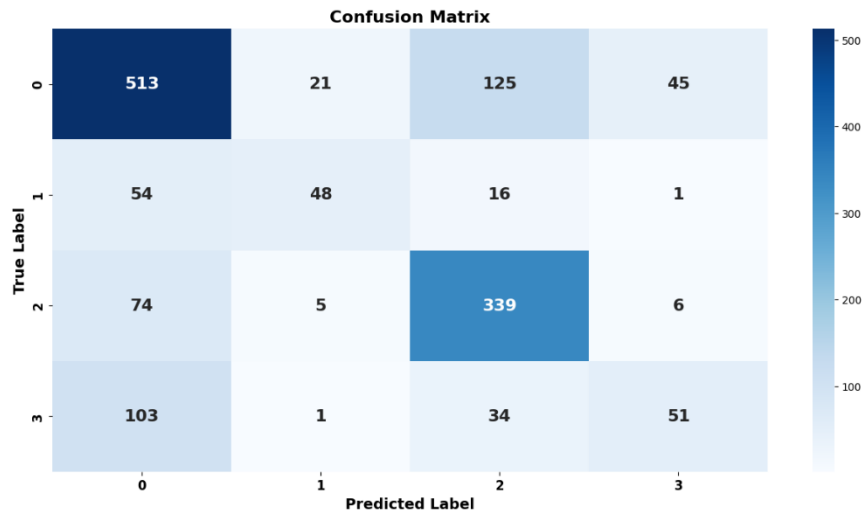


Fig. 2. Confusion matrix for LR with class weight balancing technique (Legend: Class 0: Appraisal support; Class 1: Emotional support; Class 2: Ineligible; and Class 3: Informational support)

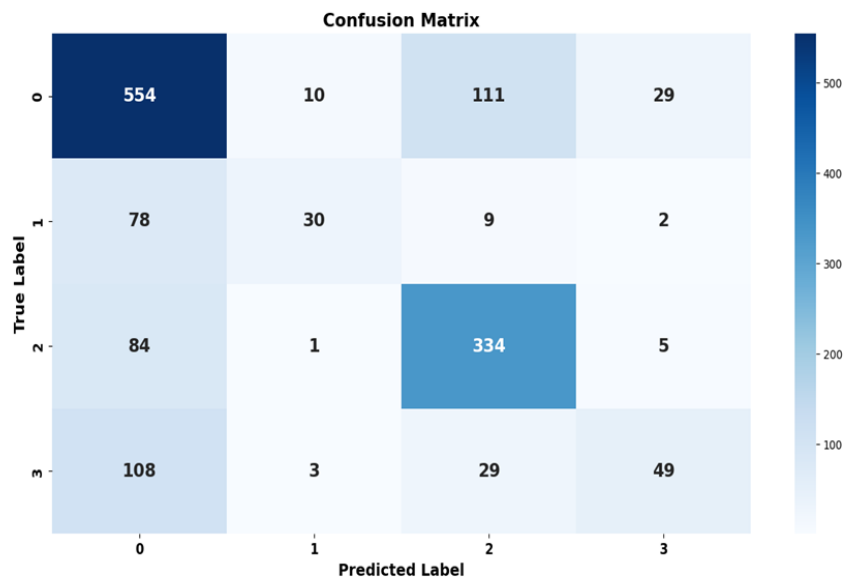


Fig. 3. Confusion matrix for SVM with class weight balancing techniques (Legend: Class 0: Appraisal support; Class 1: Emotional support; Class 2: Ineligible; and Class 3: Informational support)

6. Discussions

Our objective was to evaluate the potential of ML-NLP analysis techniques for the thematic analysis of complex health datasets. We compared ML-NLP classification models with traditional manual deductive thematic analysis. By triangulating the results from human-determined and machine-assisted analyses, we assessed the two methods in terms of time required for analysis, accuracy, and effectiveness in handling complex health datasets.

Our human analysis took 184 hours, longer than similar studies, while machine-assisted analysis took only 34 hours, aligning with other studies (Li et al., 2021; Towler et al., 2023). Complex health data sets demand more human hours due to their intricate nature, and the researchers' subjectivity, such as knowledge, experience, and prior exposure, which influence data categorization. Humans interpret data considering context, language characteristics, and tone, which is crucial for meaningful analysis of complex subjects. In contrast, most ML-NLP models need to remove punctuation and other characteristics during data preparation (Towler et al., 2023). For example, Guetterman et al. (2018) found that NLP models misinterpreted contextual data and generated slightly different themes than humans. Traditional thematic analysis also requires data checking, like team consensus, to ensure trustworthiness, making it time and resource-intensive.

