Discover Artificial Intelligence

Research

Public attitudes toward higher education using sentiment analysis and topic modeling

Ahmet Göçen[1] · **Mahat Maalim Ibrahim**[2] · **Asad Ul Islam Khan**[2]

Received: 4 June 2024 / Accepted: 5 November 2024
Published online: 14 November 2024

Abstract
This study examines higher education through data-mining methodologies, aiming to uncover key themes and sentiments in global discourse. Utilizing sentiment analysis and topic modeling, the research analyzes 157,943 tweets from 84,423 unique users over a five-month period (January to May 2023). This period was selected, coinciding with the rise of artificial intelligence (AI) tools, particularly ChatGPT. The study investigates the discussions, emotional tones, and dominant topics shaping the global narrative of higher education within X (Twitter) data. Key findings include the geographical distribution of tweets and the most frequent positive and negative perceptions. It also addresses critical issues such as affordability, accessibility, and funding in higher education. Furthermore, the data shows public reactions to AI in higher education are initially negative, while higher education tweets are primarily characterized by positivity and optimism. The higher education tweets are mainly posted on the weekend, with decreased activity during weekdays. This research provides insights into the evolving higher education landscape amid rapid technological advancements.

Keywords Higher education · Text mining · Topic modeling · X/Twitter · Sentiment analysis · Artificial intelligence · ChatGPT

1 Introduction

The global community has observed an unparalleled surge in the accumulation of data due to the advent of mobile devices and computers, through which individuals are connected to various platforms, thereby generating a substantial flow of interaction encompassing visual content, textual information, and blog entries. These data, primarily in an unstructured form, allow organizations and researchers to analyze patterns at a local and global level [1]. Textual analysis, in this regard, opens the door to utilizing unstructured textual data for extracting novel insights and uncovering significant patterns and correlations concealed within the data [2]. The classification of sentiments derived from social media content can provide valuable guidance for making informed decisions that stem from understanding people's emotions towards a particular issue [3]. In this context, the exploration of information sources and sentiments can assist policymakers in making appropriate and generalizable actions based on extensive datasets, ensuring the highest level of accuracy across various sectors, ranging from higher education to healthcare.

When dealing with large unstructured data, the main challenge lies in retrieving, classifying, and reporting the information. Researchers can rely on Text Mining (TM), also referred to as Intelligent Text Analysis (ITA), Text Data Mining (TDM),

✉ Ahmet Göçen, agocen@aku.edu.tr; Mahat Maalim Ibrahim, mahat.ibrahim@stu.ihu.edu.tr; Asad Ul Islam Khan, asad.khan@ihu.edu.tr | [1]Education Faculty, Afyon Kocatepe University, Afyonkarahisar, Türkiye. [2] School of Humanities and Social Sciences, Ibn Haldun University, Istanbul, Türkiye.

or Knowledge-Discovery in Text (KDT), to extract valuable insights from unstructured text [4]. Additionally, Educational Data Mining (EDM) techniques are available, specifically focusing on mining educational datasets to explore objectives such as enhancing teaching and learning processes, improving education quality, and analyzing students' achievements and learning patterns [5].

Text mining is increasingly prevalent in the field of big data analytics, finding applications in both academia and industry [6, 7] as well as socio-political studies [8, 9]. When properly collected and analyzed, big data from social media platforms can provide valuable predictions for intended outcomes. This approach can potentially bring about changes in the social, political, and economic spheres [3]. Text mining has proven valuable in analyzing educational data and guiding policy effectiveness across various contexts. For example, it helps uncover hidden patterns in educational data from different sources, offering insights that can improve decision-making [10]. Higher education institutions can employ text-mining techniques to gain insights into factors influencing prospective students' career choices [11] or to assess educational quality [12].

2 Higher education and social media data

In recent years, universities have undergone significant transformations, driven by technological advancements and societal trends toward digitalization [13]. This global digital transformation has created new pressures for higher education systems [14], which are now more visible on social media platforms, where users frequently share their sentiments and ideas. Analyzing these vast datasets can provide policymakers with cost-effective, time-efficient insights. For example, in a pivotal decision about whether to incorporate ChatGPT into higher education institutions, sentiment analysis revealed a predominantly positive public perception of ChatGPT, suggesting its continued and potentially increasing adoption [15]. Another study, based on data from over five and a half million users, highlights how AI tools like ChatGPT, which can transform the way we learn and communicate, is positioned as an intelligent learning partner [16]. However, these studies did not fully explore the long-term impact of ChatGPT on public perceptions of higher education and the implications for institutional understanding.

Social media research, particularly on platforms like Twitter, continues to attract significant attention from the academic community [17]. As a popular platform for microblogging, Twitter offers valuable insights into public sentiment across a wide range of topics, including higher education and politics. In the realm of higher education, mining Twitter data has demonstrated the potential for evaluating service effectiveness without directly seeking feedback from users [18]. Text mining, in particular, has proven effective in transforming unstructured social media posts into structured data through natural language processing techniques, enabling researchers to draw meaningful conclusions [19]. Tertiary-level educational institutions are increasingly recognizing the potential of these techniques to improve learning processes and outcomes, as well as inform their policy-making and evaluation of significant changes [20].

During the COVID-19 pandemic, Twitter discussions around higher education primarily focused on concerns regarding campus life and student admissions [21]. Sentiment analysis of these tweets revealed mixed public reactions to the transition to online education, with a generally more negative outlook [22]. However, in India, data mining was employed to assess public sentiment toward the 2020 New Education Policy, revealing widespread optimism about the policy's emphasis on inclusive and value-based education [23]. This study seeks to build on these findings by applying text mining and topic modeling to explore the diverse dimensions of public sentiment about higher education.

Text mining could help researchers present valuable implications from these posts on Twitter as it has demonstrated its effectiveness as a method that addresses the transformation of unstructured text documents into structured text using natural language processing techniques [19]. Educational institutions at the tertiary level are realizing the potential impact of text-mining techniques on the learning process and outcomes, enabling progress [20]. They use it to navigate their policy-making processes and evaluation of drastic changes. For example, during the COVID-19 pandemic, Twitter discussions around higher education were found to be largely centered on concerns related to campus life and student admissions [21]. Sentiment analysis of tweets in another study revealed mixed public sentiment about the forced transition to online education, though generally more negative [22]. Meanwhile, in India, data mining was used to evaluate the public's response to the New Education Policy introduced in 2020, which indicated widespread optimism regarding the plan's focus on fostering inclusive and value-based education in the higher education sector [23]. Similarly, this study tries to address this area by using text mining and topic modeling to reveal sentiments, prevailing topics, and key dimensions about higher education.

To realize this goal, we selected the time frame from January to May 2023, which coincides with the rise of tools like ChatGPT, introduced in late 2022. As we know, ChatGPT has revolutionized academia [16], making it both compelling and important to understand how it has influenced social media sentiments. To our knowledge, no such analysis has been conducted in the context of higher education since the advent of ChatGPT. By capturing this snapshot through sentiment analysis and topic modeling, the study contributes to the existing literature on higher education by providing a holistic understanding of recently discussed themes and trends. This research serves as an example of systematically extracting information from unstructured or semi-structured text data sourced from social media channels, ultimately presenting the findings for the higher education sector upon the arrival of large language models (LLMs). The answers to research questions included "the top locations for target keywords (Fig. 1) and "temporal patterns of interactions over time (Fig. 14.)". Some of the research questions can be given as follows for the period targeted in the study:

1- What was the general reaction on X (Twitter) about Higher Education?
2- How was the sentiment distribution in Higher Education tweets?
3- What topics related to Higher Education emerged in the topical cluster on X (Twitter)?

3 Data and methodology

In this study, we have employed a data mining framework to extract textual data about higher education from the Twitter platform utilizing the Twitter Academic API. Our investigation covers 157,943 tweets from 84,423 users from January 1, 2023 to May 28, 2023. The rationale underpinning our specific time frame arises from the sudden emergence of artificial intelligence (AI) tools, which have the potential to profoundly reshape the methodologies employed in higher education for both teaching and research. This period of interest was marked by the remarkable ascendancy of large language models, exemplified by ChatGPT, capable of generating text that appears to closely mirror human expression.

