Leveraging Social Media Analytics in Engineering Education Research

Ms. Sakshi Solanki, Utah State University

Sakshi Solanki is a PhD student in the Engineering Education department at Utah State University. She earned a bachelor's degree in Electrical and Electronic Engineering from ITS Engineering College, India and a master's degree in Data Science from University at Albany, New York. She worked as a Data Analyst during one of her summer internships in 2020, where she learned and gained experienced in data evaluating and validating company's huge data using the techniques based on Excel, Python, and R. She is currently working with Dr. Marissa Tsugawa on Neurodiversity Research and Education. She believes that neurodiversity can help her better understand her younger brother's condition (Asphyxiation) and respond to his basic needs because his mind works differently from everybody else's due to which he unable to express his feelings and pain.

kiana kheiri

Dr. Marissa A Tsugawa, Utah State University

Marissa Tsugawa is an assistant professor at Utah State University focusing on neurodiversity and identity and motivation. She completed her Ph.D. in Engineering Education focusing on motivation and identity for engineering graduate students.

Hamid Karimi, Utah State University

I completed my Ph.D. in Computer Science at Michigan State University (MSU) in 2021, with my primary research focus on artificial intelligence (AI) for social good. During my doctoral studies, I explored several intriguing areas, such as AI in education, computational politics, and misinformation detection. As a member of the interdisciplinary Teachers in Social Media project, I concentrated on creating innovative and efficient data mining and machine learning algorithms to enhance the quality of PK-12 education. Throughout my academic journey, I have been honored with multiple awards. These include the Best Paper Award at the IEEE-ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2018), the Outstanding Graduate Student Service Award (2019), the Dissertation Completion Award (2020), and the International Faculty Recognition Award at Utah State University (2022). In August 2021, I joined Utah State University as an Assistant Professor (tenure-track) in Computer Science, where I now lead the Data Science and Applications lab (dsa.cs.usu.edu).

Leveraging Social Media in Engineering Education Research: Latent Dirichlet Allocation Method

Abstract

In our work, we explore how social media analytics can be leveraged in engineering education research to understand lived experiences of marginalized groups outside of engineering contexts to inform research in engineering contexts. Specifically, our work explores the video-based, social media platform TikTok using latent Dirichlet allocation (LDA; a topic modeling method) to elucidate neurodivergent topics. We applied LDA to transcripts of neurodivergent TikToks and developed a four topic model to describe the text-based data. Social media is an informative environment which can be leveraged in this field with careful application of topic modeling methods. Further, topic modeling methods can be used for any large, text-based data to supplement qualitative analysis methods.

Keywords: Social Media Analysis, Latent Dirichlet Allocation, Neurodivergent

1. Introduction

The purpose of this research paper is to explore how social media can be leveraged in engineering education research and provide a step-by-step method for social media analytics. People around the world use social media platforms (e.g., Facebook, Reddit, SnapChat, TikTok, and Twitter) to share content that express their personal and professional identities and connect with others like them [1]–[4]. Social media is a public space full of rich information and conversations that can show how and who people interact with and what people publicly share about themselves. Particularly, social media has served as a platform for marginalized communities to connect, organize and collaborate, disseminate information, and negotiate their identities [5]–[11]. Social media is a rich and vast source of information that engineering education researchers can leverage to understand and integrate lived experiences of marginalized communities in larger social contexts outside of academia. Outside of engineering education, social media analytical tools such as social network analysis and natural language processing has been used to understand these social contexts. Our work presents the latent Dirichlet allocation which is a type of topic modeling method that uses machine learning and natural language processing to analyze a large amount of textual data.

2. Background

2.1. Social Media in Research

People marginalized by societal structures and institutions based on factors such as race/ethnicity, gender identity, disability, and more have found solace in digital spaces [12]. Particularly, social media platforms such as Facebook, Twitter, TikTok, and Instagram have offered a space for marginalized people to “expose the truth and drive social change” [5, p. 158]. This space is vital to marginalized people especially when traditional media (e.g., televised media) has downplayed grievances and misrepresented perspectives [12]. Further, marginalized groups have also used social media to socially construct their identities, create shared meaning,

and build communities [2], [5], [7], [9], [10], [13]. Because digital spaces such as social media has become an emancipatory platform for marginalized peoples, researchers have an opportunity to listen to and learn from the underrepresented and unheard voices. Content publicly available on social media provides a myriad of evidence for lived experiences around social identities (e.g., queer identity) that cannot be captured in a clinical or research settings. When used ethically, marginalized peoples and their lived experiences can be properly represented in research. Particularly, researchers can address the limitations of consent by taking necessary precautions to protect the privacy and autonomy of participants [14]–[16].

Researchers have only scratched the surface of the plethora of content publicly available on social media to listen to and understand marginalized groups. Researchers outside of engineering education have analyzed content posted on social media platforms to understand topics such as identity negotiation and development (professional [17], [18] and social [7], [8], [10], [19]), politics [20]–[22], (mis)information dissemination [23]–[26], and health behaviors (e.g., coping with eating disorders [27] and mental health [28]). Using social media to research different social identities (e.g., queer and racial identities) has shown that social media serves as a space for individuals to learn more about themselves, connect with similar people, and build communities [7]. This type of information can be used to inform engineering education research and pedagogical practice.

2.2. Social Media Research in Engineering Education

In engineering education research, social media platforms are underutilized data sources to understand marginalized people and their experiences. Only two studies within engineering education have used social media to understand engineering student experiences. Berdanier and colleagues generated narratives around graduate engineering student attrition from the social media platform Reddit [29]. Another study, by Chen and Gillen, also analyzed forums on Reddit to understand engineering students' learning experiences before and during COVID-19 [30]. Both studies used the application programming interface (API) from Reddit to collect posts from the Reddit forums. The studies differed in analysis where Berdanier et al. conducted qualitative narrative analysis while Chen and Gillen used quantitative machine learning, specifically natural language processing tools. Chen and Gillen's study is the first in engineering education to use machine learning as a tool to analyze qualitative data. Both studies resulted in themes that represent student experiences in engineering demonstrating the utility and novelty of social media platforms as a data source.

Engineering education researchers can also leverage social media platforms to understand engineering student experiences who experience systemic discrimination and oppression. For example, in our larger research project focusing on neurodivergent (e.g., ADHD, autism, dyslexia, anxiety) engineering students, published knowledge on what it means to be neurodivergent is limited to deficit framing and language developed by researchers [31]. However, a plethora of knowledge and first-hand accounts of neurodivergent experiences exist on social media where emancipatory, as opposed to deficit, language is used to describe their lived experiences [13], [23], [32], [33]. Researchers should immerse themselves in these online communities to join their conversations and to understand the experience of being neurodivergent. Our work presented here and elsewhere [paper under review and other ASEE

2023 submission] specifically investigates the video-based and popular social media platform, TikTok.

3. Methodological Framework: Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a popular unsupervised topic modeling method used in natural language processing (NLP). Before detailing the method in the subsequent subsections, we provided definitions of the terminology used to clarify the use of the words (see Table 1).

Table 1. Definitions of terminology used in LDA, topic modeling, and NLP [34]–[36]. Indented terms fall within the non-indented term above them.

