Detecting Behavioral and Emotional Themes Through Latent and Explicit Knowledge

Anonymous ACL submission

Abstract

Social science research increasingly employs Natural Language Processing (NLP) to analyze large-scale textual data, yet common methods like topic modeling and sentiment analysis often overlook the nuanced ways in which emotions and cultural contexts shape meaning. To address this gap, we introduce the Behavioral and Emotional Theme Detection (BET) framework—a novel approach that integrates emotional, cultural, and sociological dimensions into topic detection and emotion analysis. By applying BET to English and Hebrew datasets, we showcase its multilingual adaptability and its potential to reveal rich thematic content and emotional resonance in biographical texts. Our results demonstrate that BET not only enhances the granularity and diversity of detected themes but also tracks shifts in emotional framing over time, offering deeper insights into how individuals deploy linguistic resources to position their identities.[1]

1 Introduction

The complex interplay between language and society is a fundamental domain of inquiry within the social sciences (Alvero, 2023). Language not only reflects, but also actively shapes social structures (Bourdieu, 1993), serving as a primary mechanism through which individuals construct and perform their identities. However, capturing the social meaning embedded in language remains a challenge for NLP methods. While topic modeling and sentiment analysis provide useful tools (Franzosi et al., 2022; Zwilling, 2023), they fail to account for emotional resonance, and contextual embeddedness that shape how language operates within specific social fields and institutional settings.

One specific domain where these challenges manifest is the analysis of biographical narratives—personal reflections on significant life events that are widely used in university admissions, scholarships, hiring, and promotion that offer a rich lens for examining the dynamics between language and personal traits. In this domain, factors such as agentive positioning, rhetorical strategies, and temporal framing are often linguistic resources through which individuals position themselves within institutional contexts (Kemper, 2006; Dijk, 2009; Kleres, 2011; Berger and Packard, 2022).

However, traditional NLP analyses such as topic modeling and sentiment analysis often overlook these deeper sociolinguistic dimensions when applied separately. Despite recent attempts to integrate these tasks (Wang et al., 2017; Yin et al., 2022; Ma et al., 2023), an inclusive method that captures both topic and emotional tone—enabling the derivation of cohesive conclusions deeply embedded within narratives—remains absent.

Addressing this gap requires NLP analyses to integrate thematic and affective dimensions to better capture the nuanced ways in which individuals deploy language to construct their identities and to be able to answer the following key theoretical questions: To what extent do shared biographical experiences correspond to similar linguistic patterns? and how do sentiment-laden expressions align or diverge across narratives that describe comparable life events or circumstances?

In this paper, we introduce a novel framework for Behavioral and Emotional Theme Detection (BET) to overcome these limitations and to illuminate the inherent relationship between thematic content and emotional expression in sociocultural narratives. Our proposed methodology integrates explicit knowledge from the Linguistic Inquiry and Word Count (LIWC) lexicon with modern embedding techniques, using pre-trained language models to compute semantic similarities between latent embeddings of text and predefined emotional categories. An illustration of our framework is pre-

[1] An anonymized version of our code is available at: `https://anonymous.4open.science/r/BET-0887`.

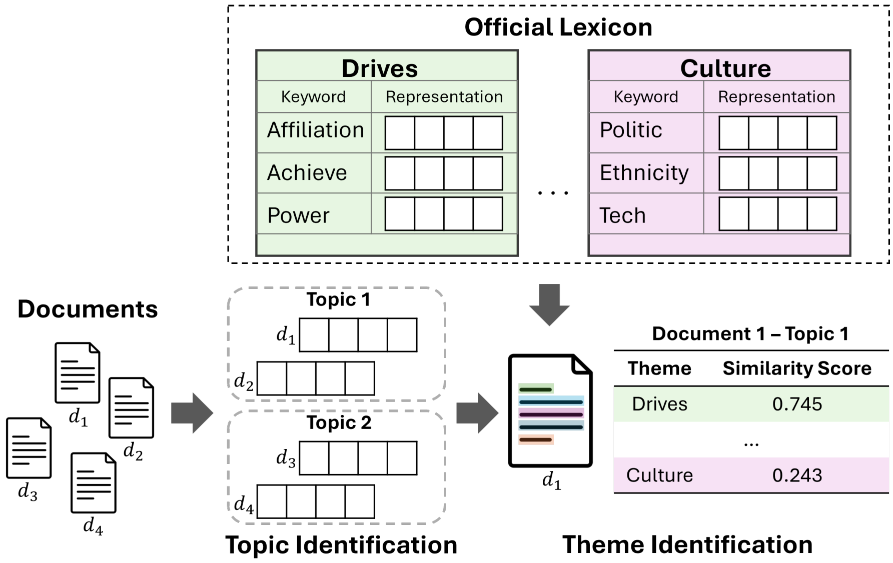

Figure 1: An illustration of the methodological framework for behavioral and emotional theme detection (BET).

sented in Figure 1.

Unlike conventional text analysis methods, BET emphasizes Affective and sociological dimensions within the thematic context. For example, analyzing narratives of previous educational experiences reveals not only the content but also the emotional frames through which these experiences are interpreted and conveyed—whether through expressions of anger, achievement, conflict, or other sentimental orientations and perspectives.

The proposed methodology is particularly compelling when applied to diverse languages. Drawing on established research (Markus and Kitayama, 1991; Stets and Turner, 2010; Jackson et al., 2019), we recognize how culture shapes emotional expression and narrative construction, reflecting the unique social contexts of their respective linguistic environments. Hence, we demonstrate the application of BET on two datasets, an English-language synthetic student profile dataset from a public Kaggle repository and a Hebrew-language financial aid application dataset provided to us by a known fellowship foundation. The cross-linguistic analysis not only enriches the variety of extracted topics but illuminates the universality of the proposed approach and its adaptability to languages with varying linguistic properties.