We used Twitter data as previous research has identified it as a widely used platform among academics and professionals. It is frequently utilized for sharing professional information, expressing academic viewpoints, and disseminating scientific content rather than communication purposes [24]. In the retrieval of data, we employed keywords including "higher education", "universities", "college," and similar terms. The intent was to efficiently capture pertinent discussions and hashtags exclusively aligned with our research scope. Subsequently, we extracted key attributes such as Date and time, Tweet ID, Username, Location, Like Count, Retweet Count, and Reply Count for a comprehensive data analysis in the domain of higher education. To ensure the integrity of our dataset, we initiated a sequence of text preprocessing steps. The foremost objective was to eliminate extraneous elements, thereby retaining a refined text conducive to thorough analysis. In this pursuit, our initial step involved the extraction of hashtags, user handles, URLs, numerical entities, emails,

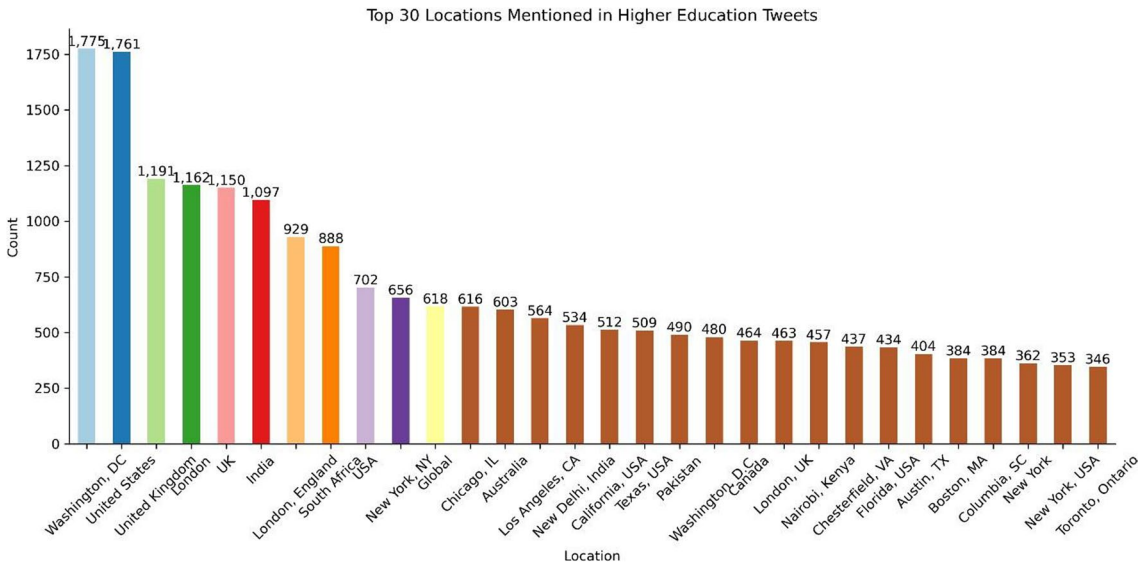


Fig. 1 Top locations for higher education tweets

emojis, HTML tags, currency symbols, and addresses. We conducted the analysis both with and without stop words and punctuation and found no significant difference in the outcomes. As a result, we chose to remove stop words and punctuation to improve computational efficiency without affecting the overall findings. The data components extracted through this process were removed, retaining only the clean and relevant text.

In addition to the transformations, we further enhanced the data quality by eliminating undesirable elements. These included the removal of bad quotes, special characters, accents to standardize text representation, punctuation marks, phone numbers, as well as instances of multiple spaces. This refinement serves to mitigate potential noise and disturbances, ensuring that the subsequent analysis is based on a clean and coherent textual dataset.

3.1 Topic Modeling Analysis

Topic modeling is a powerful technique within natural language processing (NLP) used to identify and extract hidden topics from extensive collections of text documents. We applied Latent Dirichlet Allocation (LDA) to uncover the underlying themes within the Twitter discussions. LDA operates as a generative probabilistic model, based on the assumption that each document in a text is a mixture of topics, and each topic is a mixture of words [25, 26].

The process begins with the initialization phase, where it is assumed that documents are collections of words and the entire text comprises these documents. Predefining the number of topics, LDA randomly assigns each word in every document to one of these topics. This initial random assignment serves as the starting point for the iterative process that follows.

During the iterative phase, LDA recalculates the probability of each word belonging to each topic. This calculation considers two main factors: the frequency with which the topic appears in the document (document-topic distribution) and the frequency with which the word appears in the topic (topic-word distribution). Based on these probabilities, each word is reassigned to the topic to which it most likely belongs. Subsequently, the model updates the document-topic and topic-word distributions to reflect the new assignments. This reassignment and updating process continues until the word-topic assignments stabilize, indicating that the model has converged on a good approximation of the topic distributions.

The result of the LDA process is that each document is represented as a mixture of topics, and each topic is characterized by a mixture of words. This output provides a structured understanding of the text, revealing the underlying themes that permeate the documents.

One of the significant advantages of LDA is its unsupervised nature, meaning it does not require labeled data. This makes LDA especially valuable for exploratory data analysis in contexts where labeled data may be scarce or unavailable. In addition, the advantages of LDA over other topic modeling methods are its ability to automatically identify topics in large text collections, its scalability to handle big datasets efficiently, and its flexibility to work with different types of text, from social media analysis to academic research, showcasing its versatility. The interpretability of LDA's output is another critical benefit, as the topics and associated words are easy to understand, providing clear insights into the underlying themes of the text. These features make LDA especially good at revealing hidden themes and patterns across a wide range of written material. Given these strengths, we chose to use LDA for our topic modeling rather than BERTopic, despite the latter's reputation. While BERTopic has its merits, it does not always outperform models like LDA, as acknowledged by its own creator [27]. Our decision was based on two factors: BERTopic's assumption that each document contains only one topic, which does not reflect real-world complexity, and its high computational demands compared to LDA's efficiency, particularly for large datasets. Figure 19 in Appendix (under attachments) shows the systematic workflow from data preprocessing to topic identification.

4 Findings and discussion

Figure 1 illustrates the primary geographical sources of tweets about higher education for the specified time frame. As expected, the discourse was predominantly from the US and the UK, where English is the official language, led the list. Leading the chart, Washington DC and the broader expanse of the United States collectively contribute the highest volume of tweets, tallying 1,775 originating from this region. Preeminent contributors among US cities or regions include New York, Chicago, Los Angeles, Texas, Boston, Columbia, and Florida. Meanwhile, the UK emerged as the second epicenter, with significant tweet origins attributed to locations such as the United Kingdom itself and London. Beyond these primary regions, the global discourse regarding higher education extended to other pivotal locations. New Delhi

in India, South Africa, Australia, Pakistan, Nairobi in Kenya, Austria, and Toronto in Canada constituted some of the salient centers from which discussions pertaining to higher education were launched.

It is noteworthy that a considerable proportion of these highlighted locations align with major metropolises and financial hubs, particularly within the United States. This correspondence underscores the interplay between institutions of higher education and the centers of economic activity. This linkage accentuates the connection between advanced learning and the cosmopolitan spheres of financial influence, suggesting a synergy that transcends national boundaries.

Analysis of Twitter discussions about higher education provides insight into the most common words and perceptions globally. Figure 2 includes the top 20 keywords by term frequency. It shows that the words surrounding the higher education tweets, as expected, include research, students, work, etc. One distinctive word could be "free," as the discussion on the costly sphere of higher education among students is still a key debate.

Above, Fig. 3 offers the top 20 most frequently used positive words, which again, like Fig. 2, offer a general picture of higher education in society. The top positives like "higher education," "university," "student," "college," and "institutions" feature prominently in the tweets.

There are words like "free," "public," "support," and "need," which may again suggest requested changes about affordability, accessibility, and adequate funding in higher education. Many feel higher education should be inclusive and open to all global citizens, not just the privileged few.

Words like "work," "skills," and "future" in tertiary education tweets may display its role in career preparation and economic opportunity. The vocabulary used reveals a broad picture of higher education's effects on lives and society.

In contrast to the most common positive words, Fig. 4 displays the top 20 negative words used in global Twitter discussions about higher education. Although there is some overlap with the previous list, the unique negative terms reveal particular concerns and critiques. While foundational words like "higher education," "university," "students," and "college" appear in both lists, their context here is critical rather than affirmative. The presence of additional words like "cost, debt, poor, funding, loan, expensive" may point to people voicing disapproval or frustration with the financial aspects of higher education. These words may indicate that affordability is a significant pain point, with people lamenting the high costs and limited financial assistance options. The discourse centers on deep-rooted systemic problems, fairness and equality issues, affordability challenges, and the need for academic processes to serve students and society better.

Another significant observation is the presence of 'artificial intelligence' predominantly in negative word lists, coinciding with the period when AI tools, such as OpenAI's ChatGPT, launched in November 2022, became more integrated