Term	Definition
Machine Learning (ML) [36]	Coding methods that enable computers to "learn" without programming everything.
Unsupervised Learning	The code attempts to develop a model that the data fits to; no specified outcome is identified by the researcher.
Supervised Learning	A model is provided for the code to fit the data to; a specific outcome is identified by the researcher.
Training	A set of data is provided for the code to learn or develop a model.
Natural Language Processing (NPL) [37]	A research field that uses machine learning to "obtain meaning from human language processing in text-documents" [37, p. 15170]
Topic Modeling [34]	"A statistical tool for extracting latent variables form large datasets" [34, p. 1].
Word	A unit of individual data
Document	A "string" containing *n* words; a collection of words.
Corpus	A set of *m* documents; typically, the whole dataset.
Vocabulary	A collection of distinct words also known as the dictionary.
Topic	Individual topics represent the probability distribution of vocabulary (e.g., vocabulary clusters).
Latent Dirichlet Allocation (LDA) [34], [35]	This unsupervised NPL method using topic modeling by organizing data at three levels (word, topic, and document) to generate a probabilistic model of hidden topics. This method does not rely on external knowledge bases allowing meaning to emerge from the dataset only.

3.1. Topic Modeling

Topic modeling is a powerful statistical tool used in data analytics to determine common characteristics of data points within large datasets and is particularly well-suited for analyzing textual data (the corpus) [34], [38]. Particularly, topic modeling can find hidden structures, or topical patterns, of large textual datasets to aid researchers extract meaningful information [38]. The outcome of topic modeling is a set of topics where each topic generated contains a collection of words with high probability of appearing together in the same context. Topic modeling serves a variety of purposes, one of which is to allow users to dig deeper into documents and develop links between previously unrelated topics. Topic modeling has been widely used from analyzing bioinformatics data to social data to environmental data [34], [39].

3.2. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a popular unsupervised topic modeling method that has been used in fields such as business, medical, and road traffic [40]–[42]. LDA is similar to cluster analysis where LDA is used to identify latent topics while cluster analysis is used to group items based on their observable similarities [43]. Specifically, LDA identifies hidden patterns, or topics, in the data by generating a probabilistic model based on the probability of words appearing in documents of a large corpus [37]. The model assumes that words in a document are related so the probabilities of words appearing in documents can be calculated. Further, individual words are assigned to topics based on the probability of similar words appearing together in a document. Two model parameters (the Dirichlet parameters) indicate how many topics are represented in each document ($0 < \alpha < 1$) and how many words are represented in each topic ($0 < \beta < 1$) where a high α value makes the document appear more similar to one another and high β value makes topics appear more similar to one another. For our study, we chose lower α and β values because the documents are relatively small and we want to look for nuances across the corpus for neurodivergence as the main topics discussed in the data are autism, ADHD, and neurodivergence. Figure 1 demonstrates how the LDA model creates a probabilistic model of the topics from documents.

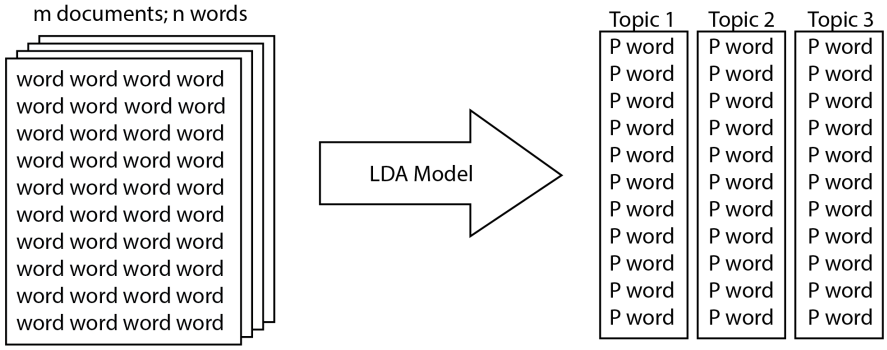

Fig. 1. The corpus contains m documents with n words (left). An LDA model (center) generates the probability, P, of words being in a topic (right).

For the model to assign words to topics, the researcher must input the number of topics as a parameter [44]. The researchers might not know the number of topics in the corpus, so the researcher can test a range of numbers of topics in the model then calculate a coherence score for each number of topics. The coherence score is based on the likelihood of words being related to one another in a topic and is calculated by word pairs and word pair probabilities [44]. The number of topics can then be chosen by looking at a coherence score vs number of topics plot where the optimal number of topics is at an inflection point in the graph while also considering the researchers prior knowledge on the corpus [45]. After the model is generated, the researchers then explore the words in each topic to determine what each topic represents. Overall, the LDA method is like qualitative content analysis and qualitative coding methods.

4. Positionality

Before delving into this study, we provide our positionality statement here as our positions influence how we conceptualize the theory and methodological choices [46], [47]. We provide

an overview of the research team's reflexivity for this study. Although the research team is diverse in many ways, we share a common goal of highlighting the voices of neurodivergent and other marginalized peoples in our research. All authors have backgrounds in engineering of different disciplines. A majority of the research team are neurodivergent or disabled and leveraged their lived expertise in this research. The team is also diverse in race/ethnicity, gender, and country of origin which also provided diverse perspectives in approaching this research.

5. Method

To demonstrate the LDA method in engineering education research, we conducted small-scale social media analysis on neurodivergent content shared on TikTok. The main goal of our study was to identify underlying topics related to neurodivergence that help describe neurodivergent lived experiences. Overall, this research design involved manually downloading content from TikTok; transcribing, cleaning and preprocessing the data; and training the LDA model. Figure 2 provides an overview of the design process.

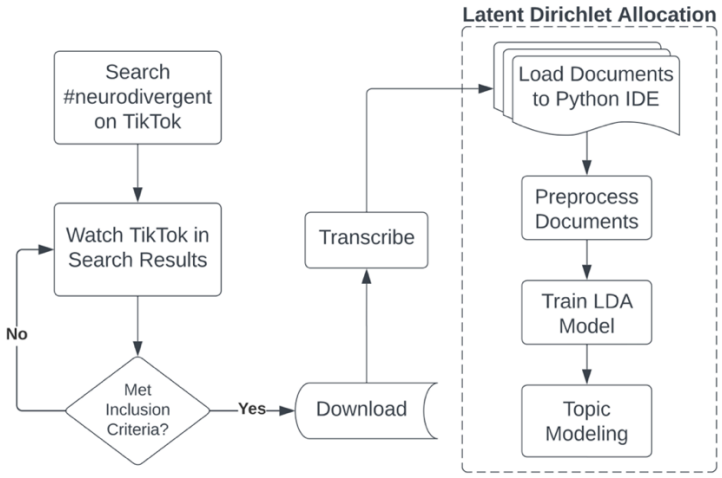

Fig. 2. Overview of the research design.

5.1. Research Context – TikTok

TikTok (https://tiktok.com/) is a popular video-based, social media platform [48] that gained popularity during the COVID-19 pandemic [48]. TikTok continues to be one of the most downloaded apps on both the Apple App Store and Google Play (as of February 2023) with approximately 85% of its users below 35 years old [49]. The TikTok algorithm personalizes users' For You Page based on the users' activity which results in the emergence of niche TikTok communities [49], [50]. Content Creators (TikTokers) can quickly create and edit a few second- to three-minute-long video content (TikToks) using smartphones to share lived experiences, communicate with and respond to other users, and build community. TikToks may fit into a number of different genres including acting, animated infographic, documentary, news, oral speech, pictural slideshow, and TikTok dance [51]. Other users interact with the TikTokers' posts by liking, sharing, bookmarking, or commenting. Further, TikTokers can interact with their audiences' comments through video responses to seamlessly continue conversations with commenters. TikTokers can also "stitch" or "duet" other TikTokers' videos to respond or interact

to other TikTokers' content. This interactive approach encourages community interaction and engagement and is unique to TikTok [52].