In our study, the LR classification model with class weight balancing techniques achieved an accuracy of 0.66, which is lower than the accuracy reported in similar studies (Gurcan, 2018; Joshi & Abdelfattah, 2021; Kadhim, 2019; Parmar et al., 2018). One possible explanation for the lower accuracy is that our data set was highly imbalanced, creating bias toward the majority class despite the use of class balancing techniques (Padurariu & Breaban, 2019; Thabtah et al., 2020). Thematic analysis categorizes data into meaningful units rather than focusing on numbers, facilitating the identification of patterns or themes within complex data sets and providing deeper insights and understanding. By focusing on themes rather than numbers in each category, thematic analysis provides a rich, detailed account of the data, making it a powerful tool for understanding complex phenomena and generating new knowledge (Braun & Clarke, 2006). In contrast, ML-NLP approaches focus on numbers, which may lead to superficial data analysis and overlook the intricacies of complex matters.

In terms of effectiveness in handling complex health datasets, supervised ML-NLP is mostly suitable for descriptive studies that require extracting higher-level themes from large datasets. It may not be as beneficial for complex datasets that require inductive analysis and human interpretations beyond numerical data (Towler et al., 2023). For example, similar data analyzed inductively by research teams provided insights beyond social support, focusing more on stigma-related themes (Virani et al., 2024).

We noted that supervised ML-NLP may not be suitable for complex health datasets available on social media. However, we suggest a combined approach, where ML-NLP techniques assist with thematic analysis, which can be beneficial. Social media data is not typically collected for research purposes, and contributors provide data based on their interests, resulting in a significant amount of ineligible data. For instance, in our study, 2,262 comments were deemed ineligible by the research team. ML-NLP approaches can help remove ineligible data, allowing researchers to focus on valuable data. This method can save time and improve the efficiency of traditional manual thematic analysis. Guetterman et al. (2018) experimented with a small dataset (126 text messages) using three approaches: NLP-only, qualitative-only, and a combined NLP and qualitative approach. They found no significant difference between the combined and qualitative-only

approaches but noted that the combined analysis was more comprehensive and provided a better understanding of the data.

This study's limitations include a small, highly imbalanced dataset, which restricts the generalizability of the findings. Future research should aim to use larger datasets with slightly imbalanced classes to better explore the potential of NLP in analyzing complex health datasets. Moreover, investigating the potential of unsupervised ML approaches could provide valuable insights. By addressing these limitations, future studies can enhance the robustness and applicability of ML-NLP techniques in complex health data analysis.

7. Conclusion

In this study, we explored the feasibility and effectiveness of using ML-NLP approaches to augment the resource intensive thematic analysis of social media data related to complex health issues. We utilized five NLP classification models, incorporating class balancing techniques to enhance performance. Our findings revealed that the intricacies of social media data on complex health issues, imbalanced data across themes, human subjectivity, and the human ability to consider context for analysis present challenges to the optimal use of supervised ML-NLP approaches. Despite these challenges, social media data holds significant potential for providing timely insights into public health needs, exploring current health trends, and evaluating public responses to health promotion interventions and policy implementations. Future research should focus on further exploring AI tools to augment thematic analysis, thereby supporting researchers in maximizing the potential of social media data analysis. This could lead to more efficient and insightful analyses, ultimately contributing to better-informed public health strategies and interventions.

Author Statement

The authors declare that there is no conflict of interest.

Acknowledgements

We are thankful to Thompson Rivers University (TRU) for providing the Internal Research Fund for this project. We also acknowledge the work of the research assistants and the other team members who helped us with human-determined analysis, and the second author, AM, as part of this work was derived from his Master's project, completed to fulfill the requirements for an MSc in Data Science degree at TRU. Preliminary analysis of this work is also presented at the Twentieth International Conference on Technology, Knowledge, and Society, in March 2024.