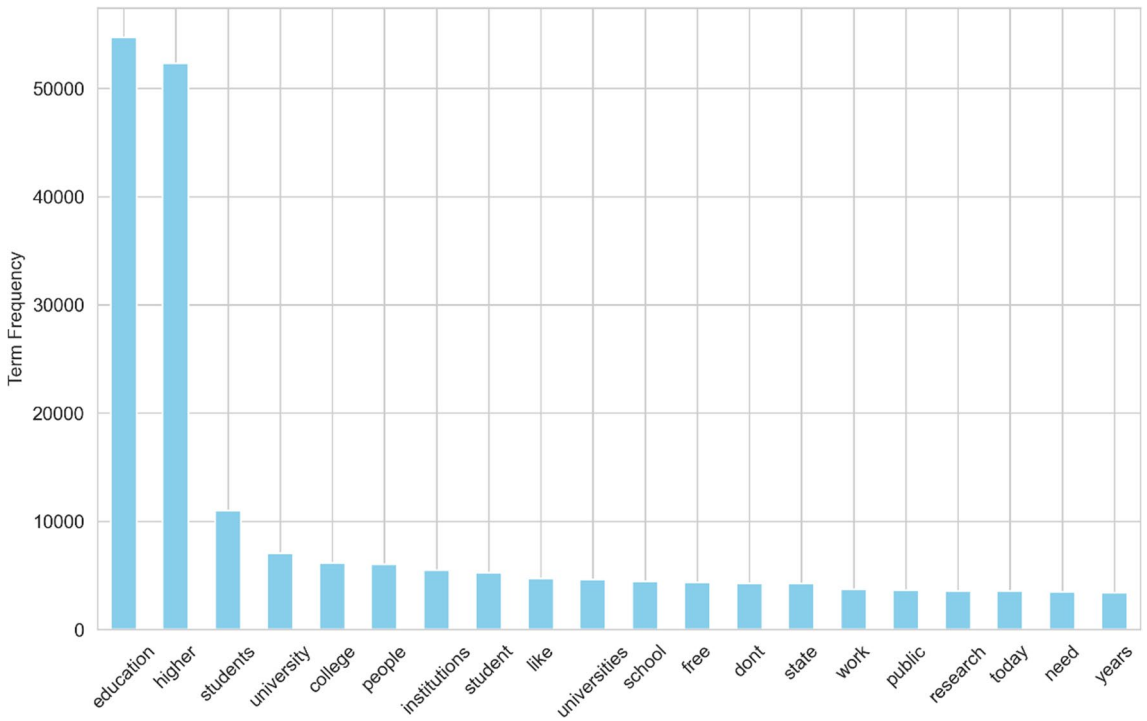


Fig. 2 Top 20 keywords by term frequency

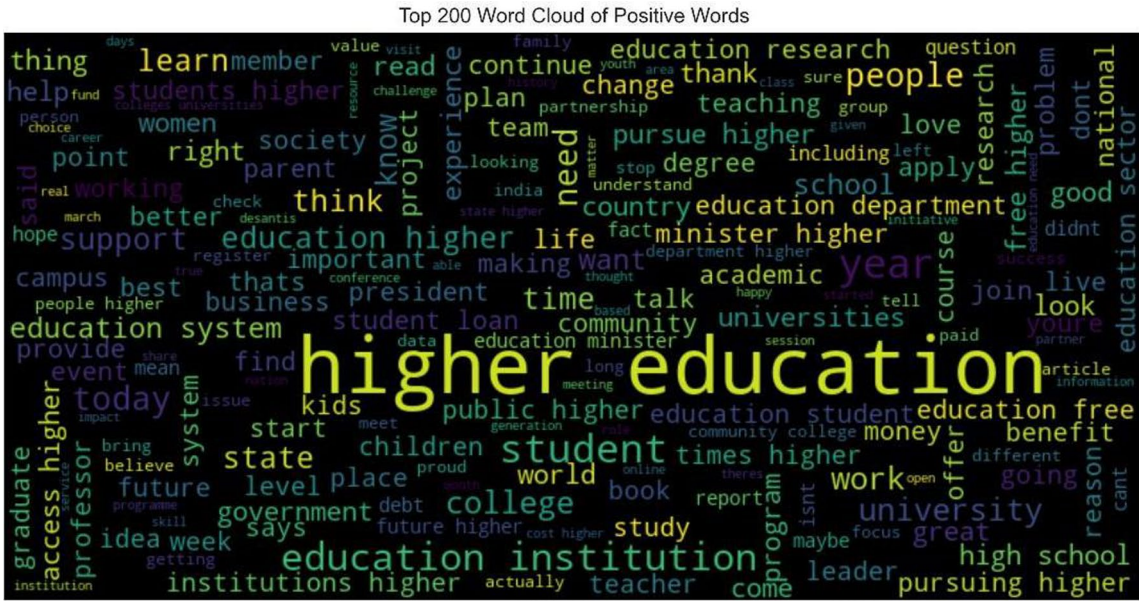


Fig. 3 Top 200 positive word clouds

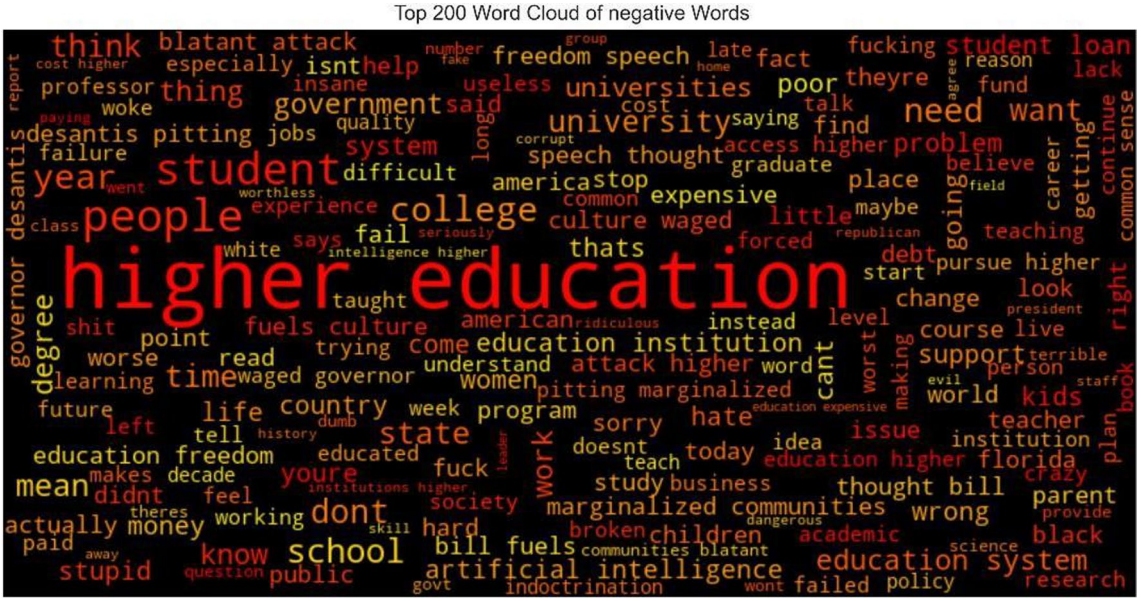


Fig. 4 Top 200 negative word clouds

into everyday life. Notably, 'AI' was absent from the first positive word cloud, suggesting that initial public sentiment towards AI in higher education was largely negative, as evidenced by its more frequent mention in negative contexts."

The words "hate, black, marginalized, indoctrination" point out the negative criticisms prevalent for higher education for years. They are still a point of discussion in negative terms.

Figure 5 displays the top 20 neutral words used in Twitter conversations about higher education, which also overlap with the previous positive and negative lists. However, the unique terms here reveal discussions of a more informational nature. Foundational words like "education," "university," and "college" appear neutrally without strong sentiment. Meanwhile, words like "research," "know," "read," and "programs" may suggest fact-based discussions about academic topics, publications, and educational offerings. Terms such as "help" and "join" likely indicate conversations around practical details like scholarship opportunities, application procedures, and program requirements rather than commentary on the state of higher education. The discourse centers on research, programs, practical details, and educational-employment

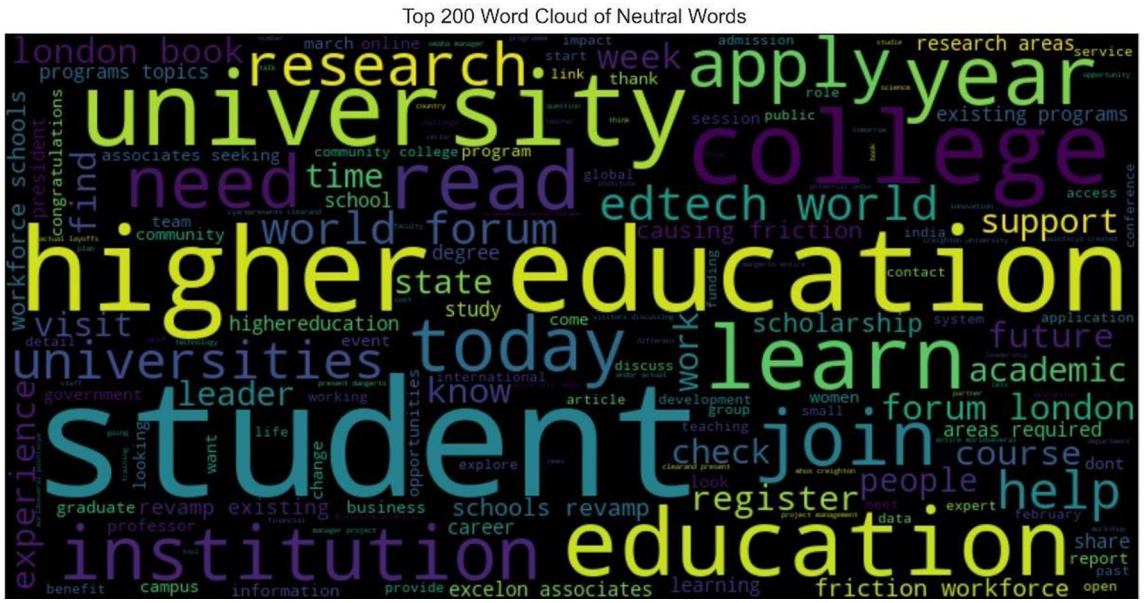


Fig. 5 A word cloud of the top 200 neutral words

links rather than systemic critiques or affirmations. The overlap with previous lists affirms foundational higher education terminology while the unique neutral terms demonstrate more pragmatic, informational dialog. This suggests that while sentiment-laden conversations may grab attention, a great deal of Twitter talk on higher education simply shares practical knowledge.