5.2. Data collection

TikTok operates with a recommendation algorithm meaning it learns from a user's TikTok activity such as likes, shares, and follows [53]. To avoid personal recommendation algorithms affecting the data collection, we collected TikToks solely through the search feature of the app. We searched the Top TikTok Page using the hashtag #neurodivergent first, then #depression and #anxiety which are common traits of being neurodivergent [54], [55]. Future work will explore the intersection of neurodivergent and engineering as this work was intended to share how to use the LDA method in engineering education research. We then downloaded TikToks after watching, liking and bookmarking, and determining if each met our inclusion criteria where TikToks must be available to the public (public TikToks are downloadable while private TikToks are not); an informational, narrative genre (excludes TikTok dances and jokes); and educational such that they describe lived experiences, discuss scenarios or hypotheticals, or spread awareness of social interactions.

In total, we downloaded 100 TikToks from the #neurodivergent search, 50 from the #depression search, and 50 from the #anxiety search. We collected a total of 200 TikToks where each TikTok ranged from nine seconds to five and a half minutes long and totaled two hours and 25 minutes' worth of TikTok content. We dropped four TikToks after re-reviewing the TikToks and determined they did not meet the inclusion criteria with a final number of 196 TikToks. Basic TikTok information such as file name, TikTok length in seconds, hashtags, video descriptions, and dates were documented in a spreadsheet for organization. We did not summarize the demographic data of the TikTokers as they do not report such information in a standard form and we chose not to assume the demographics. However, we can analyze creator demographics in later studies that use data mining techniques. Further, we chose to manually collect TikToks and not mine the data so we can compare the LDA method to a qualitative thematic analysis (paper under preparation).

5.3. Data Handling and Preprocessing

After the TikToks were downloaded to a secure folder, the research team transcribed the audio and any added text (e.g., captions or additional text added by the creators). Transcriptions were checked for accuracy and followed a standardized transcription system that accounts for visual aspects of the TikToks. Specifically, the transcription system distinguishes what was verbally said from text appearing on TikToks, closed caption edits to what was said, and any contextual information. Each transcript represents a document of words for the LDA analysis. The documents (transcripts) were then loaded to a Jupyter Notebook (Python) to clean and preprocess for the LDA analysis.

Two key inputs are needed to train an LDA model: the corpus (collection of documents) and the dictionary. To create a corpus and dictionary, the documents need to be cleaned then preprocessed (demonstrated in Table 2). Cleaning the documents involved expanding contraction words; removing any punctuation (e.g., ‘, ., ?), symbols, (e.g., #, @, %), emojis (e.g., 😁, 😎, ✨),

and accented letters (e.g., â, ñ); and converting capital letters to lower case letters to facilitate more accurate analysis [56]. Next, preprocessing the documents consists of three main processes: tokenizing, stopping, and stemming. Tokenizing the documents converts the documents to their atomic elements [56], which identifies the words as words in each document. Stopping removes any words in the documents that do not provide significant meaning to a sentence (stopwords) such as 'the,' 'this,' 'that,' and 'it' are removed from the documents. Common English stopwords are defined in the natural language processing toolkit and the researchers can add other stopwords [57]. Our additional stopwords included words that were common on TikTok that had no meaning like "part" and "story" and "TikTok." Finally, stemming is the process of converting words of similar meaning to their stem word (e.g., "changing" and "changed" have a stem of "chang"). However, we used the lemmatizing process which similarly finds the stem of a word but uses the dictionary word rather than the stem (e.g., "change" rather than "chang").

Table 2. Demonstration of the text being preprocessed.

Process	Data
Original Data	`So you're breaking up with me because I'm too… Blond? [So you're breaking up with me because I'm too…✨Neurodivergent✨]`
Expand Contractions	`So you are breaking up with me because I am too… Blond? [So you are breaking up with me because I am too…✨Neurodivergent✨]`
Remove Punctuation and Symbols	`So you are breaking up with me because I am too Blond So you are breaking up with me because I am too Neurodivergent`
Convert to Lower Case	`so you are breaking up with me because i am too blond so you are breaking up with me because i am too neurodivergent`
Remove Stopwords	`breaking blond breaking neurodivergent`
Lemmatize	`break blond break neurodivergent`
Tokenize	`WordList(['break', 'blond', 'break', 'neurodivergent'])`

5.4.3. Describing the Corpus

Prior to training the LDA model, we first explored the contents of the corpus. This step is analogous to descriptive statistics describing the sample before conducting analysis with the variables (e.g., linear regression). Describing textual data provides an overview of the most frequent words that appear in the corpus which can be visually represented in plots such as word clouds [58] of frequency plots. By highlighting the most frequent words in the corpus, we can identify the most relevant and important terms used in the documents. For our corpus, these words give us insight to the words used by neurodivergent people to describe their lived experiences.

5.4.4. Training and Interpreting the LDA Model

A predetermined number of topics from the corpus must be provided to train an LDA model which can be found by calculating the coherence score from testing the LDA model over a range of topics. For our study, we tested the LDA model from two to twelve topics. The coherence score was then plotted by the number of topics where the optimal number of topics is identified

at the first inflection point on the plot. After the number of topics was chosen, we then trained the LDA model with our corpus using low Dirichlet parameters ($\alpha = 0.01$ and $\beta = 0.01$). We then used a popular visualization tool, *PyLDAvis*, to interactively visualize the results of the LDA model. The *PyLDAvis* package allows the researchers to interactive with the topics for visualizing the variety of topics generated by the LDA model [59]. One of the PyLDAvis's key features is the intertopic distance map which displays the topics in a 2D space. As such, this tool allowed us to see the words that make up each topic and see how unique each topic is where topics that overlap are less unique and topics further apart of more unique. To understand and interpret individual topics, each topic be manually chosen to examine its most frequent or relevant terms using λ parameter values and give each topic a label or "meaning" that can be understood by people. Values closer to 1 indicate more often recurring terms in the document that might not be related to the topic, while values closer to 0 indicate more exclusive terms for the selected topic [60].

6. Results

6.1. Description of the Neurodivergent TikTok Corpus

In total, the corpus consisted of 6,156 words (after preprocessing) with 2,392 unique terms. We generated a word cloud of the corpus as shown in Fig. 3 displays the frequency of the most frequent 30 words in the corpus as both a word cloud (left) and frequency plot (right). The top three most frequent words in the corpus were ADHD (count is 151), autistic (count is 89), and feel (count is 55). This result demonstrated that a majority of the neurodivergent conversation on TikTok is about ADHD and autism. Common words used to describe neurodivergent lived experiences include "forget," "different," "mask," "understand," "trait," "function," "trauma," and "anxiety." These terms demonstrate how autism and ADHD can overlap. Although "symptom" was a common word used, it was less common than the words "different" and "trait." This finding supports the shift away from pathologized thinking of neurodivergent people to accepting neurodivergent people for who they are and that part of them is not curable.