ORCID

Anila Virani  <https://orcid.org/0000-0001-8751-0756>

Ahsan Mollani  <https://orcid.org/0009-0002-7854-3905>

Piper Jackson  <https://orcid.org/0000-0002-7025-5063>

References

- Abdelhamid, M., & Desai, A. (2024). Balancing the scales: A comprehensive study on tackling class imbalance in binary classification. *arXiv: 2409.19751* <https://doi.org/10.48550/arXiv.2409.19751>
- Archer, C., & Kao, K. T. (2018). Mother, baby and Facebook makes three: Does social media provide social support for new mothers? *Media International Australia*, 168(1), 122–139. <https://doi.org/10.1177/1329878X18783016>
- Auxier, B., Anderson, M., Perrin, A., & Turner, E. (2020). *Parents' attitudes – and experiences – related to digital technology*. Pew Research Center. Retrieved from <https://www.pewresearch.org/internet/2020/07/28/parents-attitudes-and-experiences-related-to-digital-technology/>
- Azizi, M. (2022). *A comparison of machine learning techniques to classify tweets relevant to people impacted by dementia and COVID-19*. Master's thesis, University of Saskatchewan, Canada. Retrieved from <https://harvest.usask.ca/items/765755da-a18d-470e-b392-b32601110bb0>
- Bacsu, J. D., Fraser, S. A., Jamali, A. A., Conanan, C., Chasteen, A. L., Vellani, S., ... Spiteri, R. J. (2024). Navigating awareness and strategies to support dementia advocacy on social media during World Alzheimer's Month: Infodemiology study. *JMIR Infodemiology*, 4: e63464. <https://doi.org/10.2196/63464>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Castleberry, A., & Nolen, A. (2018). Thematic analysis of qualitative research data: Is it as easy as it sounds? *Currents in Pharmacy Teaching and Learning*, 10(6), 807–815. <https://doi.org/10.1016/j.cptl.2018.03.019>
- Chen, Q., Li, W., Xiong, J., & Zheng, X. (2022). Prevalence and risk factors associated with postpartum depression during the COVID-19 pandemic: A literature review and meta-analysis. *International Journal of Environmental Research and Public Health*, 19(4): 2219. <https://doi.org/10.3390/ijerph19042219>
- Dadi, A. F., Miller, E. R., & Mwanri, L. (2020). Postnatal depression and its association with adverse infant health outcomes in low-and middle-income countries: A systematic review and meta-analysis. *BMC Pregnancy and Childbirth*, 20: 416. <https://doi.org/10.1186/s12884-020-03092-7>
- Dol, J., Richardson, B., Grant, A., Aston, M., McMillan, D., Tomblin Murphy, G., & Campbell-Yeo, M. (2021). Influence of parity and infant age on maternal self-efficacy, social support, postpartum anxiety, and postpartum depression in the first six months in the Maritime Provinces, Canada. *Birth*, 48(3), 438–447. <https://doi.org/10.1111/birt.12553>
- Gage-Bouchard, E. A., LaValley, S., Mollica, M., & Beaupin, L. K. (2017). Cancer communication on social media: Examining how cancer caregivers use Facebook for cancer-related communication. *Cancer Nursing*, 40(4), 332–338. <https://doi.org/10.1097/NCC.0000000000000418>
- Gopalakrishnan, A., Gururajan, R., Venkataraman, R., Zhou, X., Ching, K. C., Saravanan, A., & Sen, M. (2023). Attribute selection hybrid network model for risk factors analysis of postpartum depression using social media. *Brain Informatics*, 10: 28. <https://doi.org/10.1186/s40708-023-00206-7>
- Guetterman, T. C., Chang, T., DeJonckheere, M., Basu, T., Scruggs, E., & Vydiswaran, V. G. V. (2018). Augmenting qualitative text analysis with natural language processing: Methodological study. *Journal of Medical Internet Research*, 20(6): e231. <https://doi.org/10.2196/jmir.9702>