Twitter engagement regarding higher education is evident in Fig. 6, which depicts metrics such as count, retweet count, and reply count. The graph shows the predominance of likes over retweets and replies. This may suggest a prevailing inclination toward passive endorsement and consensus rather than active discourse, indicating potential factors like social dynamics or platform constraints influencing engagement patterns. This phenomenon underscores the need for further exploration into the dynamics of social media engagement in the context of higher education.

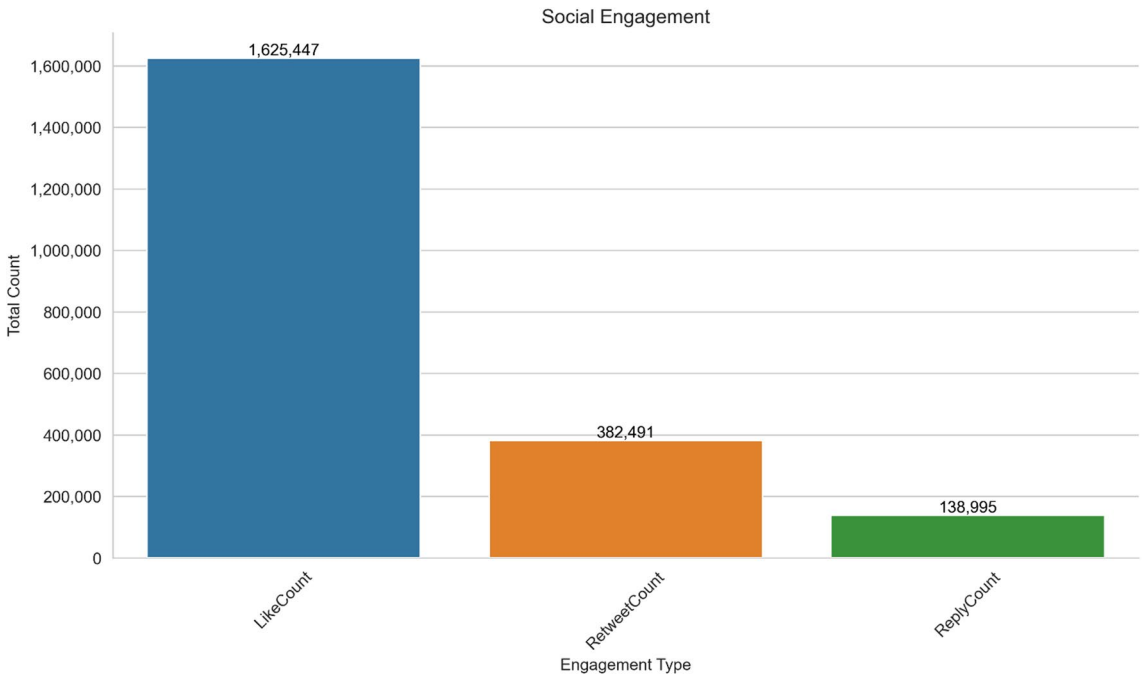


Fig. 6 Social engagement

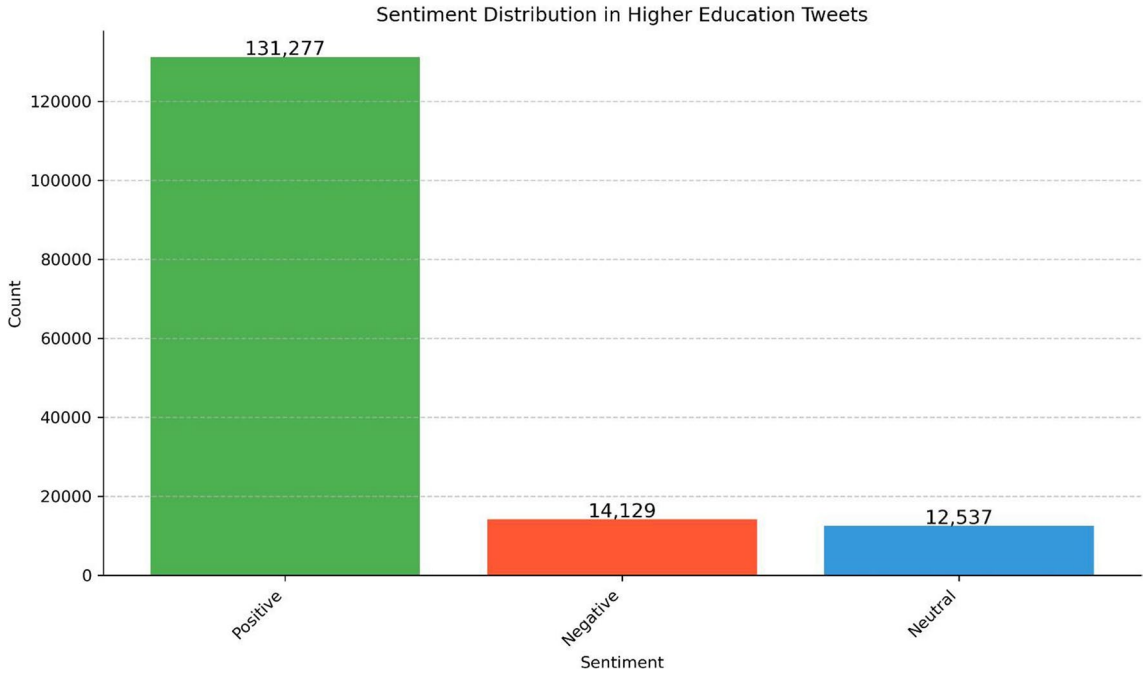
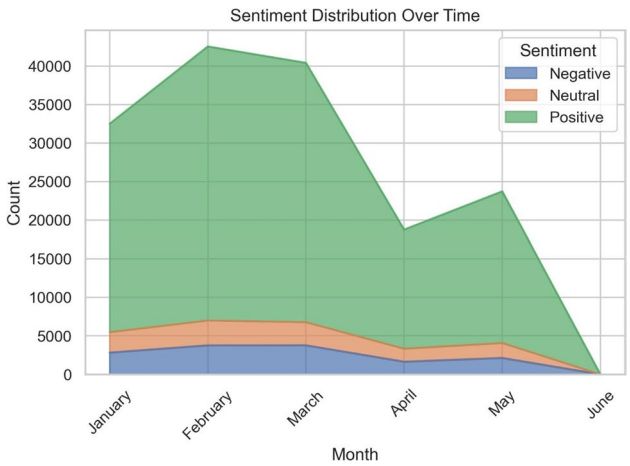


Fig. 7 Sentiment distribution in higher education tweet

Fig. 8 Sentiment distribution overtime (Monthly)



In our analysis, we have charted the sentiment distribution within higher education-related tweets, as seen in Fig. 7. Sentiments categorized as positive, negative, and neutral were evaluated, with a noteworthy predominance of positive sentiment observed. This dominance strongly suggests that the discourse surrounding higher education on tweets is primarily characterized by positivity and optimism, signifying a favorable overall perception.

5 Time series analysis

Figure 8, through their time series analysis of tweets, provides valuable insights into the dynamics of discussions during the specified timeframe. The tweets about tertiary education were higher in numbers between January and March. Additionally, the graph displays the distribution of sentiments (positive, neutral, and negative) in tweets from January to June 2023. Positive sentiment was the most dominant during this time, peaking in February and March 2023 with over 40,000 tweets. This surge likely corresponds with the increased attention and discussions in higher education due to LLMs like ChatGPT. However, from March onwards, a sharp decline in overall tweet volume is observed, with sentiment levels dropping significantly by May, potentially indicating a period of acclimatization as the novelty of these tools wore

off. A small resurgence in positive discussions appears towards the end of May, hinting at renewed interest, possibly due to the emergence of newer AI variants like Bird and Claude AI. Throughout the period, negative and neutral sentiments remained relatively stable but had lower volume compared to positive sentiments.

Figure 9 illustrates the distribution of sentiments in higher education-related tweets over a week (Monday to Sunday). The analysis categorizes the tweets into three sentiment types: positive, neutral, and negative. The data reveals that positive sentiment is consistently the most prominent across all days, with significant peaks on Friday and Sunday, suggesting increased engagement and favorable discourse toward higher education on these days. Conversely, there is a notable decline in tweet activity, particularly in positive sentiment, on Tuesday and Wednesday, indicating a midweek lull in higher education discussions. However, activity rises again on Thursday, culminating in a resurgence during the weekend, especially on Saturday and Sunday. Both neutral and negative sentiments display relatively stable patterns throughout the week, with slight increases towards the weekend. This consistency suggests that while the volume of discourse changes, the overall proportion of negative and neutral sentiments remains fairly steady. Additionally, people interested in higher education generally tweet from Friday to Monday, with lesser activity seen from Tuesday to Thursday.