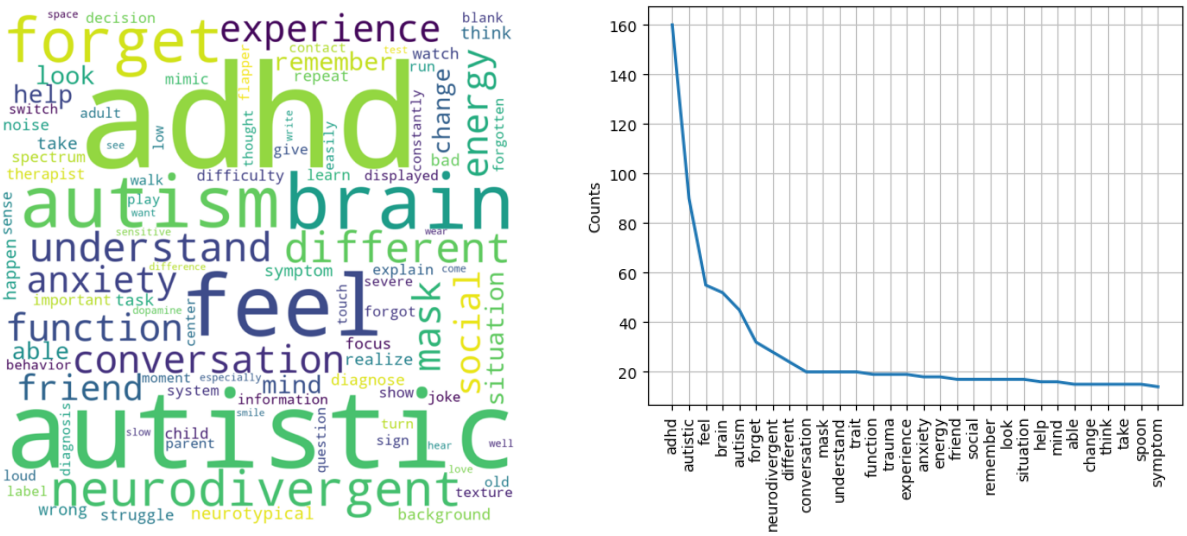

Fig. 3. Word cloud visualization of the corpus.

6.2. LDA Model for Neurodivergent TikTok

An optimal number of topics within the corpus must be determined as an input to the LDA model, which can be identified through coherence scores and researcher knowledge on the topic. Coherence scores range from 0 to 1 with higher scores indicating more meaningful and consistent topics [61], [62]. We calculated the coherence score for the number of topics ranging from two to twelve as shown in Fig. 4. We chose four topics for this study (coherence score of 0.43) as it represents the first peak in Fig. 4 and serves as a simple model for this paper. This score suggests that the four topics generated by the model had a moderate level of coherence. We then trained the LDA model using four topics and Dirichlet parameters of $\alpha = 0.01$ and $\beta = 0.01$.

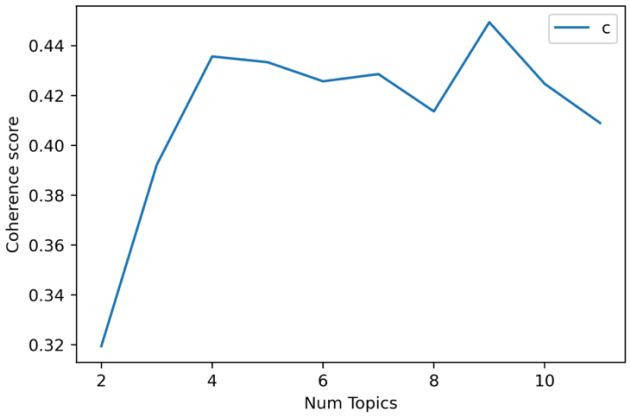

Fig. 4. Coherence score versus number of topics.

To explore the topics and the distribution of words within these four topics, we generated word cloud images for each topic. Figure 5 shows the top 15 words for each topic as a word cloud where the larger words represent more frequency of that word for that topic. Within each topic, there is overlap of the words "ADHD," "autistic," and "autism." We considered the two terms "autistic" and "autism" as separate and unique because both have different cultural meanings within the autistic community. The term "autistic" is preferred over "person with autism" because it is seen as an integral part of their identity [63]–[65]. Further, in listening to the data audio files, the autistic community tended to use "autistic" or "neurodivergent" when talking about their identities and "autism" when talking about the condition as a larger concept.

We then plotted the word counts and their weights for each topic as shown in Fig. 6. For Topic 1, the most frequent words are "ADHD," "brain," and "forgotten," and this topic may represent ADHD experiences with memory. Topic 2's top three words are "autistic," "forget," and "ADHD" which overlap with Topic 1. However, the other most frequent words included in Topic 2 may represent the more nuanced experience of undiagnosed autistic women. Topic 3's most important words are "autistic," "autism," and "clutter" and may represent the autistic experience. Topic 4 included "spoon," "ADHD," and "wave" as its top three words and may represent ADHD experiences again in relation to "spoons." Spoons stem from spoon theory which is used by neurodivergent people to describe how much energy and the type of energy they have or do not have for completing tasks [66]. This result is interesting as autistic people tend to leverage spoon theory more than other neurodivergent people to describe their energy levels. However, it is widely used across neurodivergent and disabled people.

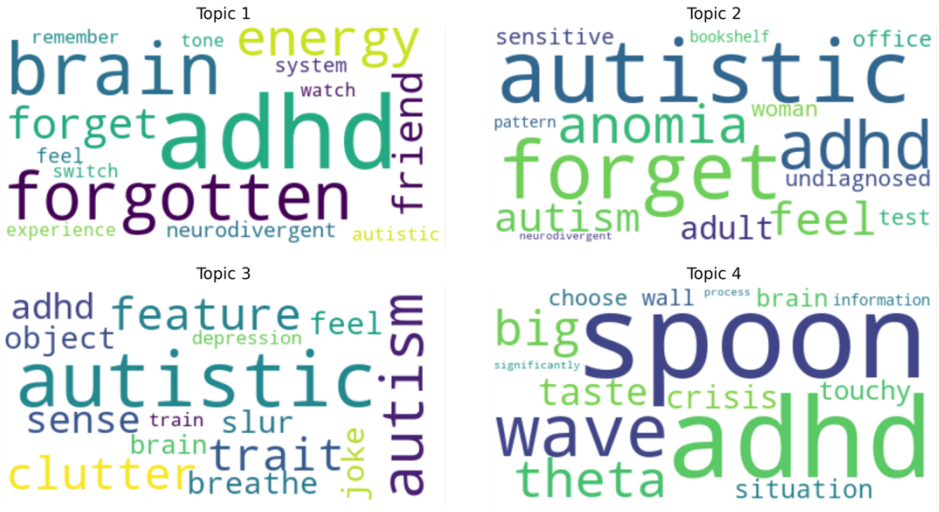

Fig. 5. Word clouds of the top 15 words for each of the four topics.