Our contributions are as follows: (1) We introduce a novel framework for detecting behavioral and emotional themes by integrating topic detection with linguistic patterns; (2) We develop a hybrid methodology that combines the LIWC lexicon with embedding techniques for fine-grained emotional theme detection; (3) We demonstrate the method's adaptability using datasets in English and Hebrew, addressing varying morphological and syntactical complexities; and (4) We illustrate how the method improves the content analysis and provides new insights for social analysis.

2 Previous Work

The outputs of topic detection, emotion recognition, and sentiment analysis provide valuable insights for research about thematic contents and emotional tones of texts. We review key developments in these areas, highlighting their relevance to the framework presented in this study.

2.1 Topic Detection

Traditional topic detection models use supervised and unsupervised methods to identify the topics and main themes in texts (Blei et al., 2003; Griffiths et al., 2003; Teh et al., 2004; Blei and Lafferty, 2006; Mcauliffe and Blei, 2007; Ramage et al., 2009; Meng et al., 2019). Recent unsupervised methods, such as BERTopic (Grootendorst, 2022) and Top2Vec (Angelov, 2020), integrate contextual embeddings to improve topic coherence and applicability to diverse datasets. In this work, we utilize BERTopic for both document-level and sentence-level topic modeling. Document-level analysis assigns a single topic to each document, offering a high-level thematic overview of the text. Sentence-level analysis, if employed, allows for detecting fine-grained subtopics within individual documents, capturing nuanced thematic elements.

2.2 Sentiment and Emotion Analysis

Sentiment analysis models identify the emotional tones expressed in texts, often focusing on simple polarity (positive, neutral, negative). Emotion analysis is a sub-task in sentiment analysis that offers a finer granularity sentiment level compared with polarity analysis. The most frequent emotion categories in the literature are based on Ekman's model (Ekman et al., 1987) which are anger, fear, sadness, joy, disgust, and surprise. Plutchik's model (Plutchik, 2001), also adds trust and anticipation.

Another relevant task in this realm is Aspect-Based Sentiment Analysis (ABSA), where a model identifies the sentiment associated with specific aspects or entities in a text, enabling a more detailed analysis compared to document-level approaches (Liang et al., 2022; Li et al., 2023). For example, in a restaurant review, "The food was delicious, but the service was terrible," ABSA would identify Food (aspect) with a positive sentiment and Service (aspect) with a negative sentiment.

Despite recent advancements, ABSA models primarily focus on polarity detection, often overlooking the nuanced emotional spectrum associated with various aspects. This highlights the necessity for integrating emotion analysis. In this work, we propose a model to identify emotions and other sociological concepts related to identified topics in the text. Our approach can also be applied to the extracted aspects from a text, resulting in a more refined output than traditional ABSA.

2.3 Topic-Sentiment Detection

Topic-sentiment detection integrates topic modeling and sentiment analysis to uncover the sentiments tied to specific topics (Garcia and Berton, 2021; Qi et al., 2024; Zhang et al., 2024). This approach not only identifies what people are discussing but also their sentimental perspectives. For example, in social media posts about climate change, topic-sentiment detection would identify the overall sentiment toward climate change as a topic, and it may also track how sentiments change over time, revealing shifts in public opinion (Dahal et al., 2019; Rosenberg et al., 2023).

Early methods, such as the Topic Sentiment Mixture (TSM) model (Mei et al., 2007) and the Joint sentiment/topic (JST) model (Lin and He, 2009), link thematic content with sentiment, providing initial insights into this integration. Tang et al. (2019) extends this approach by proposing the Hidden Topic-Emotion Transition Model, capturing dynamic interactions between topics and multi-level emotions. More recent advancements, introduce knowledge-aware transformers, linking contextual emotions with topics for dialogue analysis (Zhu et al., 2021). Our work builds on these advances with emotional theme detection which extends topic-based sentiment analysis by integrating fine-grained emotion recognition. Unlike previous methods that focus on broad sentiment categories, our method links varied emotions and sociological aspects to themes, enabling a richer understanding of emotional expressions in multilingual datasets.

3 Methodology

Our method comprises four primary steps. The process begins with transforming the documents into dense embeddings that capture their latent semantic structures. These embeddings form the foundation for unsupervised topic identification, which is conducted using clustering techniques. Next, emotions and sociocultural aspects are detected by comparing sentence embeddings with predefined categories from an official psychological/sociological lexicon. Finally, the detected emotions are combined with the document topics to generate a distribution of emotional themes that describes each document. Figure 1 illustrates our framework, and the following subsections provide a detailed description of the steps.

3.1 Latent Representation

To capture the latent semantic structure of the text, we transform each document, d_i, into a dense vector representation, or embedding, $e_i \in \mathbb{R}^d$, where d is the dimensionality of the embedding space, using Sentence-Transformers (Reimers and Gurevych, 2019), a framework for producing dense embeddings: $e_i = \textsc{SentenceTransformer}(d_i)$. These embeddings encode the semantic meaning of the text, allowing a comprehensive analysis of the relationships within and between documents. The resulting embeddings are subsequently used as input for topic and emotion detection models.

3.2 Topic Detection

We identify document topics by clustering the document embeddings obtained during the latent representation phase, $E = \{e_1, \ldots, e_n\}$, where n is the number of documents. For this purpose, we employ BERTopic (Grootendorst, 2022), a topic modeling framework that leverages dynamic topic representation and unsupervised clustering to uncover meaningful patterns within the data. The embedding step can be applied at the sentence level, allowing multiple topics to be assigned to each document.

The clustering is held via the HDBSCAN algorithm, which groups embeddings e_i into a cluster C_j if they satisfy the density threshold determined by the algorithm parameters. Embeddings that do not meet the density criteria are labeled as noise. Clusters C_j are formed by grouping points where the local density exceeds a threshold. The local density at e_i is inversely proportional to the mutual reachability (MR) distance:

$$\textsc{Density}(e_i) = \frac{1}{\sum_l \text{MR}(e_i, e_l)},$$

where $\text{MR}(e_i, e_l) = \max\{\|e_i - e_l\|, \|e_i - e_{k_{min}}\|, \|e_l - e_{k_{min}}\|\}$, and k_{min} is the minimal number of points required to form a dense region (a hyperparameter of HDBSCAN).