6 Topic modeling

Topic modeling is key for understanding themes in text analysis. In this study, we applied Latent Dirichlet Allocation (LDA) to analyze the underlying themes within the Twitter discussions. The visual representations in Figs. 10, 11, as well as Figs. 15, 16, 17, 18, and 19 in the appendix illustrate the first ten topics identified by the LDA model. Each of these topics is broken down into subtopics, offering a detailed look at the finer points of educational discourse on Twitter. These visualizations enable us to understand the predominant themes and subthemes in the data, offering insights into public opinion and trends in education-related conversations on the platform. We found two main topic clusters:

1. Cluster: Financial aspects of education (topics 0, 2, 4): costs, loans, and debts
2. Cluster: Educational institutions and research (topics 1, 3)

We utilized the KNIME Topic Extractor (Parallel LDA) node, which implements Latent Dirichlet Allocation for topic modeling. After extensive experimentation, we found the following parameters to be optimal for our dataset: 40 words per topic, alpha of 0.1 (document-topic density), beta of 0.02 (word-topic distribution), and 1000 iterations. A lower alpha makes documents contain fewer topics, while a higher alpha allows for more topics per document. Similarly, a lower beta yields to topics with fewer words, while a higher beta allows for more words per topic. These settings provided a good balance between topic interpretability and granularity. To determine the ideal number of topics, we tested various values, assessing the coherence and distinctiveness of the resulting topics. The use of 16 parallel processing threads enhanced computational efficiency. This configuration of the KNIME node allowed us

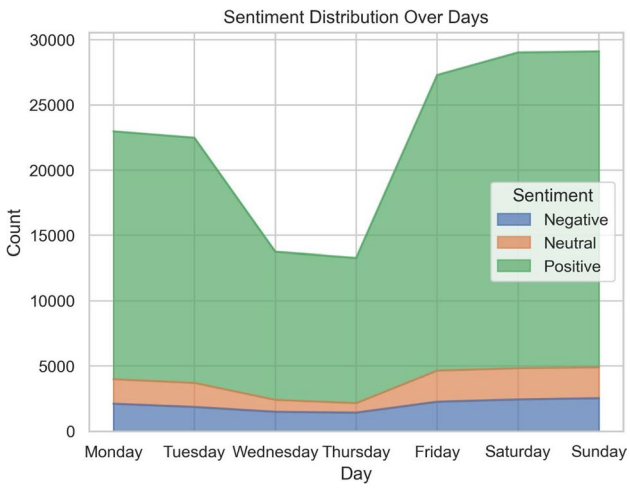


Fig. 9 Sentiment distribution over time (Daily)