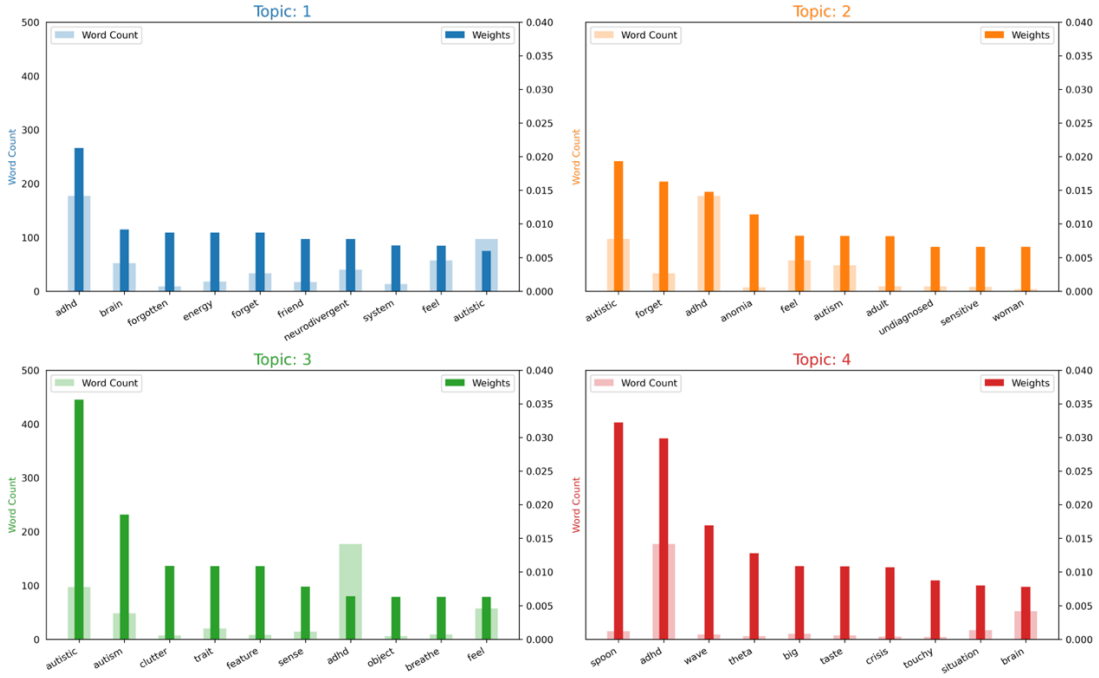

Fig. 6. Topic word counts and weights.

Next, we used the *PyLDAvis* tool to visualize how much the LDA model topics overlap with one another as shown in Fig. 7. This visual is composed of the two panels. On the left panel, multiple of circles are plotted using a multidimensional scaling algorithm. The LDA model topics are more unique and unrelated if there are more dispersed or non-overlapping circles in the intertopic distance map. This approach reduces a large number of dimensions to a manageable amount [60]. Each circle represents one single topic from the corpus. Each circle is numbered from 1 to 4 according to their decreasing order of frequency [67]. The distance between each circle tells how close the topics are in meaning. Over-lapping circles represent topics have close meaning. On the right panel, a bar chart displays the most relevant terms for understanding the selected topic in the left panel [59]. The top 30 most salient words automatically selected represent collection of

texts which are most effective at identifying topics. Word saliency value tells us how useful it is for recognizing a certain topic. A higher value means the word is more salient in that topic and a lower value means the word is less salient in that topic [60]. The word “ADHD” shows highest saliency value which means this word is the most salient word in the corpus. By selecting each topic in this interactive graphic, we can see which words are most salient for each topic relative to the corpus rearranged by salience for the topic (demonstrated in Fig. 8 for Topic 1).

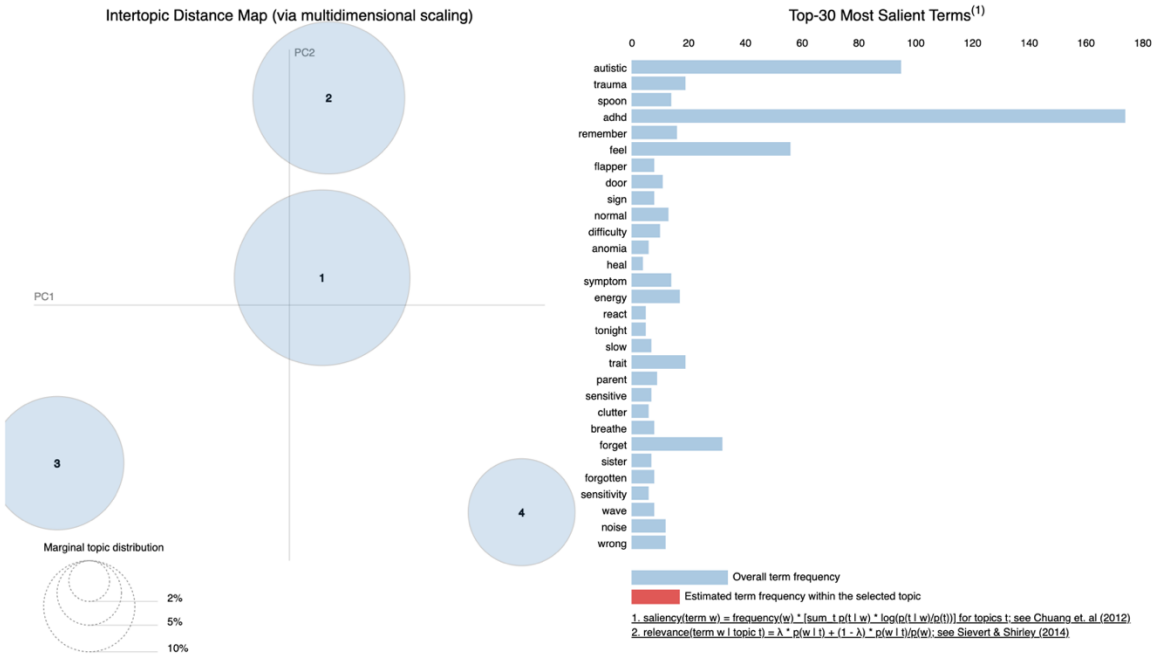

Fig. 7. Intertopic distance map for the corpus.

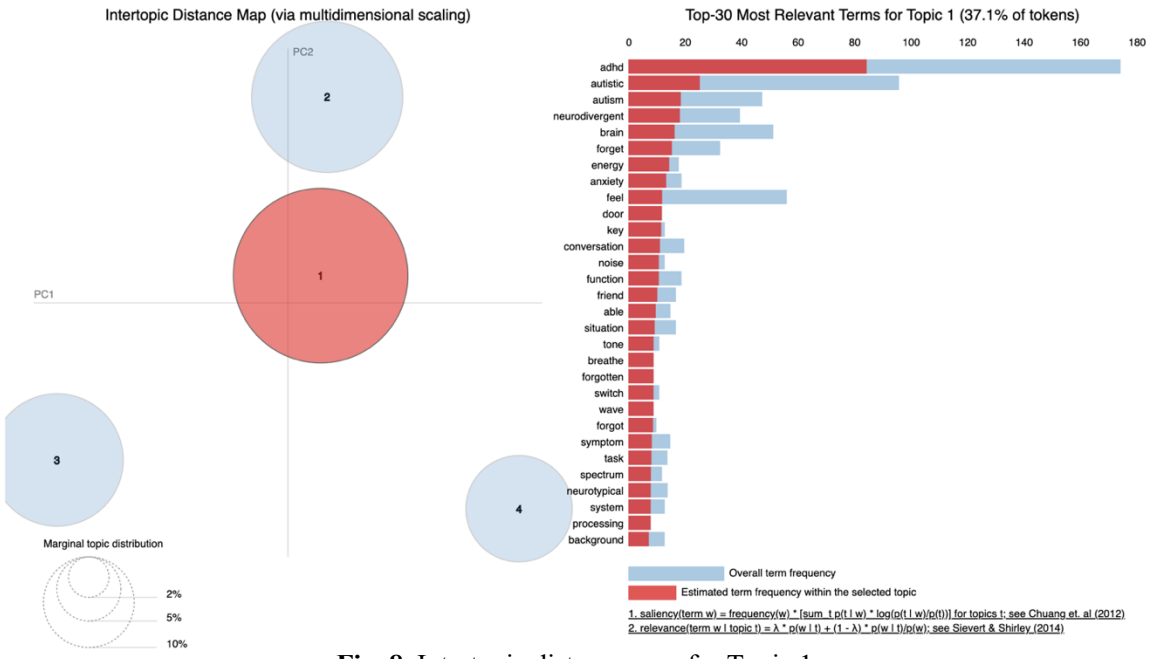

Fig. 8. Intertopic distance map for Topic 1.