A notable feature of BERTopic is its ability to support dynamic topic modeling, which involves adjusting the topic representations over time across different contexts. This adaptability ensures that the extracted topics remain accurate and relevant even when applied to datasets that evolve or exhibit temporal variations. Dynamic topic modeling is particularly useful for capturing trends, shifts in discourse, and changes in thematic emphasis.

3.3 BET: Behavioral and Emotional Theme Detection

To detect emotional themes we utilize the LIWC lexicon (Pennebaker et al., 2022), a psycholinguistic tool that organizes keywords into predefined emotional, cognitive, and linguistic categories. LIWC is a valuable resource for understanding the psychological and sociological aspects of the text, providing a structured framework to connect textual content with thematic categories such as “Drives,” “States,” and “Emotions.” Table 1 presents the categories and keywords employed in this study to identify emotional themes for English.[2] The analysis incorporates $k = 10$ validated lexical categories (with 62 keywords) derived from the English version of LIWC2022. For other languages, we recommend either using existing multilingual versions of LIWC[3] or translating the keywords and categories from the English LIWC. For the Hebrew corpus we analyze in this paper, we use the Hebrew version of LIWC defined by Shapira et al. (2021), and in consultation with a domain expert, we draw upon $k = 42$ contextually adapted semantic categories (with 7019 keywords). A full list of the categories we chose appears in Appendix B.

In our proposed methodology, each keyword w within a category k from the lexicon is represented with an embedding using the same pre-trained embedding model applied to generate document and sentence embeddings, i.e., $e_w = \textsc{SentenceTransformer}(w)$. Let $W = \{w_1, \ldots, w_m\}$ be the set of lexicon keywords for category k and let $E = \{e_{w_1}, \ldots, e_{w_m}\}$ be their corresponding embeddings. For each document d_i, the similarity score between a sentence embedding $e_{i,s}$, and a keyword embedding e_w is computed using cosine similarity. The strength of theme category k in document d_i is determined by the highest similarity score across all sentences:

$$S_{i,k} = \max_s \textsc{Sim}(e_{i,s}, e_k).$$

For example, in document d_1, if the cosine similarity scores for the keywords "Affiliation," "Achieve," and "Power" are $S_{1,\text{Affiliation}} = 0.545$, $S_{1,\text{Achieve}} = 0.422$, and $S_{1,\text{Power}} = 0.745$, the emotional theme corresponding to the "Drive" category in this document would be determined by the highest score, 0.745. This approach enables context-aware identification of thematic elements across documents, effectively capturing the nuanced emotional and cognitive dimensions within the text.

By combining the explicit structured knowledge of LIWC with latent embeddings, we are able to create a methodology for emotional theme detection. This integration allows us to bridge the gap between traditional lexicon-based approaches and modern contextualized embeddings, enhancing the adaptability of the results. Moreover, the combination ensures that emotional themes are also sensitive to contextual subtleties present in the text.

The methodological framework yields a multidimensional matrix that interweaves topics with their associated affective dimension, with emotional strength represented through similarity scores. Using this output, a domain expert, such as a social science researcher, can create a thorough analysis that connects both the thematic content and the emotional undercurrents, offering a nuanced understanding of the data and its implications within the

[2] Categories and keywords were selected in consultation with a social science domain expert.

[3] LIWC translations are available in https://www.liwc.app/dictionaries/liwc-translations.

Category	Keywords
Drives	Affiliation, Achieve, Power
States	Need, Want, Acquire, Lack, Fulfil, Fatigue
Motive	Reward, Risk, Curiosity, Allure
Time Orientation	Focus past, Focus present, Focus future
Culture	Politic, Ethnicity, Tech
Lifestyle	Leisure, Home, Work, Money, Religious
Physical	Health, Illness, Wellness, Mental, Substances, Sexual, Food, Death
Social	Prosocial, Polite, Conflict, Moral, Communication
Positive Emotions	Amused, Anticipation, Calm, Contentment, Enthusiastic, Interested, Joy, Proud, Surprise, Trust, Vigor
Negative Emotions	Anger, Anxiety, Ashamed, Confusion, Disgust, Fatigue, Guilt, Hostile, Nervous, Sad, Crying, Sarcasm, Smirk, Swear

Table 1: Selected emotional theme categories and their corresponding keywords from LIWC 2022.

context of the research. In the next section, we exemplify the application of our analysis and provide a potential interpretation of the results.

4 Experimental Setup

4.1 Datasets

We examine our method with two datasets.

Financial Aid Application Dataset. The financial aid application dataset was collected by a non-governmental organization that facilitates educational access for socioeconomically disadvantaged students through monetary support mechanisms. The dataset comprises 28,424 financial aid applications spanning 2012-2024, with an annual submission rate of approximately 2,200 applications. A key feature of these applications is the requirement for students to write personal narratives describing significant life challenges, their responses to these challenges, and their future goals.[4] These narratives average 474 words in length (median 473, range 23-1,632) and are written in Hebrew. To protect the privacy of participants and maintain data security, access to this dataset is restricted.

Synthetic Student Profile Dataset A Kaggle dataset[5] that provides a comprehensive collection of student profiles, showcasing a wide spectrum of academic and personal characteristics. Each profile encapsulates demographic information, academic details, hobbies, unique qualities, and personal narratives. From this dataset, consisting of 23,236 synthetically generated observations, we used the 'Story' field that includes a narrative or background story about a student to extract sociocultural themes about the student's background story. The average number of words in the stories is 2073. Example stories from this dataset appear in Appendix §A.