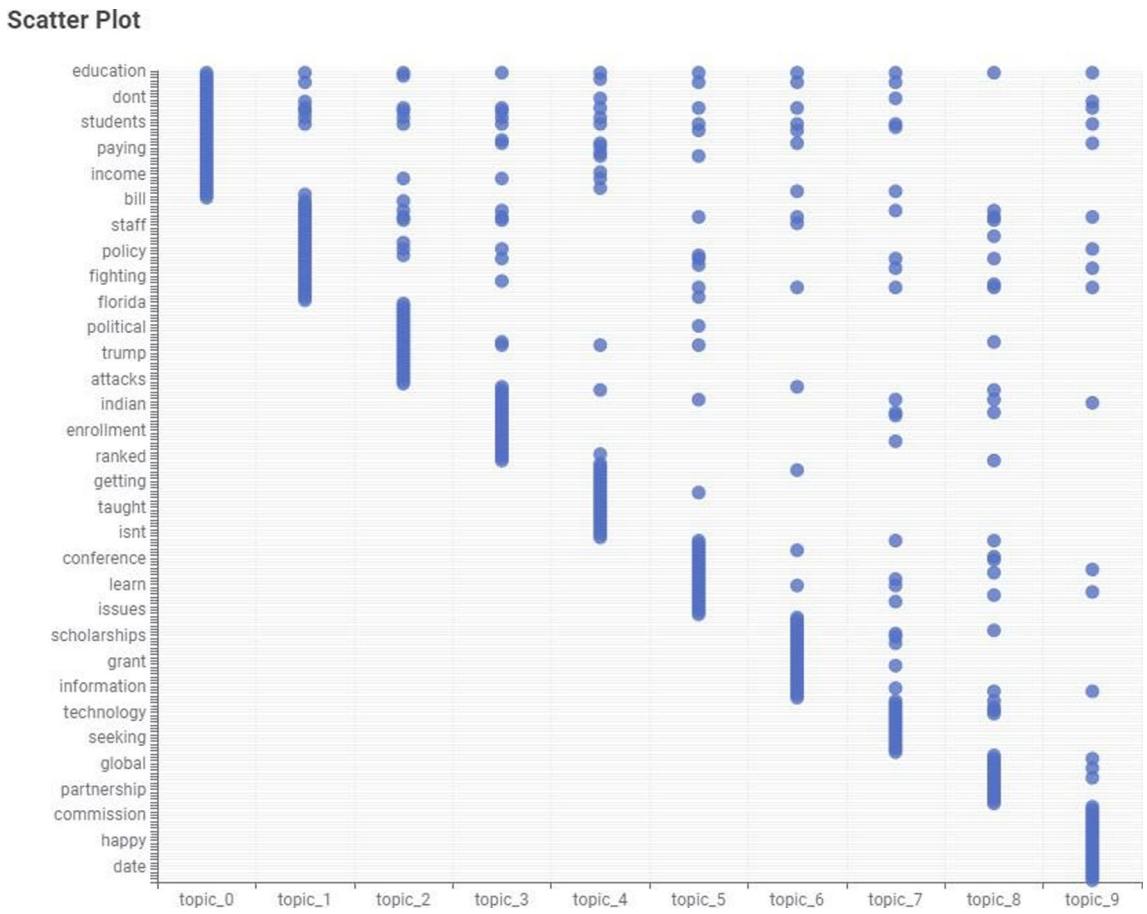


Fig. 10 Topic modeling sunburst on scatter plot

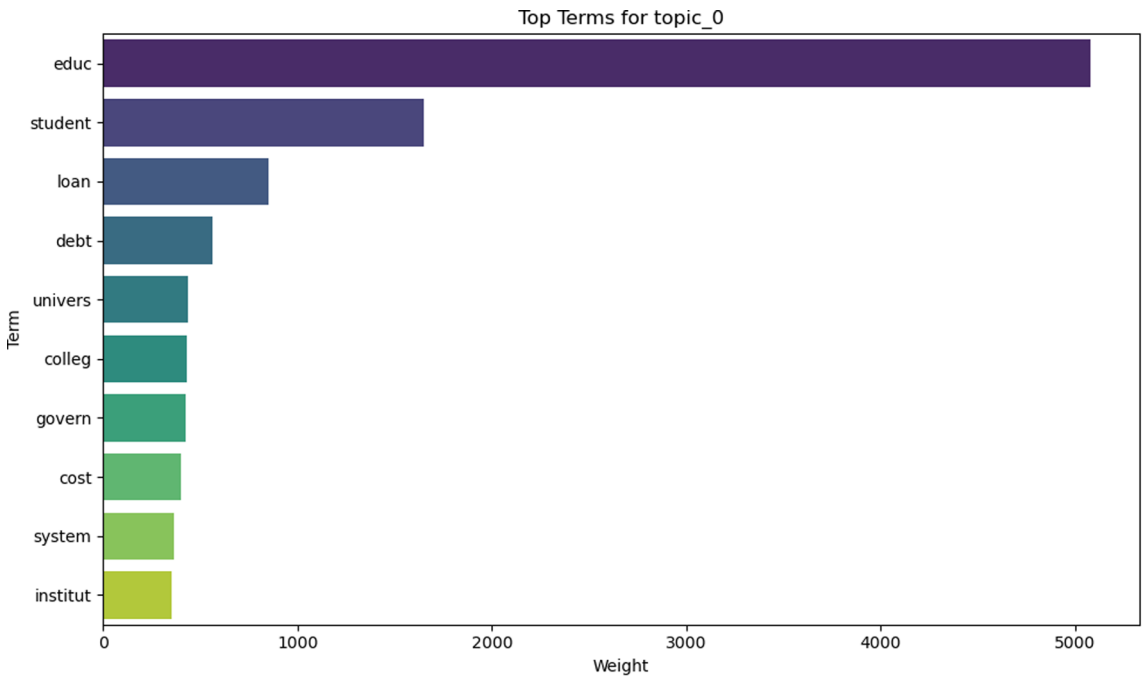


Fig. 11 Topic modeling- top terms for topic 0

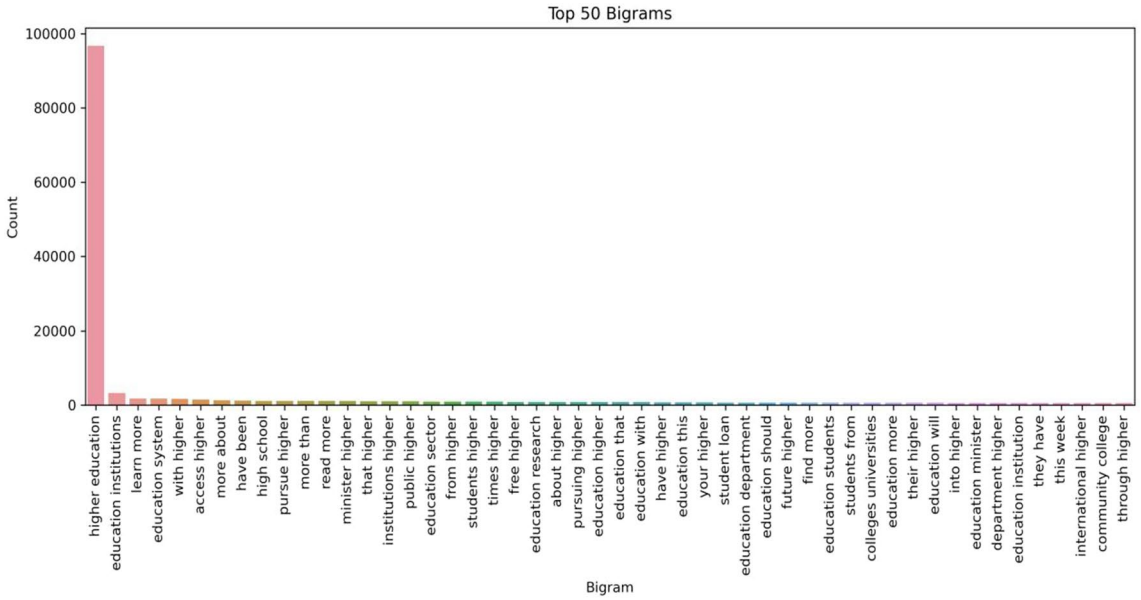


Fig. 12 Top 20 bigrams

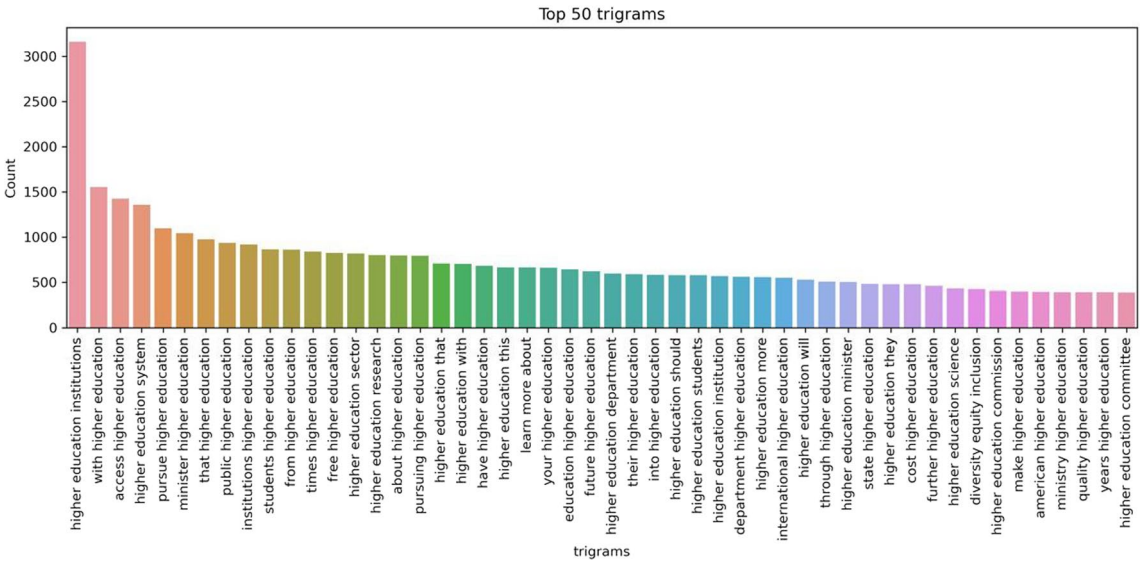


Fig. 13 Top 20 trigrams

to generate stable and meaningful topic models that captured the underlying themes in our data. Figure 19 shows the workflow.

Figure 11 displays the first topic among 10. for which Figs. 15, 16, 17, 18, and 19 in the appendix may be referred for the 5 chosen topics. To enable smooth reading, not all of the topics under separate figures are listed here.

Figures 12 and 13 depict the presence of bigrams and trigrams upon text analysis. Bigrams reveal the associations between two consecutive terms in the text, while trigrams unveil connections among three consecutive terms.

The incorporation of bigrams and trigrams in our analysis significantly enhances our understanding of the text. These methods allow us to uncover intricate relationships between terms, shedding light on the nuanced aspects of the discourse. For example, bigrams reveal associations like "higher education," "education institutions," "pursue higher education," and "learning more." Similarly, trigrams exhibit analogous patterns such as "higher education institutions," "access to higher education," "pursue higher education," and others.

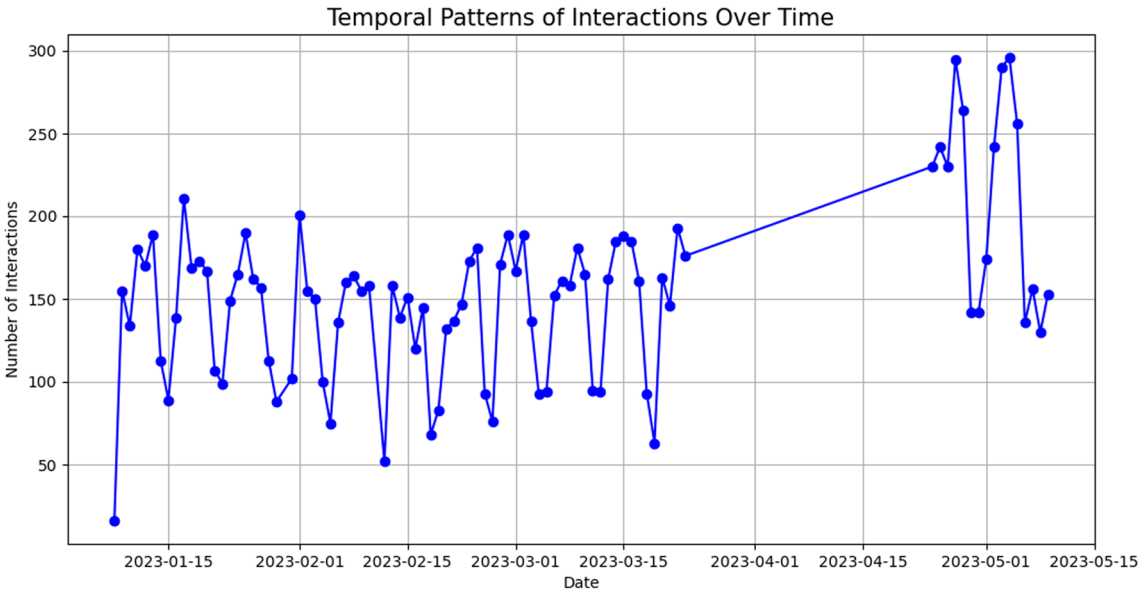


Fig. 14 Temporal patterns of interactions over time

The advantage of utilizing bigrams and trigrams in text analysis lies in their ability to capture context and semantics more effectively. They help us grasp the subtleties and intricacies of language, allowing for a more profound understanding of the underlying themes and discussions within the text.

Furthermore, Fig. 14 depicts the evolving trends in interactions within the realm of higher education on Twitter, spanning from January 15, 2023, to May 15, 2023. Interactions are gauged by the cumulative count of retweets, replies, and likes.

A discernible upward trajectory is observed in the graph, suggesting a progressive surge in interactions concerning higher education throughout this period. Several factors likely contribute to this surge, including the expanding Twitter user base engaging in higher education discussions, the growing participation of academic institutions and related organizations in the discourse, and the heightened societal significance of higher education. Figure 14 further highlights sporadic peaks in interaction levels on specific dates. These peaks may be attributed to significant events within the higher education landscape, such as the commencement of academic semesters, the publication of college rankings, appearance of new technologies, or the announcement of new government policies affecting higher education.

The remarkable spike of March 2023 coincides with the surge in popularity and extensive discussions surrounding AI chatbots, including ChatGPT started in late November 2022, and their integration into higher education discourse. This peak reflects the initial wave of intense curiosity and later normalization as users may have adapted to this innovative technology.

7 Results

The current study applies data-mining techniques, including sentiment analysis, topic modeling, and textual analysis, to identify key themes and trends in higher education. The primary goal is to provide a comprehensive analysis of discussions, sentiments, and prevalent topics related to the global landscape of higher education. This analysis examines unstructured data from Twitter (now X), spanning January to May 2023, a period marked by the rapid emergence of artificial intelligence (AI) tools like ChatGPT.

This study analyzed English-language data sourced from Twitter. Based on the results, we identified the locations of tweets about higher education, the most common positive, negative, and neutral words, as well as the top keywords based on term frequency. Additionally, we generated word clouds for each sentiment category, examined social engagement, analyzed the distribution of sentiment in higher education tweets, conducted topic modeling, explored top bigrams and trigrams, and studied temporal interaction patterns in tweets about higher education.

The primary geographical sources of tweets about higher education during the target period are predominantly from the US and the U, both of which hold the top positions and are English-speaking countries. Following closely are key locations in India, South Africa, Australia, Pakistan, Canada, and Kenya. Since tweets were collected in English and higher education institutions are often tied to global economic centers with significant financial influence, this result is unsurprising. As influential economic powers and hosts to leading higher education institutions [28], the US and UK are frequently discussed in the global higher education discourse. Interestingly, the countries with the most positive perceptions of the education system include India, Canada, Pakistan, Australia, South Africa, and Kenya, as identified in Mouronte-López et al.'s [29] study, which examines educational perceptions based on Twitter content.

Tweets in this study provide insight into the most common positive and negative perceptions of higher education globally. The most frequently used words in a positive context are "higher education," "university," "student," "college," and "institutions". Words like "education," "university," "students," and "college" appear in both positive and negative lists; their context in the negative list is critical rather than affirmative.