7. Discussion

7.1. Using LDA in Engineering Education Research

Our work demonstrated how the latent Dirichlet allocation (LDA) topic modeling method can be used in engineering education research to identify patterns (topics) in large text-based data [68]. Essentially, LDA is a useful tool to quantitatively analyze qualitative data and identify patterns (themes) from large, text-based data. LDA should not and does not replace qualitative analysis, but this method can assist in analyzing large datasets. For example, this method can also be used to help analyze interview data by showing researchers patterns they may have missed through qualitative analysis. Depending on the type of qualitative analysis, an LDA analysis can be strategically conducted after inductive analysis to compare patterns or before deductive analysis to generate codebooks. Using LDA on its own in engineering education can benefit the field by highlighting the vocabulary used by participants, especially those in marginalized groups. We should integrate vocabulary from these participants in research to best represent the lived experiences of our participants.

The only precaution we have for using LDA and other topic modeling methods for large datasets in engineering education is to be familiar with the context of the data and that researchers should still familiarize themselves with the data. LDA has the potential to be misused or perpetuate bias by not understanding the data and the community they are researching. For example, researchers may be bias in choosing stop words during the data preprocessing stage (identified meaningless words that are removed from the corpus). To mitigate these types of risks, researchers should include participants or other researchers who are a part of the community of study who can provide insight when selecting data sources, identify important terms, and evaluate the sensitivity of results [69], [70]. Researchers can better interpret the findings from their LDA models by being familiar with the data and collaborating with their participants.

7.2. Topics from Neurodivergent TikTok

We used a four topic LDA model to identify the hidden patterns in neurodivergent TikToks. The topics that emerged are 1) ADHD experiences with memory, 2) undiagnosed autistic women's experiences, 3) autistic experience in general, and 4) ADHD spoons. These emergent topics are each unique and provide insight to the discussions happening on TikTok about being neurodivergent. As expected, the topics are more related to ADHD and autism as they have been at the forefront of the neurodiversity movement. These topics are also nuanced to ADHD and autistic experiences such as Topic 2 which may relate to undiagnosed autistic women's experiences. This topic is important to further explore to determine specific experiences related to undiagnosed autistic women who tend to be undiagnosed. Particularly, undiagnosed autistic women and AFAB people are underrepresented or even not represented in clinical research and diagnostic tools (e.g., DSM-V) [71], [72]. Thus, TikTok serves as an emergent space for this community to share their own experiences and potentially educate others who are similar and do not know. Further, Topic 2 differs from Topic 3, which may represent the general autistic experience, by relating to women's experiences. Although Topic 1 and Topic 4 relate to ADHD experiences, they are also distinct from one another and describe different types of ADHD experiences.

8. Future Work and Limitations

A major limitation to this work was the small sample size used to generate an LDA model as an LDA model is more accurate with more data for training. While the ideal sample size for LDA modeling can vary, previous literature suggest using a sample size of at least fifty with preference for larger sample sizes for more accurate and stable results [73], [74]. We plan to address this limitation in our future work by using APIs to data mine social media platforms. By data mining social media posts, we will be able to gather large amounts of data to train our LDA model for more accuracy. Further, the LDA method alone analyzes words and does not consider context or syntax of those words (e.g., the order of words matters). Other methods can be paired with LDA to analyze word combinations and sentiment to capture the context of the words in a corpus. Our future research will include sentiment analysis to assign positive, negative, and neutral meanings to the words. This analysis will be helpful in identifying strengths and challenges that come with being neurodivergent in lay terms. Other future work we plan to conduct is comparing this work's LDA results with the thematic analysis we conducted using inductive coding methods. We qualitatively generated a codebook from the same dataset that contained 56 codes. We can compare the LDA model at 56 topics with the codebook to determine how the two methods differ or result in similar outcomes. Further, a larger topic model can show how some of the topics overlap where clusters of topics may represent specific neurodivergent experiences (e.g., ADHD experiences are close together while autistic experiences are clustered but farther away from ADHD experiences).a This method comparison can provide triangulation of large text-based data both qualitatively and quantitatively.

9. Summary

In this paper, we explored how latent Dirichlet allocation (LDA), a topic modeling method that identifies hidden topics in large text-based data, can be used as a method in engineering education research. To demonstrate the method, we explored the lived experiences of neurodivergent people as posted on the video-based, social media platform TikTok. First we downloaded 200 TikToks that provided descriptions of being neurodivergent then we transcribed the TikToks. After transcription, the transcript data was loaded to a Jupyter Notebook to clean and preprocess the text. Cleaning the text involved removing symbols and converting letters to lower case while preprocessing the text consisted of removing stop words, lemmatizing words, and tokenizing the text. We generated a four topic LDA model from 196 TikToks related to being neurodivergent (four TikToks were dropped due to not meeting the inclusion criteria). The topics generated from the model related to ADHD, autism, and neurodivergent. Overall, LDA is a useful topic modeling method that can be implemented in engineering education research as it can find hidden topics in large text-based data that may be difficult to find with only qualitative methods.

References

[1] J. Kasperiuniene and V. Zydziunaite, "A Systematic Literature Review on Professional Identity Construction in Social Media," *SAGE Open*, vol. 9, no. 1, p. 2158244019828847, Jan. 2019.

[2] Gündüz, "The effect of social media on identity construction," *Mediterr. J. Soc. Sci.*, 2017.

[3] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a Feather: Homophily in Social Networks," *Annu. Rev. Sociol.*, vol. 27, pp. 415–444, 2001.

[4] K. Z. Khanam, G. Srivastava, and V. Mago, "The homophily principle in social network analysis: A survey," *Multimed. Tools Appl.*, Jan. 2022.

[5] Vizcaíno-Verdú and Aguaded, "# ThisIsMeChallenge and Music for Empowerment of Marginalized Groups on TikTok," *Media Commun.*, 2022.

[6] D. Bamman, J. Eisenstein, and T. Schnoebelen, "Gender identity and lexical variation in social media," *J. socioling.*, vol. 18, no. 2, pp. 135–160, Apr. 2014.

[7] J. Fox and R. Ralston, "Queer identity online: Informal learning and teaching experiences of LGBTQ individuals on social media," *Comput. Human Behav.*, vol. 65, pp. 635–642, Dec. 2016.

[8] M. A. DeVito, A. M. Walker, and J. Birnholtz, "'Too Gay for Facebook': Presenting LGBTQ+ Identity Throughout the Personal Social Media Ecosystem," *Proc. ACM Hum.-Comput. Interact.*, vol. 2, no. CSCW, pp. 1–23, Nov. 2018.