4.2 Preprocessing and Parameters

Preprocessing. The preprocessing pipeline was designed to clean the textual data and prepare it for further analysis. We begin this process by tokenizing the text and removing stop words based on the Hebrew/English stop word list from the NLTK library. We then filter tokens to retain only alphabetic characters and numbers, while excluding tokens with a length of one character and specific cases such as line breaks and dashes.

Sentence Embeddings. Embeddings were generated using the Sentence Transformer of paraphrase-multilingual-MiniLM-L12-v2.[6] To assess the semantic similarity between sentences and emotional categories, cosine similarity was employed.

Topic Detection. We employ BERTopic with all-MiniLM-L6-v2 for English and paraphrase-multilingual-MiniLM-L12-v2 for Hebrew as the underlying embedding model and use HDBSCAN for document clustering. The HDBSCAN configuration includes a minimum cluster size of 5, a minimum sample size of 5, Euclidean distance, and the "excess of mass" clustering method. Additionally, we utilize a custom vectorizer with dynamically determined document frequency thresholds based on the dataset size, with a maximum threshold set to 0.9. The vectorizer incorporates both unigrams and bigrams to capture nuanced contextual relationships. For the Financial Aid Application Dataset,

[4] Applicants were asked to share their life story, describe a meaningful challenge and its impact, explain lessons learned from that experience, and outline their ten-year goals across personal, professional, and social domains.

[5] https://www.kaggle.com/datasets/anthonytherrien/synthetic-student-profiles-dataset?resource=download

[6] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

we set the number of topics to 100, guided by multiple internal experiments and domain expertise in sociology. In contrast, for the Synthetic Student Profiles Dataset, the number of topics is determined automatically by the HDBSCAN clustering method and the chosen hyperparameters.

5 Results

5.1 Synthetic Student Profile Dataset

5.1.1 Topic and Emotion Detection

The analysis of this dataset uncovers distinct thematic and emotional patterns within the narratives. BERTopic identifies 126 topics, with professional trajectories emerging as the most prevalent theme (12.2%), followed by narratives centered on marine and ocean exploration (5.2%), fashion and design (4.1%), environmental consciousness (4.0%), and dance and ballet (3.9%). This distribution reflects a diverse spectrum of student self-representations, ranging from career-oriented aspirations to narratives emphasizing lifestyle, leisure, and personal identity formation.

Given that the texts are synthetically generated rather than penned by real individuals, we expect a generally low emotional valence, as the focus is on creating diverse content rather than expressing emotional perspectives. In line with our expectations, the semantic analysis of the emotional keywords ($k = 62$) that appear in Table 1 reveals relatively low mean affect scores across the dataset. The Drives category exhibits the highest salience (0.25 cosine similarity). Within this category, the drive for achievement (0.32) outweighs both affiliation (0.24) and power orientations (0.19), underscoring the centrality of individualistic success narratives. To demonstrate the analytical utility of our methodological approach, we examine four cases exhibiting heightened emotional articulation, subjecting them to detailed BET analysis.

5.1.2 Behavioral and Emotional Theme Analysis

Figure 2 presents a radar plot illustrating the maximum similarity scores of four selected documents from the synthetic student profile dataset. Each document pertains to a distinct topic: document #546 is associated with Photography & Film, document #16140 with Nutrition & Wellness, document #3676 with Dance & Ballet, and document #8366 with Yoga & Certification Training.

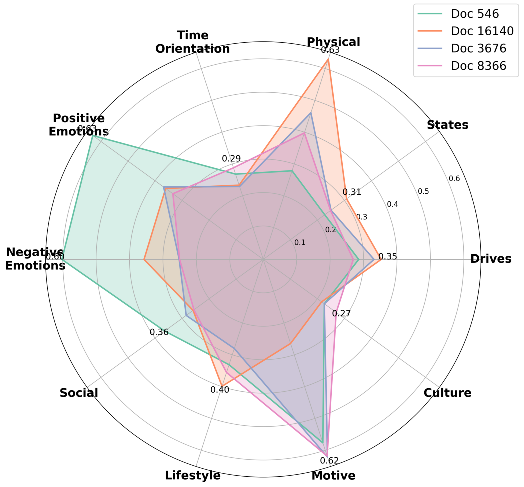

Figure 2: Maximum semantic similarity scores between embeddings of sentences from four selected student profile documents from the synthetic student profile dataset and embeddings of English emotional theme keywords from Table 1.

Focusing on document #546 (Photography & Film), we observe a combination of positive and negative emotions alongside motivational elements. This emotional distribution is reflected in sentences such as “He was nervous but hopeful” and “Bruce remained humble and continued to pursue his studies with the same dedication and passion.” Additionally, the document aligns with its designated topic, as evidenced by sentences like “But what truly set him apart was his love for photography” and “Bruce was a photography aficionado, always carrying his camera with him wherever he went.”

The analysis of this document with BET allows us to demonstrate how emotional and motivational themes relate to the topic of Photography & Film. Similarly, document #16140 (Nutrition & Wellness) exhibits a strong association with the Physical theme, as reflected in its high similarity score (0.63). This correlation is likely influenced by text such as “She believed that taking care of one’s physical and mental well-being was crucial for success and happiness.”

Furthermore, we find that the Motive theme is strongly linked to documents #3676 (0.62) and #8366 (0.61). For document #3676 (Dance & Ballet), this association can be explained by sentences such as “One day, while walking to his Environmental Science class, William heard loud music coming from the auditorium. Curiosity got the better of him, and he decided to check it out.” Similarly, document #8366 (Yoga & Certification Training) also aligns with the Motive theme, likely due to excerpts like “He was thrilled to see how his passion for yoga had brought people together and helped them find inner peace.” Interestingly, both Dance & Ballet and Yoga & Certification Training are disciplines that emphasize achievement within competitive or structured training environments. Both documents demonstrate high scores in the Motive theme, which may suggest a broader connection between motivation and engagement in structured physical and artistic pursuits.