- Gurcan, F. (2018, October). Multi-class classification of Turkish texts with machine learning algorithms. In *Proceedings of the 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*. IEEE. <https://doi.org/10.1109/ISMSIT.2018.8567307>
- Han, C. J., Lee, Y. J., & Demiris, G. (2018). Interventions using social media for cancer prevention and management: A systematic review. *Cancer Nursing*, 41(6), E19–E31. <https://doi.org/10.1097/NCC.0000000000000534>
- Haslam, D. M., Tee, A., & Baker, S. (2017). The use of social media as a mechanism of social support in parents. *Journal of Child and Family Studies*, 26(7), 2026–2037. <https://doi.org/10.1007/s10826-017-0716-6>
- Hemphill, L., Schöpke-Gonzalez, A., & Panda, A. (2022). Comparative sensitivity of social media data and their acceptable use in research. *Scientific Data*, 9: 643. <https://doi.org/10.1038/s41597-022-01773-w>
- Iyortsuun, N. K., Kim, S. H., Jhon, M., Yang, H. J., & Pant, S. (2023). A review of machine learning and deep learning approaches on mental health diagnosis. *Healthcare*, 11(3): 285. <https://doi.org/10.3390/healthcare11030285>
- Izadirad, H., Niknami, S., Zareban, I., & Hidarnia, A. (2017). Effects of social support and self-efficacy on maternal prenatal cares among the first-time pregnant women, Iranshahr, Iran. *Journal of Family & Reproductive Health*, 11(2), 67–73. Retrieved from <https://jfrh.tums.ac.ir/index.php/jfrh/article/view/545>
- Jakhotiya, A., Jain, H., Jain, B., & Chaniyara, C. (2022). Text pre-processing techniques in natural language processing: A review. *International Research Journal of Engineering and Technology*, 9(2), 878–880. Retrieved from <https://www.irjet.net/archives/V9/i2/IRJET-V9I2150.pdf>
- Jalal, N., Mehmood, A., Choi, G. S., & Ashraf, I. (2022). A novel improved random forest for text classification using feature ranking and optimal number of trees. *Journal of King Saud University – Computer and Information Sciences*, 34(6), 2733–2742. <https://doi.org/10.1016/j.jksuci.2022.03.012>
- Johnson, J., Black, R. W., & Hayes, G. R. (2020, October). Roles in the discussion: An analysis of social support in an online forum for people with dementia. In *Proceedings of the ACM on Human-Computer Interaction (Vol. 4, No. CSCW2)*. Association for Computing Machinery. <https://doi.org/10.1145/3415198>
- Joshi, S., & Abdelfattah, E. (2021, May). Multi-class text classification using machine learning models for online drug reviews. In *Proceedings of the 2021 IEEE World AI IoT Congress (AIIoT)*. IEEE. <https://doi.org/10.1109/AIIoT52608.2021.9454250>
- Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review*, 52, 273–292. <https://doi.org/10.1007/s10462-018-09677-1>
- Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, 82(3), 3713–3744. <https://doi.org/10.1007/s11042-022-13428-4>
- Leahy-Warren, P. (2014). Social support theory. In J. J. Fitzpatrick & G. McCarthy (Eds.), *Theories Guiding Nursing Research and Practice: Making Nursing Knowledge Development Explicit* (pp.85–101). Springer.
- Le Glaz, A., Haralambous, Y., Kim-Dufor, D. H., Lenca, P., Billot, R., Ryan, T. C., ... Lemey, C. (2021). Machine learning and natural language processing in mental health: Systematic review. *Journal of Medical Internet Research*, 23(5): e15708. <https://doi.org/10.2196/15708>
- Li, Y., Shyr, C., Borycki, E. M., & Kushniruk, A. W. (2021). Automated thematic analysis of health information technology (HIT) related incident reports. *Knowledge*