[9] S. Harlow and A. Benbrook, "How #Blacklivesmatter: exploring the role of hip-hop celebrities in constructing racial identity on Black Twitter," *Inf. Commun. Soc.*, vol. 22, no. 3, pp. 352–368, Feb. 2019.

[10] J. Chan, "Racial identity in online spaces: Social media's impact on students of color," *J. Stud. Aff. Res. Pract.*, vol. 54, no. 2, pp. 163–174, Apr. 2017.

[11] Y. Zhao, J. Zhang, and M. Wu, "Finding Users' Voice on Social Media: An Investigation of Online Support Groups for Autism-Affected Users on Facebook," *Int. J. Environ. Res. Public Health*, vol. 16, no. 23, Nov. 2019.

[12] J. Ortiz, A. Young, M. Myers, R. T. Bedeley, and D. Carbaugh, "Giving voice to the voiceless: The use of digital technologies by marginalized groups," 2019.

[13] J. Logan, "Queer and Neurodivergent Identity Production within the Social Media Panopticon," *The Macksey Journal*, vol. 1, no. 177, 2020.

[14] E. Buchanan, "Ethical decision-making and internet research," *Association of Internet Researchers*, 2012.

[15] *Datafication and empowerment: How the quantified self movement can help us be more human*. Big Data & Society.

[16] A. S. Franzke, A. Bechmann, M. Zimmer, C. Ess, and the Association of Internet Researchers, "Internet Research: Ethical Guidelines 3.0," 2020.

[17] J. P. Carpenter, R. Kimmons, C. R. Short, K. Clements, and M. E. Staples, "Teacher identity and crossing the professional-personal divide on twitter," *Teaching and Teacher Education*, vol. 81, pp. 1–12, May 2019.

[18] E. Heidari, G. Salimi, and M. Mehrvarz, "The influence of online social networks and online social capital on constructing a new graduate students' professional identity," *Interact. Learn. Environ.*, pp. 1–18, May 2020.

[19] A. Cavalcante, "Tumbling Into Queer Utopias and Vortexes: Experiences of LGBTQ Social Media Users on Tumblr," *J. Homosex.*, vol. 66, no. 12, pp. 1715–1735, 2019.

[20] J. C. Medina Serrano, O. Papakyriakopoulos, and S. Hegelich, "Dancing to the Partisan Beat: A First Analysis of Political Communication on TikTok," in *12th ACM Conference on Web Science*, Southampton, United Kingdom, 2020, pp. 257–266.
[21] A. Jungherr, "Twitter in Politics: A Comprehensive Literature Review," *Available at SSRN 2402443*, 27-Feb-2014.
[22] I. Literat and N. Kligler-Vilenchik, "How Popular Culture Prompts Youth Collective Political Expression and Cross-Cutting Political Talk on Social Media: A Cross-Platform Analysis," *Social Media + Society*, vol. 7, no. 2, p. 20563051211008820, Apr. 2021.
[23] A. Yeung, E. Ng, and E. Abi-Jaoude, "TikTok and Attention-Deficit/Hyperactivity Disorder: A Cross-Sectional Study of Social Media Content Quality," *Can. J. Psychiatry*, vol. 67, no. 12, pp. 899–906, Dec. 2022.
[24] V. Suarez-Lledo and J. Alvarez-Galvez, "Prevalence of Health Misinformation on Social Media: Systematic Review," *J. Med. Internet Res.*, vol. 23, no. 1, p. e17187, Jan. 2021.
[25] W. Ahmed, J. Vidal-Alaball, J. Downing, and F. López Seguí, "COVID-19 and the 5G Conspiracy Theory: Social Network Analysis of Twitter Data," *J. Med. Internet Res.*, vol. 22, no. 5, p. e19458, May 2020.
[26] D. A. González-Padilla and L. Tortolero-Blanco, "Social media influence in the COVID-19 Pandemic," *Int. Braz J Urol*, vol. 46, no. suppl.1, pp. 120–124, Jul. 2020.
[27] S. S. C. Herrick, L. Hallward, and L. R. Duncan, "'This is just how I cope': An inductive thematic analysis of eating disorder recovery content created and shared on TikTok using #EDrecovery," *Int. J. Eat. Disord.*, vol. 54, no. 4, pp. 516–526, Apr. 2021.
[28] S. Arora, S. Bindra, M. Ahmad, and T. Ahmad, "An Analysis of Depression Detection Model Applying Data Mining Approaches Using Social Network Data," in *2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 2021, pp. 1–7.
[29] C. G. P. Berdanier, C. Whitehair, A. Kirn, and D. Satterfield, "Analysis of social media forums to elicit narratives of graduate engineering student attrition," *J. Eng. Educ.*, vol. 109, no. 1, pp. 125–147, Jan. 2020.
[30] Z. Chen and A. Gillen, "How Do Engineering Students Characterize Their Educational Experience on a Popular Social Media Platform Before and During the Covid-19 Pandemic?," in *2022 ASEE Annual Conference & Exposition*, 2022.
[31] H. B. Rosqvist, N. Chown, and A. Stenning, *Neurodiversity Studies: A New Critical Paradigm*. Taylor & Francis Group, 2020.
[32] T. Eagle, "Exploring Collective Medical Knowledge and Tensions in Online ADHD Communities," in *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing*, Virtual Event, Taiwan, 2022, pp. 245–250.
[33] J. Egner, "#ActuallyAutistic: Using Twitter to construct individual and collective identity narratives," *Stud. Soc. Justice*, vol. 16, no. 2, pp. 349–369, Mar. 2022.
[34] I. Vayansky and S. A. P. Kumar, "A review of topic modeling methods," *Inf. Syst.*, vol. 94, p. 101582, Dec. 2020.
[35] D. Andrzejewski and X. Zhu, "Latent Dirichlet allocation with topic-in-set knowledge," in *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing - SemiSupLearn '09*, Boulder, Colorado, 2009.
[36] Q. Bi, K. E. Goodman, J. Kaminsky, and J. Lessler, "What is Machine Learning? A Primer for the Epidemiologist," *Am. J. Epidemiol.*, vol. 188, no. 12, pp. 2222–2239, Dec. 2019.