5.2 Financial Aid Application Dataset

5.2.1 Topic and Emotion Detection

Our analysis of this dataset reveals distinct patterns in human-generated narratives. Among the 100 extracted topics, education emerges as the dominant theme, appearing in 1193 (11.8%) documents. These narratives particularly emphasize high school experiences, matriculation certificates, and academic achievements. National service constitutes the second most prevalent topic, present in 811 (8.0%) documents, where applicants detail their military service experiences and related life events. The narratives also prominently feature immigration stories, with rich accounts from diverse ethnic communities, including Ethiopian, former Soviet Union, and French immigrants. Emotional analysis of all applications yields an average cosine similarity score of 0.530, characterized primarily by expressions of trust and interest, with notably low levels of hostility.

5.2.2 Behavioral and Emotional Theme Analysis

The intersection of topics ($j = 100$) and emotional categories ($k = 42$) generates a comprehensive analytical matrix that reveals notable patterns at the intersection of emotions and themes. Figure 3 visualizes the similarity scores of a subset of 10 topics and 10 emotional valences, from which we analyze two topics in-depth: Ballet and Dance and Economic Hardship.

Ballet & Dance Within the broader landscape of extracurricular engagement, the Ballet and Dance topic represents one compelling window into how students articulate their academic trajectories. Analysis of semantic patterns, as presented in Figure 3, shows correlations between the Ballet/Dance topic and multiple positive affect indicators. The Joy marker exhibits the highest mean similarity coefficient (0.593), with three additional dimensions ranking high: absence of hostility (0.582), non-confusion (0.579), and enthusiasm (0.556).

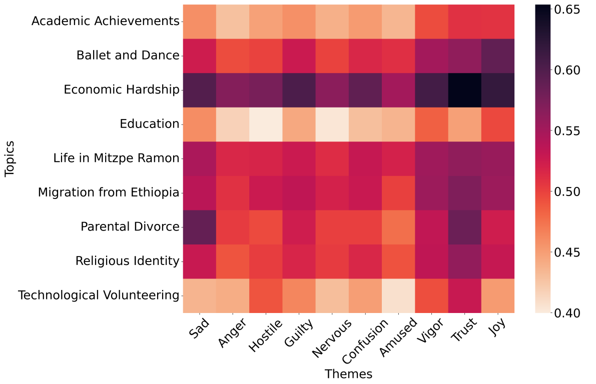

Figure 3: Topic-Theme Semantic Similarity: A sample of the mean score heatmap for the dataset of financial aid applications.

Analysis of participant narratives reveals a consistent conceptualization of dance practice through two primary mechanisms: first, as a structured environment fostering discipline, self-efficacy, and expressive capabilities; and second, as a pathway facilitating access to advanced educational opportunities through professional training. These dual functions align with the previously noted affect markers of reduced hostility and non-confusion. This pattern is exemplified in one participant’s reflection: “The dance teacher’s demands for persistence and dedication are extremely demanding. From a young age, he has instilled in us values of discipline, punctuality, professionalism, and teamwork. This is while teaching us that there are no limits to what we can accomplish if we only desire it. Since the troupe is built in part on hidden competition between members for solo roles, and I don’t always get the lead role, I learn to experience both successes and disappointments—this toughens me for the future.”

Economic Hardship Analysis of the Economic Hardship topic also reveals pronounced affective dimensions. Quantitative findings demonstrate significant correlations across multiple emotional valences such as: not vigor (0.595), not joy (0.623), guilt (0.624), disgust (0.604), and trust (0.670). Participant’s narratives consistently articulate internalized perceptions of familial burden, manifesting in assumed financial responsibilities and labor participation to support household economies. Accounts of witnessing parental navigation of economic precarity, health challenges, and financial obligations emerge as significant catalysts for emotional dis-

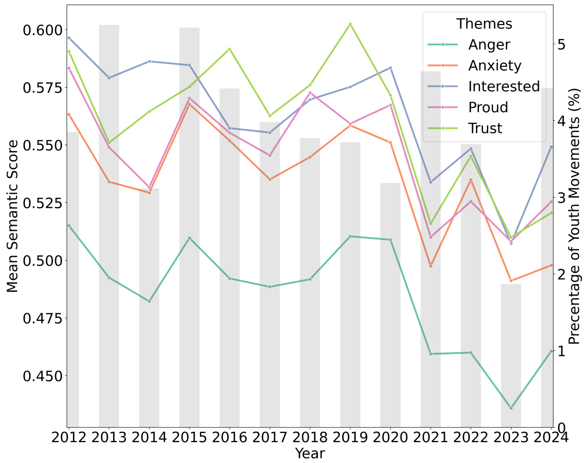

Figure 4: Thematic evolution of the Youth Movements topic over time. The left y-axis indicates the mean cosine similarity score for each theme depicted in the line plot, while the right y-axis denotes the proportion of documents clustered to the Youth Movements topic for each year depicted by the gray bars.

tress, correlating with elevated measures of sadness (0.621) and anger (0.594).

The data indicates heightened confusion markers (0.614) regarding educational trajectory maintenance amid financial instability, particularly in relation to the complex negotiation of competing demands across employment, academic pursuits, and familial obligations. However, the findings simultaneously reveal countervailing narratives of aspiration and achievement. These manifest in expressions of pride (0.627) associated with first-generation college attendance and the anticipated disruption of intergenerational poverty cycles.

5.2.3 Temporal Dynamics of Emotional Themes

To examine the temporal dynamics of the emotional landscape, we analyze five key emotions with relation to the topic of Youth Movements. We focus on pride, anxiety, anger, interest, and trust. Figure 4 illustrates the mean similarity score for each emotional theme per year and the distribution of documents related to youth movements over the years using bar charts.