Words in the positive list, such as "free, fund, public, support, need," alongside keywords from the negative list, including "cost, debt, poor, funding, loan, expensive," suggest both progress and concerns surrounding affordability, accessibility, and adequate funding. The presence of negative financial terms highlights disapproval and frustration regarding the financial aspects of higher education. Similarly, Samuels [30] characterizes universities as giant investment banks or hedge funds, focusing on athletics and administration while raising tuition and reducing the quality of undergraduate education.

Higher education should be inclusive and equally accessible to all citizens. However, the higher education sector in both the USA and UK continues to face challenges related to ethnicity [31, 32]. Given that much of the data in this study originates from these countries, the frequent occurrence of the word "black" in the content highlights the significance of these issues. This underscores the need for a more democratic and modernized education system in both nations.

Positive words like "project, partnerships, work, skills, and future" may indicate tertiary education and its role in career preparation and economic opportunity, while the negative word list includes "lack, broken, marginalized, hate, indoctrination, etc." suggest persistent problems. While the positive vocabulary emphasizes higher education's potential, the negative terms reveal where people believe reality falls short of their expectations. The discourse centers on deep-rooted systemic problems, fairness and equality issues, affordability challenges, and the need for academic processes to better serve students and society.

Several words are placed in negative and positive, along with the top neutral listing, which suggests that while sentiment-laden conversations may grab attention, a great deal of Twitter talk on higher education simply shares practical knowledge as well.

As a platform Twitter reveals a significant social engagement in the field of higher education, with over 1.6 million tweets on higher education receiving likes despite the short time frame. This highlights interest in the topic. The prevalence of positive sentiment strongly indicates that discussions surrounding higher education on social media are predominantly characterized by optimism. However, there is a slight discrepancy in the viewpoint between mainstream media and reports in the United States regarding this positive opinion [33, 34]. On the other hand, trust and confidence in universities in Australia have increased following the pandemic [35].

In the text analysis of Twitter contents, we spotted that the time between January and March 2023 was marked by heightened activity and intensity on higher education. This surge coincided with the widespread publicity garnered by ChatGPT, eliciting concerns from educators and researchers regarding its potential implications on higher education. Subsequently, as April and May unfolded, heightened activity declined as the controversies over ChatGPT seemed to go into thin air, and both educators and researchers became more accustomed to its presence, thus resulting in reduced discourse. However, a resurgence in discussions is observed towards the end of May 2023, which could be due to the introduction of new versions of AI tools, including ChatGPT and Cladue.ai. This is not to mention the simple fact that students are utilizing these breakthroughs for their homework assignments [36]. Similar to the sudden surge in tweets about higher education during the period covered in the study, there has also been an unprecedented increase in the

literature regarding "ChatGPT" and "education". While there were only 730 results on Google Scholar that included both terms in the year 2022, there are, as of this writing, 18,900 results for 2023 [37]. However, the data in our study places initial participants' opinions of AI mainly in a negative sentiment context in relation to higher education tweets, as shown in Fig. 4.

In the sentiment distribution of higher education-related tweets throughout the week, positive sentiment dominates the discourse, with noticeable peaks on Friday and Sunday, suggesting that discussions surrounding higher education tend to become more positive towards the weekend. In contrast, Tuesday and Wednesday exhibit the lowest levels of sentiment activity across all categories, indicating a midweek decline in engagement. Negative and neutral sentiments remain relatively stable across the week, with slight increases toward the weekend, though they are significantly overshadowed by positive sentiments. Overall, the data indicates that Twitter discussions on higher education are predominantly positive, with higher tweet activities from Friday to Sunday and lower activities from Tuesday to Thursday Mondays mark the beginning of the work week when workers are more focused on their tasks. As the week progresses, this focus gradually lessens until the arrival of the weekend. It is worth noting that research on the optimal time and day for posting on Twitter to maximize engagement yields different results for different sectors. For reference, SocialPilot [38] and SproutSocial [39] provide further information on these variations. Our study suggests that higher education institutions should concentrate their promotional efforts, conduct opinion polls, and share their research earlier in the week and at the weekend.

We tried to find the most popular (sub)topics of higher education that people are discussing on Twitter. Regarding topic modeling, the first five topics, with each topic revealing its associated subtopics, are shown in the findings. Upon closer examination, it becomes evident that the topics can be broadly categorized into two main clusters (see Figs. 10 and 11, as well as Figs. 15, 16, 17, 18, and 19 in the Appendix). The first cluster encompasses topics that revolve around the financial aspects of education, including issues related to education costs, loans, and debts. Funding for higher education systems varies and differs across different countries [40]. Regrettably, over the past two decades, many governments worldwide have transferred the financial burden of higher education from the state to the students under cost-sharing [41]. Thus, the ongoing debates regarding fairness and accessibility for all continue based on the data from our sample. A more concerning situation arises when individuals who aspire to obtain a degree are unable to complete their education, accumulate significant debt, and sometimes face greater financial difficulties in early adulthood, particularly if they come from lower socio-economic backgrounds with colors [42].

The second cluster delves into matters about educational institutions, teaching, and research. Higher education institutions are said to be founded to serve the public good and advancement through teaching, community partnership, and research among other emerging new callings [43]. The data in the study pinpoints the same foundational messages as universities are expected to deliver the best teaching for their students, conduct research, and help community growth. Upon examination of higher education journals' abstracts in the last decades, Takei et al. [44] in parallel with these clusters, found that the research topics that became more popular over time are all student related: identities and experiences, college access, financial aid, student experiences with diversity, and student success, while the topics that became less popular over time include academic misconduct, research usage and research methodology, and academic careers.

The use of bigrams and trigrams in the study yielded several noteworthy associations such as "higher education", "access higher", "higher education institutions", "pursue higher education" "public higher education" and "free higher education". The later trigrams show some of the most repeated words altogether in tweet texts.

Temporal patterns of interactions over time, which are calculated by the cumulative count of retweets, replies, and likes, suggest a progressive surge concerning higher education throughout the target period. The progressive surge in the interactions based on tweets from January to March 2023 can be explained by important higher education topics in discussion over these days, significant events within the higher education landscape, the publication of college rankings, new technologies, and new policies or regulations on higher education. However, the remarkable spike in March 2023 seems to coincide with AI chatbots, including ChatGPT, and their integration into higher education discourse. OpenAI released the ChatGPT-3.5 language model on 30 November 2022 and then the ChatGPT-4 on 14 March 2023 [45], which sparked many discussions on higher education level [46]. This was followed by the normalization of interactions in May 2023, as users may have adapted to this innovative technology. While the surge on higher education tweets and its connection to AI is a plausible explanation, we should be cautious in this interpretation as well. AI tools became more

prevalent in everyday life in the study period, but the most commonly used words in this study do not frequently mention AI tools, with "AI" appearing only in the negative word list. This suggests that initial public reactions to AI were predominantly negative. Other factors besides AI and new technology may also have influenced these patterns of interaction and public sentiment toward higher education.

8 Limitations

Our study focused on English-language tweets about higher education. English was chosen for its global academic prevalence and status as an international language, allowing us to capture a significant portion of worldwide higher education discourse. While this decision was practical, given the vast number of languages on Twitter and the complexity of multi-language analysis, we recognize it may not capture all global perspectives. While we agree that Twitter data may not fully represent the broader population's views on higher education, Twitter remains a valuable platform for capturing real-time public discourse, especially among young and more digitally active users. As of April 2024, X (formerly Twitter) boasted an audience reach of 106.23 million users in the USA. Japan ranked second with over 69 million users, followed by India in third place with more than 25 million users [47]. Despite these limitations, our approach still offers a meaningful understanding of key trends and issues in higher education, as English remains central to international academic communication. Future research could benefit from including other prominent languages to provide a more comprehensive global view (Figs. 15, 16, 17, 18, and 19 in Appendix).

Acknowledgements We used ChatGPT for proofreading this paper.

Author contributions Dr. Ahmet Göçen wrote the inital draft, validated process and final format. Mahat Maalim Ibrahim performed text mining, text analysis and contributed to method. Dr. Asad Ul Islam contributed to inital draft and reviewed the manuscript.

Funding This is not applicable as we have not received any financial support or funding from anywhere for this study.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

Appendix

See Figs. 15, 16, 17, 18, and 19.