[37] H. Jelodar *et al.*, “Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey,” *Multimed. Tools Appl.*, vol. 78, no. 11, pp. 15169–15211, Jun. 2019.
[38] B. V. Barde and A. M. Bainwad, “An overview of topic modeling methods and tools,” in *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2017, pp. 745–750.
[39] R. Churchill and L. Singh, “The Evolution of Topic Modeling,” *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–35, Nov. 2022.
[40] J. Williams, S. J. Chinn, and J. Suleiman, “The value of Twitter for sports fans,” *Journal of Direct, Data and Digital Marketing Practice*, vol. 16, no. 1, pp. 36–50, Jul. 2014.
[41] A. F. Hidayatullah and M. R. Ma’arif, “Road traffic topic modeling on Twitter using latent dirichlet allocation,” in *2017 International Conference on Sustainable Information Engineering and Technology (SIET)*, 2017, pp. 47–52.
[42] D. Valdez, M. Ten Thij, K. Bathina, L. A. Rutter, and J. Bollen, “Social Media Insights Into US Mental Health During the COVID-19 Pandemic: Longitudinal Analysis of Twitter Data,” *J. Med. Internet Res.*, vol. 22, no. 12, p. e21418, Dec. 2020.
[43] Giri, “Is Latent Dirichlet Allocation (LDA) A clustering algorithm?,” *HDS*, 02-May-2021. [Online]. Available: https://highdemandskills.com/lda-clustering/. [Accessed: 28-Mar-2023].
[44] S. Syed and M. Spruit, “Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation,” in *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2017, pp. 165–174.
[45] M. Hasan, A. Rahman, M. R. Karim, M. S. I. Khan, and M. J. Islam, “Normalized Approach to Find Optimal Number of Topics in Latent Dirichlet Allocation (LDA),” in *Proceedings of International Conference on Trends in Computational and Cognitive Engineering*, 2021, pp. 341–354.
[46] S. Secules *et al.*, “Positionality practices and dimensions of impact on equity research: A collaborative inquiry and call to the community,” *J. Eng. Educ.*, vol. 110, no. 1, pp. 19–43, Jan. 2021.
[47] B. Bourke, “Positionality: Reflecting on the research process,” *The Qualitative Report*, Oct. 2014.
[48] J. Feldkamp, “The Rise of TikTok: The Evolution of a Social Media Platform During COVID-19,” in *Digital Responses to Covid-19: Digital Innovation, Transformation, and Entrepreneurship During Pandemic Outbreaks*, C. Hovestadt, J. Recker, J. Richter, and K. Werder, Eds. Cham: Springer International Publishing, 2021, pp. 73–85.
[49] A. Bhandari and S. Bimo, “Why’s everyone on TikTok now? The algorithmized self and the future of self-making on social media,” *Soc. Media Soc.*, vol. 8, no. 1, p. 205630512210862, Jan. 2022.
[50] E. Simpson, A. Hamann, and B. Semaan, “How to Tame ‘Your’ Algorithm: LGBTQ+ Users’ Domestication of TikTok,” *Proc. ACM Hum. Comput. Interact.*, vol. 6, no. GROUP, pp. 1–27, Jan. 2022.
[51] Y. Li, M. Guan, P. Hammond, and L. E. Berrey, “Communicating COVID-19 information on TikTok: a content analysis of TikTok videos from official accounts featured in the COVID-19 information hub,” *Health Educ. Res.*, vol. 36, no. 3, pp. 261–271, Jul. 2021.
[52] K. R. MacKinnon, H. Kia, and A. Lacombe-Duncan, “Examining TikTok’s Potential for Community-Engaged Digital Knowledge Mobilization With Equity-Seeking Groups,” *J. Med. Internet Res.*, vol. 23, no. 12, p. e30315, Dec. 2021.

[53] P. Wang, “Recommendation algorithm in TikTok: Strengths, dilemmas, and possible directions,” *Int. J. Soc. Sci. Stud.*, vol. 10, no. 5, p. 60, Sep. 2022.
[54] C. M. Cummings, N. E. Caporino, and P. C. Kendall, “Comorbidity of anxiety and depression in children and adolescents: 20 years after,” *Psychol. Bull.*, vol. 140, no. 3, pp. 816–845, May 2014.
[55] C. H. Basch, L. Donelle, J. Fera, and C. Jaime, “Deconstructing TikTok Videos on Mental Health: Cross-sectional, Descriptive Content Analysis,” *JMIR Form Res*, vol. 6, no. 5, p. e38340, May 2022.
[56] J. Barber, “Latent Dirichlet Allocation (LDA) with Python.” [Online]. Available: https://rstudio-pubs-static.s3.amazonaws.com/79360_850b2a69980c4488b1db95987a24867a.html. [Accessed: 27-Feb-2023].
[57] K. Shakeel, G. R. Tahir, I. Tehseen, and M. Ali, “A framework of Urdu topic modeling using latent dirichlet allocation (LDA),” in *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, 2018, pp. 117–123.
[58] S. Kaleru and S. R. Dhanikonda, “Exploratory Data Analysis and Latent Dirichlet Allocation on Yelp Database,” *Int. J. Eng. Res. Appl.*, vol. 13, no. 21, pp. 15035–15039, 2018.
[59] C. Sievert and K. Shirley, “LDAvis: A method for visualizing and interpreting topics,” in *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, Baltimore, Maryland, USA, 2014.
[60] “Getting to the Point with Topic Modeling,” *Alteryx Community*, 05-Aug-2020. [Online]. Available: https://community.alteryx.com/t5/Data-Science/Getting-to-the-Point-with-Topic-Modeling-Part-3-Interpreting-the/ba-p/614992. [Accessed: 28-Feb-2023].
[61] M. Röder, A. Both, and A. Hinneburg, “Exploring the space of topic coherence measures,” in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM ’15*, Shanghai, China, 2015.
[62] T. L. Griffiths and M. Steyvers, “Finding scientific topics,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101 Suppl 1, no. suppl_1, pp. 5228–5235, Apr. 2004.
[63] American Psychiatric Association, *Diagnostic and statistical manual of mental disorders*. American Psychiatric Association, 2013.
[64] D. E. Milton, “Autistic expertise: a critical reflection on the production of knowledge in autism studies,” *Autism*, vol. 18, no. 7, pp. 794–802, Oct. 2014.
[65] S. Silberman, *NeuroTribes: The Legacy of Autism and the Future of Neurodiversity*. Penguin, 2015.
[66] C. Miserandino, “The Spoon Theory written by Christine Miserandino.” But you don’t look sick. https://web. archive. org/web/20191117210039/https …, 2003.
[67] S. Tran, “Topic Modelling for Absolute Beginners,” *The Digital Skye*, 07-Nov-2020. [Online]. Available: https://thedigitalskye.com/2020/11/07/topic-modelling-for-absolute-beginners/. [Accessed: 28-Feb-2023].
[68] Textrics, “Topic Modelling: How can you use it for analysing unstructured data?,” *Medium*, 09-Jul-2021. [Online]. Available: https://textrics.medium.com/topic-modelling-how-can-you-use-it-for-analysing-unstructured-data-28045dd502cc. [Accessed: 28-Mar-2023].
[69] R. Albalawi, T. H. Yeap, and M. Benyoucef, “Using topic modeling methods for short-text data: A comparative analysis,” *Front. Artif. Intell.*, vol. 3, p. 42, Jul. 2020.

[70] J. P. Martin, S. K. Stefl, and A. E. Slaton, “LEARNING IN PUBLIC AND A PATH TOWARDS METHODOLOGICAL ACTIVISM: A CONVERSATION ON EQUITY RESEARCH,” *JWM*, vol. 28, no. 1, 2022.
[71] M. J. Ouellette, K. Rowa, N. Soreni, A. Elcock, and R. E. McCabe, “Exposure to stressful and traumatic life events in hoarding: Comparison to clinical controls,” *J. Clin. Psychol.*, vol. 77, no. 10, pp. 2216–2227, Oct. 2021.
[72] K. Barker and T. R. Galardi, “Diagnostic domain defense: Autism spectrum disorder and the DSM-5,” *Soc. Probl.*, vol. 62, no. 1, pp. 120–140, Feb. 2015.
[73] R. Deveaud, E. SanJuan, and P. Bellot, “Accurate and effective latent concept modeling for ad hoc information retrieval,” *Doc. numér.*, vol. 17, no. 1, pp. 61–84, Apr. 2014.
[74] D. Blei, “Probabilistic topic models,” in *Proceedings of the 17th ACM SIGKDD International Conference Tutorials*, San Diego California, 2011.