Throughout the examined period, trust and interest emerge as dominant emotional markers, highlighting the aspirational and engaged nature of youth movements. In contrast, anger exhibits the lowest scores over the years, suggesting its limited role in the core experiences and contributions of these movements. The marked decline in emotional expression across all categories during 2020-2021 aligns with the imposition of COVID-19 restrictions and the subsequent shift to digital mobilization. Although a modest resurgence followed, emotional intensity remained subdued compared to pre-pandemic levels. This longitudinal analysis not only demonstrates the methodology's capacity to capture nuanced emotional dynamics that conventional topic-based approaches might overlook, but also situates the analyzed narratives within the context of macro-level societal events.

6 Conclusion

The exponential growth of large-scale textual data within the social sciences presents significant methodological challenges regarding the extraction of meaningful insights through empirically validated analytical frameworks. While contemporary computational methods offer an array of analytical tools, they remain insufficient for researchers seeking to derive nuanced interpretations that capture the full complexity of social phenomena. Current topic modeling approaches, for instance, predominantly emphasize content-based thematic classification while failing to account for the writer's sociocultural positionality, which is a critical dimension through which meaning is constructed and interpreted, since social locations and subjective standpoints can imbue ostensibly similar topics with distinctly different meanings.

Addressing this methodological gap, this paper proposes a comprehensive framework for identifying and analyzing emotional themes within topic structures, while simultaneously addressing the complexities inherent in textual analysis. By integrating explicit knowledge from official lexicons with latent semantic representations, our approach evaluates the thematic composition of a document based on semantic similarity. The empirical findings reveal intricate thematic structures that have been largely overlooked in prior research. Furthermore, our method is language-agnostic and adaptable to any official lexicon, as evidenced by its successful application in both Hebrew, a morphologically rich language, and English.

Future research can expand our method to include not only granular analysis of sociocultural themes but also integrate intersectional dimensions of social stratification—including gender, race, and class—to more comprehensively theorize and empirically examine the situated nature of human-generated text.

7 Limitations and Ethical Considerations

This study presents a promising method for emotional theme detection; however, it is not without its limitations.

First, due to privacy concerns, we cannot make the financial aid dataset publicly available or provide specific examples from it. While this ensures compliance with ethical and legal obligations, it limits the reproducibility of our study and the ability of other researchers to analyze the method in the context of this dataset. To address this, we also present results on the English dataset of Kaggle student profiles. However, this dataset has its own limitations, as the texts are synthetic and not written by real people. Consequently, the emotional valence expressed in these texts is lower than what we would expect in real-world data and what we observe in the financial aid texts, potentially limiting the generalizability of our findings.

Second, the quantitative insights generated through this methodological framework demand interpretation through disciplinary expertise. Content specialists can illuminate how writers' narrative strategies and emotional repertoires reflect broader cultural, historical, and social forces within their respective fields..

Finally, while the LIWC lexicon serves as a powerful foundation for this study, its full English version remains a closed corpus. This restricts access to the complete set of words associated with each category, potentially limiting the accuracy of category definitions in our study. Free access to the complete LIWC corpus would enable more precise and comprehensive category definitions by including the wide variety of words used to describe each thematic dimension.

Despite these limitations, our method provides a solid foundation for future research, and we hope it inspires further advancements in emotional theme detection and interdisciplinary collaboration.

References

AJ Alvero. 2023. Sociolinguistic perspectives on machine learning with text data. In *The Oxford Handbook of the Sociology of Machine Learning*. Oxford University Press.

Dimo Angelov. 2020. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Jonah Berger and Grant Packard. 2022. Using natural language processing to understand people and culture. *American Psychologist*, 77(4):525.

David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Pierre Bourdieu. 1993. *Language and symbolic power*. Harvard University Press.

Biraj Dahal, Sathish AP Kumar, and Zhenlong Li. 2019. Topic modeling and sentiment analysis of global climate change tweets. *Social network analysis and mining*, 9:1–20.

Teun A. van Dijk. 2009. *Society and Discourse: How Social Contexts Influence Text and Talk*. Cambridge University Press.

Paul Ekman, Wallace V Friesen, Maureen O'sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, et al. 1987. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4):712.

Roberto Franzosi, Wenqin Dong, and Yilin Dong. 2022. Qualitative and quantitative research in the humanities and social sciences: how natural language processing (nlp) can help. *Quality & Quantity*, 56(4):2751–2781.

Klaifer Garcia and Lilian Berton. 2021. Topic detection and sentiment analysis in twitter content related to covid-19 from brazil and the usa. *Applied soft computing*, 101:107057.

Thomas Griffiths, Michael Jordan, Joshua Tenenbaum, and David Blei. 2003. Hierarchical topic models and the nested chinese restaurant process. *Advances in neural information processing systems*, 16.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Joshua Conrad Jackson, Joseph Watts, Teague R Henry, Johann-Mattis List, Robert Forkel, Peter J Mucha, Simon J Greenhill, Russell D Gray, and Kristen A Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. *Science*, 366(6472):1517–1522.

Theodore D Kemper. 2006. Power and status and the power-status theory of emotions. In *Handbook of the sociology of emotions*, pages 87–113. Springer.

Jochen Kleres. 2011. Emotions and narrative analysis: A methodological approach. *Journal for the Theory of Social Behaviour*, 41(2):182–202.

Hengyun Li, XB Bruce, Gang Li, and Huicai Gao. 2023. Restaurant survival prediction using customer-generated content: An aspect-based sentiment analysis of online reviews. *Tourism Management*, 96:104707.

Bin Liang, Hang Su, Lin Gui, Erik Cambria, and Ruifeng Xu. 2022. Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. *Knowledge-Based Systems*, 235:107643.

Chenghua Lin and Yulan He. 2009. Joint sentiment/-topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 375–384.

Yongchao Ma, Ying Teng, Zhongzhun Deng, Li Liu, and Yi Zhang. 2023. Does writing style affect gender differences in the research performance of articles?: An empirical study of bert-based textual sentiment analysis. *Scientometrics*, 128(4):2105–2143.