- Management & E-Learning*, 13(4), 408–420. <https://doi.org/10.34105/j.kmel.2021.13.022>
- Mukhamediev, R. I., Popova, Y., Kuchin, Y., Zaitseva, E., Kalimoldayev, A., Symagulov, A., ... Yelis, M. (2022). Review of artificial intelligence and machine learning technologies: Classification, restrictions, opportunities and challenges. *Mathematics*, 10(15): 2552. <https://doi.org/10.3390/math10152552>
- Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with Python: A guide for data scientists*. O'Reilly.
- Padurariu, C., & Breaban, M. E. (2019). Dealing with data imbalance in text classification. *Procedia Computer Science*, 159, 736–745. <https://doi.org/10.1016/j.procs.2019.09.229>
- Parmar, P. S., Biju, P. K., Shankar, M., & Kadiresan, N. (2018, September). Multiclass text classification and analytics for improving customer support response through different classifiers. In *Proceedings of the 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE. <https://doi.org/10.1109/ICACCI.2018.8554881>
- Probst, P., Boulesteix, A. L., & Bischl, B. (2019). Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research*, 20, 1–32. Retrieved from <https://jmlr.org/papers/volume20/18-444/18-444.pdf>
- Rahman, M. A., & Akter, Y. A. (2019, May). Topic classification from text using decision tree, K-NN and multinomial naïve bayes. In *Proceedings of the 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*. IEEE. <https://doi.org/10.1109/ICASERT.2019.8934502>
- Safa, R., Bayat, P., & Moghtader, L. (2022). Automatic detection of depression symptoms in Twitter using multimodal analysis. *Journal of Supercomputing*, 78(4), 4709–4744. <https://doi.org/10.1007/s11227-021-04040-8>
- Saif, M. A., Medvedev, A. N., Medvedev, M. A., & Atanasova, T. (2018, June). Classification of online toxic comments using the logistic regression and neural networks models. In *Proceedings of the 44th International Conference on Applications of Mathematics in Engineering and Economics (AMEE'18)*. AIP Publishing. <https://doi.org/10.1063/1.5082126>
- Sajno, E., Bartolotta, S., Tuena, C., Cipresso, P., Pedrolì, E., & Riva, G. (2023). Machine learning in biosignals processing for mental health: A narrative review. *Frontiers in Psychology*, 13: 1066317. <https://doi.org/10.3389/fpsyg.2022.1066317>
- Singh, A., & Shashi, M. (2019). Vectorization of text documents for identifying unifiable news articles. *International Journal of Advanced Computer Science and Applications*, 10(7), 305–310. <https://doi.org/10.14569/ijacsa.2019.0100742>
- Statistics Canada. (2019). *Maternal mental health in Canada, 2018/2019*. Retrieved from <https://www150.statcan.gc.ca/n1/daily-quotidien/190624/dq190624b-eng.htm>
- Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., & Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification: An experimental review. *Journal of Big Data*, 7: 70. <https://doi.org/10.1186/s40537-020-00349-y>
- Tatman, R. (2018). *What you can, can't and shouldn't do with social media data*. Paper presented at Joint Statistical Meetings. Vancouver BC. Retrieved from <https://makingnoiseandhearingthings.com/2018/08/31/what-you-can-cant-and-shouldnt-do-with-social-media-data/>
- Taylor, B., Henshall, C., Kenyon, S., Litchfield, I., & Greenfield, S. (2018). Can rapid approaches to qualitative analysis deliver timely, valid findings to clinical leaders? A mixed methods study comparing rapid and thematic analysis. *BMJ Open*, 8(10):

- e019993. <https://doi.org/10.1136/bmjopen-2017-019993>
- Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. *Information Sciences*, 513, 429–441. <https://doi.org/10.1016/j.ins.2019.11.004>
- Towler, L., Bondaronek, P., Papakonstantinou, T., Amlôt, R., Chadborn, T., Ainsworth, B., & Yardley, L. (2023). Applying machine-learning to rapidly analyze large qualitative text datasets to inform the COVID-19 pandemic response: Comparing human and machine-assisted topic analysis techniques. *Frontiers in Public Health*, 11: 1268223. <https://doi.org/10.3389/fpubh.2023.1268223>
- Virani, A., Nagra, B., O'Mahony, J., Bacsu, J., Ghatore, J., & Panda, S. (2024). Using YouTube comments data to explore postpartum depression in social media: An infodemiology study. *International Journal of Environmental Research and Public Health*, 21(11): 1526. <https://doi.org/10.3390/ijerph21111526>
- World Health Organization. (2022). *Statement on the tenth meeting of the International Health Regulations (2005) Emergency Committee regarding the coronavirus disease (COVID-19) pandemic*. World Health Organization. Retrieved from [https://www.who.int/news/item/19-01-2022-statement-on-the-tenth-meeting-of-the-international-health-regulations-\(2005\)-emergency-committee-regarding-the-coronavirus-disease-\(covid-19\)-pandemic](https://www.who.int/news/item/19-01-2022-statement-on-the-tenth-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-coronavirus-disease-(covid-19)-pandemic)
- World Health Organization. (2023). *Statement on the fifteenth meeting of the IHR (2005) Emergency Committee on the COVID-19 pandemic*. World Health Organization. Retrieved from <https://www.who.int/news/item/05-05-2023-statement-on-the-fifteenth-meeting-of-the-international-health-regulations-%282005%29-emergency-committee-regarding-the-coronavirus-disease-%28covid-19%29-pandemic>
- Xu, S., Li, Y., & Wang, Z. (2017, May). Bayesian multinomial naïve bayes classifier to text classification. In *Proceedings of the 11th International Conference on Multimedia and Ubiquitous Engineering (MUE2017) and the 12th International Conference on Future Information Technology (FutureTech2017)*. Springer. https://doi.org/10.1007/978-981-10-5041-1_57
- Zhang, Y., Zhou, Y., & Yao, J. (2020, June). Feature extraction with TF-IDF and game-theoretic shadowed sets. In *Proceedings of the 18th International Conference of Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2020)* (pp. 722–733). Springer. https://doi.org/10.1007/978-3-030-50146-4_53