Hazel R. Markus and Shinobu Kitayama. 1991. Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, 98(2):224–253.

Jon Mcauliffe and David Blei. 2007. Supervised topic models. *Advances in neural information processing systems*, 20.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*, pages 171–180.

Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2019. Weakly-supervised hierarchical text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6826–6833.

James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2022. *LIWC2022: Linguistic Inquiry and Word Count*. Pennebaker Conglomerates, Inc., Austin, TX. Software and dictionary available at https://liwc.app/.

Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.

Weihong Qi, Jinsheng Pan, Hanjia Lyu, and Jiebo Luo. 2024. Excitements and concerns in the post-chatgpt era: Deciphering public perception of ai through social media analysis. *Telematics and Informatics*, page 102158.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. 2009. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 248–256.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Emelie Rosenberg, Carlota Tarazona, Fermín Mallor, Hamidreza Eivazi, David Pastor-Escuredo, Francesco Fuso-Nerini, and Ricardo Vinuesa. 2023. Sentiment analysis on twitter data towards climate action. *Results in Engineering*, 19:101287.

Natalie Shapira, Dana Atzil-Slonim, Daniel Juravski, Moran Baruch, Dana Stolowicz-Melman, Adar Paz, Tal Alfi-Yogev, Roy Azoulay, Adi Singer, Maayan Revivo, et al. 2021. Hebrew psychological lexicons. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 55–69.

JAN E. Stets and JONATHAN H. Turner. 2010. Handbook of emotions. In *The Oxford Handbook of the Sociology of Machine Learning*. Guilford Press.

Donglei Tang, Zhikai Zhang, Yulan He, Chao Lin, and Deyu Zhou. 2019. Hidden topic–emotion transition model for multi-level social emotion detection. *Knowledge-Based Systems*, 164:426–435.

Yee Teh, Michael Jordan, Matthew Beal, and David Blei. 2004. Sharing clusters among related groups: Hierarchical dirichlet processes. *Advances in neural information processing systems*, 17.

Bo Wang, Maria Liakata, Adam Tsakalidis, Agelos Georgakopoulos, Orestis Papadopoulos, Lazaros Apostolidis, Arkaitz Zubiaga, Rob Procter, and Yiannis Kompatsiaris. 2017. Totemss: Topic-based, temporal sentiment summarisation for twitter. *Proceedings of the IJCNLP 2017*, pages 21–24.

Hui Yin, Xiangyu Song, Shuiqiao Yang, and Jianxin Li. 2022. Sentiment analysis and topic modeling for covid-19 vaccine discussions. *World Wide Web*, 25(3):1067–1083.

Zhouqing Zhang, Kongmeng Liew, Roeline Kuijer, Wan Jou She, Shuntaro Yada, Shoko Wakamiya, Eiji Aramaki, et al. 2024. Differing content and language based on poster-patient relationships on the chinese social media platform weibo: Text classification, sentiment analysis, and topic modeling of posts on breast cancer. *JMIR cancer*, 10(1):e51332.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. *arXiv preprint arXiv:2106.01071*.

Moti Zwilling. 2023. Big data challenges in social sciences: an nlp analysis. *Journal of Computer Information Systems*, 63(3):537–554.

A Example Appendix

Examples from the Kaggle synthetic student profile dataset:

Box A.1: Document 546

Bruce Keller was a sophomore at the University of South Carolina, majoring in Earth Science. He had always been fascinated by the natural world and was determined to understand it better. With a GPA of 3.4, he was a dedicated student, always striving for excellence in his studies. But Bruce wasn't just a bookworm. He had a creative side that he loved to explore. His hobbies included writing and painting, two activities that allowed him to express his thoughts and emotions in a different way. But what truly set him apart was his love for photography. Bruce was a photography aficionado, always carrying his camera with him wherever he went. One day, while walking around the campus, Bruce stumbled upon a beautiful garden filled with colorful flowers and lush green trees. Without hesitation, he pulled out his camera and started taking pictures. He was completely lost in the moment, capturing the beauty of nature through his lens. As he was taking pictures, he noticed a group of students huddled around a bulletin board. Curiosity got the better of him, and he walked over to see what was going on. To his surprise, it was a call for submissions to a photography contest organized by the university. Bruce's heart skipped a beat. It was the perfect opportunity for him to showcase his passion and talent for photography. Without wasting any time, Bruce gathered his best shots and submitted them to the contest. He was nervous but hopeful. To his delight, a few weeks later, he received an email congratulating him on winning first place in the contest. His photo, "The Enchanting Garden," had captured the attention of the judges and had been chosen as the winner. Bruce was over the moon. Not only had he won the contest, but he had also found a way to combine his love for Earth Science and photography. From that day on, he became known as the "Photography Guru" among his friends and classmates. But Bruce remained humble and continued to pursue his studies with the same dedication and passion. He even started a photography club on campus, where he shared his knowledge and skills with others who shared his love for capturing the world through a lens. Years later, as a successful Earth Science researcher, Bruce looked back on his college days with fondness. He was grateful for the opportunities he had been given and the friends he had made. But most of all, he was thankful for discovering his unique quality as a photography aficionado, which had opened up a world of possibilities for him.

Box A.2: Document 16140

Allison Osborne was a sophomore at the University of North Carolina, majoring in Sociology with a GPA of 3.18. She was a bright and driven student, determined to make a difference in the world through her studies. But there was more to Allison than just her academic achievements. Allison was a health and wellness advocate, always promoting a balanced and mindful lifestyle. She believed that taking care of one's physical and mental wellbeing was crucial for success and happiness. Her passion for health and wellness stemmed from her own struggles with anxiety and depression, which she had overcome through exercise, meditation, and a healthy diet. When she wasn't studying or attending classes, Allison could be found jogging around campus. She loved the feeling of the wind in her hair and the rush of endorphins that came with each run. It was her way of relieving stress and staying physically fit. But jogging wasn't her only hobby. Allison was also a talented tarot reader. She had inherited the skill from her grandmother and had been practicing it since she was a teenager. She found solace in the cards and enjoyed helping others gain insight into their lives through tarot readings. One day, while jogging around campus, Allison ran into her friend, Lily, who was struggling with a difficult decision. Lily was torn between two job offers, and she didn't know which one to choose. Allison could sense her friend's anxiety and offered to do a tarot reading for her. As she laid out the cards, Allison explained the meaning behind each one and how they related to Lily's situation. The reading revealed that one job offered more financial stability, while the other allowed for more personal growth and fulfillment. After much contemplation, Lily decided to take the job that aligned with her personal goals and aspirations. Impressed by Allison's tarot reading, Lily suggested that she offer her services to other students on campus. Allison took her friend's advice and started offering tarot readings to her peers. Word quickly spread about her accurate readings and soon, Allison's schedule was filled with appointments. As her reputation as a tarot reader grew, so did her passion for helping others. She found joy in using her unique skill to guide and inspire her peers. It was a fulfilling and meaningful way for her to share her passion for health and wellness. Allison's journey as a college student was not just about academics, but also about discovering her unique qualities and using them to make a positive impact on others. She had found her passion in tarot reading and her purpose in promoting health and wellness, and she was determined to continue spreading positivity and light wherever she went.

Box A.3: Document 3676

William Thompson was a freshman at the University of Saskatchewan, majoring in Environmental Science. He was a tall and lean young man with a passion for fitness and callisthenics. Growing up in the small town of Saskatchewan, William was always fascinated by the beauty of nature and how it could be preserved for future generations. Despite his love for the environment, William's grades were not the best. He had a GPA of 2.81, which constantly worried him. He knew he had to work harder to achieve his dream of becoming an environmentalist. In his free time, William would hit the gym and train rigorously. He was determined to maintain a healthy and fit lifestyle. His peers were amazed by his strength and agility, and some even asked him to train them. William was happy to share his knowledge and passion for fitness with others. But what set William apart from others was his remarkable dancing skills. He had been taking dance classes since he was a child and was known for his smooth and graceful moves. His friends often joked that he could dance his way out of any problem. One day, while walking to his Environmental Science class, William heard loud music coming from the auditorium. Curiosity got the better of him, and he decided to check it out. To his surprise, there was a dance competition happening, and the winner would get a scholarship for their college fees. Without hesitation, William signed up for the competition. His friends were shocked and asked him why he was taking part when he had a low GPA. William simply replied, "I have to give it a shot. It’s about saving the environment, and I believe my dancing can make a difference." The day of the competition arrived, and William gave a stellar performance. His dance was a fusion of modern and traditional moves, and the audience was captivated. When the results were announced, William was declared the winner, and he received a scholarship for his college fees. From that day on, William’s confidence grew, and he worked even harder to improve his grades. He became a role model for his peers as he balanced his love for the environment, fitness, and dancing while excelling in his studies. Years later, William graduated with flying colors and landed his dream job as an environmentalist. He often looks back at his college days and remembers how his unique qualities helped him achieve his goals. He continues to dance, inspiring others to do their part in preserving the environment.

Box A.4: Document 8366

Keith Steele was a sophomore at the University of California, Berkeley, majoring in Biochemistry. He was a hardworking student with a GPA of 3.75 and had always been passionate about science and research. Keith was originally from California, USA, and had always dreamed of attending one of the top universities in the state. Aside from his academic pursuits, Keith had a few hobbies that he enjoyed in his free time. He was an avid board game enthusiast and loved spending hours strategizing and playing with his friends. He also had a keen interest in photography and would often go on long walks around the campus, capturing the beauty of the surrounding nature through his lens. But what made Keith stand out from his peers was his unique quality “he was a certified yoga instructor. Keith had been practicing yoga since high school and had even completed a yoga teacher training course during his gap year. He found solace in the practice and believed that it not only improved his physical health but also his mental well-being. Keith was known as the go-to yoga instructor among his friends and classmates. He would often hold free yoga sessions on the weekends, and many students would attend to distress from their busy academic lives. His calming voice and expert guidance made him a popular instructor, and his classes were always full. One day, as Keith was walking around campus, he stumbled upon an old abandoned building. Curiosity got the better of him, and he decided to explore it. As he entered the building, he noticed that the walls were covered with graffiti, and the windows were shattered. But what caught his eye was a large open space in the center of the building. It was the perfect spot for a yoga studio. Keith shared his idea with his friends, and they were all on board. Together, they cleaned up the space and turned it into a makeshift yoga studio. Keith started holding regular classes there, and soon enough, he had a loyal following of students. Word of mouth spread, and soon, even students from other universities would come to attend Keith’s classes. He was thrilled to see how his passion for yoga had brought people together and helped them find inner peace. Keith’s unique quality had not only made a positive impact on his own life but also on the lives of those around him. He had found his calling and was determined to continue spreading the benefits of yoga to as many people as possible. And as he continued to excel in his academics, Keith was grateful for the balance that his hobbies and unique quality brought to his life.

B Hebrew LIWC

The full list of categories and keywords in Hebrew LIWC appear in https://github.com/natalieShapira/HebrewPsychologicalLexicons/blob/master/src/hepsylex/Lexicons.py.

From that list, we chose the following 42 categories: Amused, Anger, Anticipation, Anxiety, Ashamed, Calm, Confusion, Contentment, Disgust, Enthusiastic, Fatigue, Guilty, Hostile, Interest, Joy, Nervous, Not Amused, Not Anger, Not Anticipation, Not Anxiety, Not Ashamed, Not Calm, Not Confusion, Not Contentment, Not Disgust, Not Enthusiastic, Not Fatigue, Not Guilty, Not Hostile, Not Interested, Not Joy, Not Nervous, Not Proud, Not Sad, Not Surprise, Not Trust, Not Vigor, Proud, Sad, Surprise, Trust, Vigor.