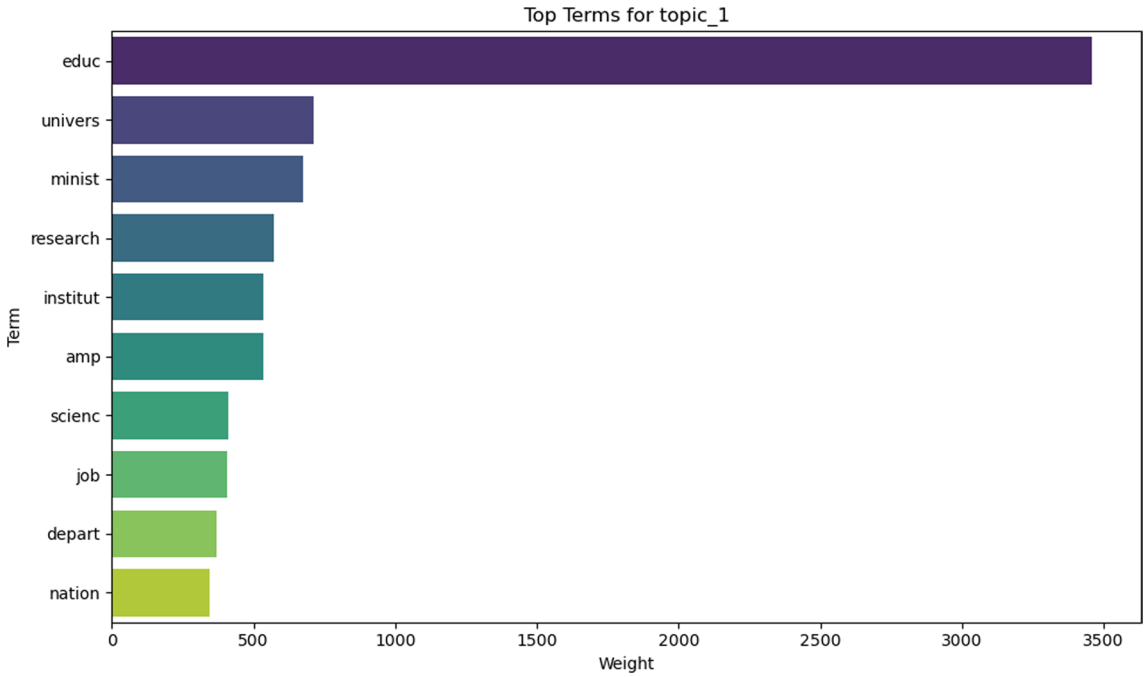


Fig. 15 Topic modelling: Top Terms for Topic 1

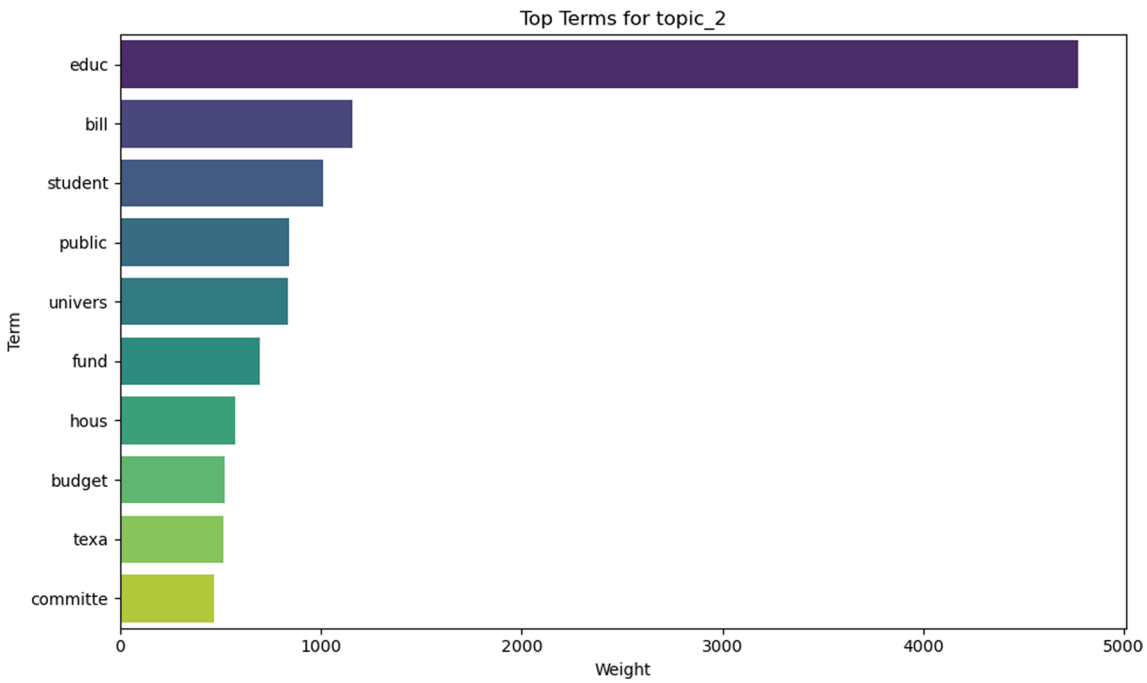


Fig. 16 Topic modelling: Top Terms for Topic 2

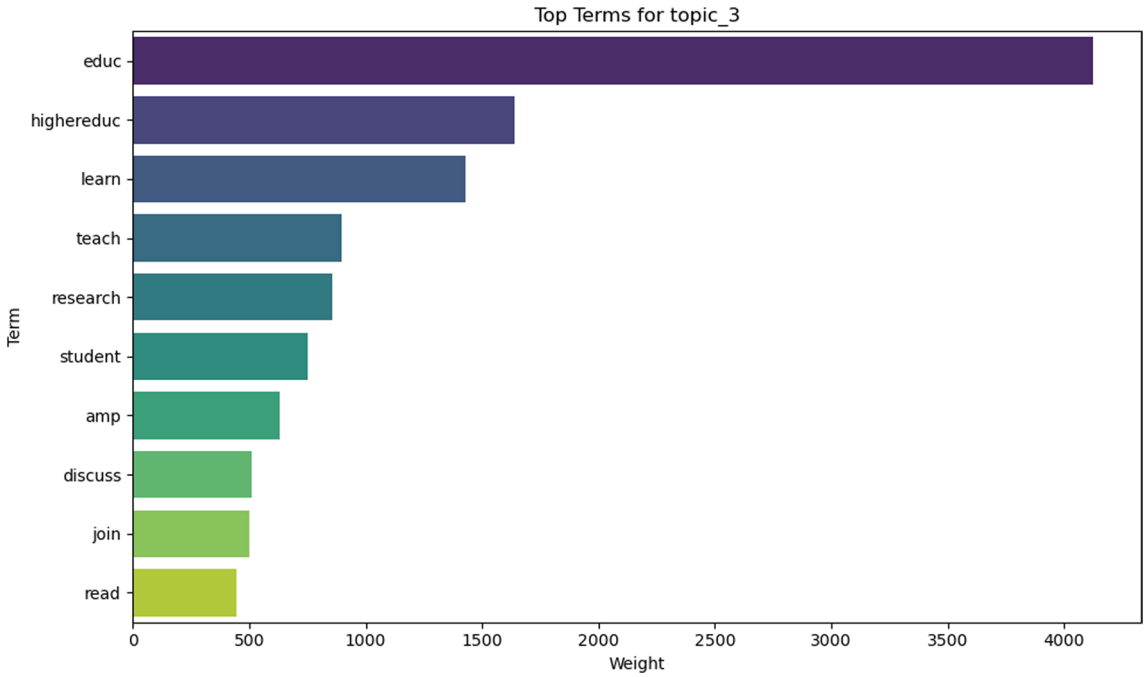


Fig. 17 Topic modelling: Top Terms for Topic 3

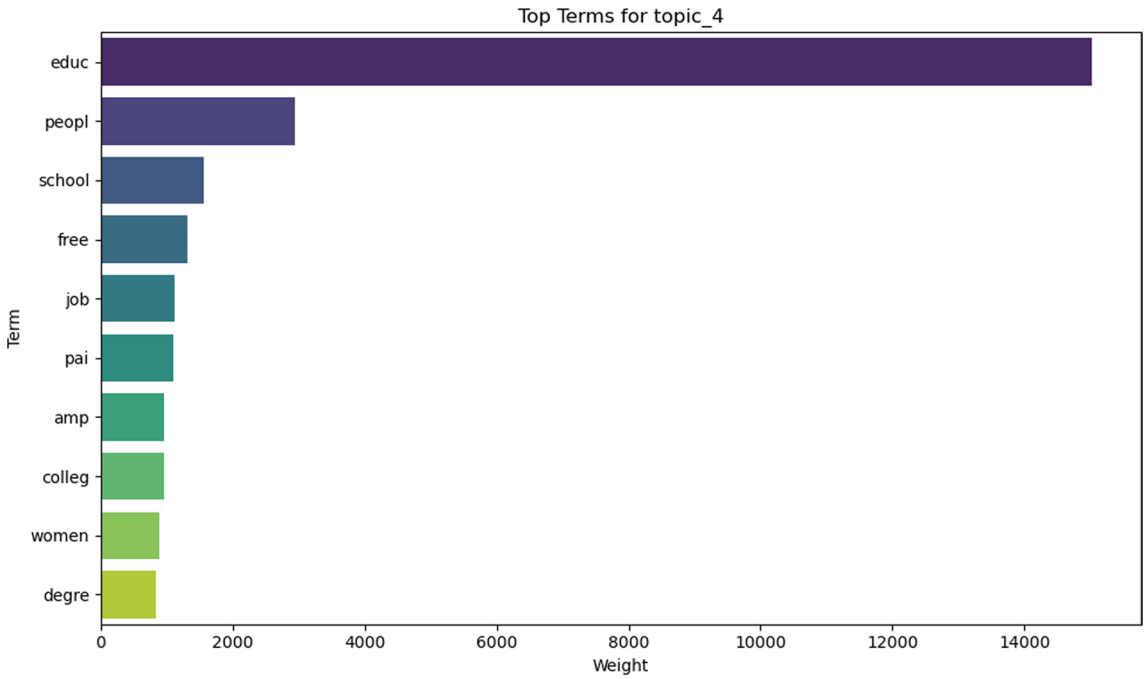


Fig. 18 Topic modelling: Top Terms for Topic 4

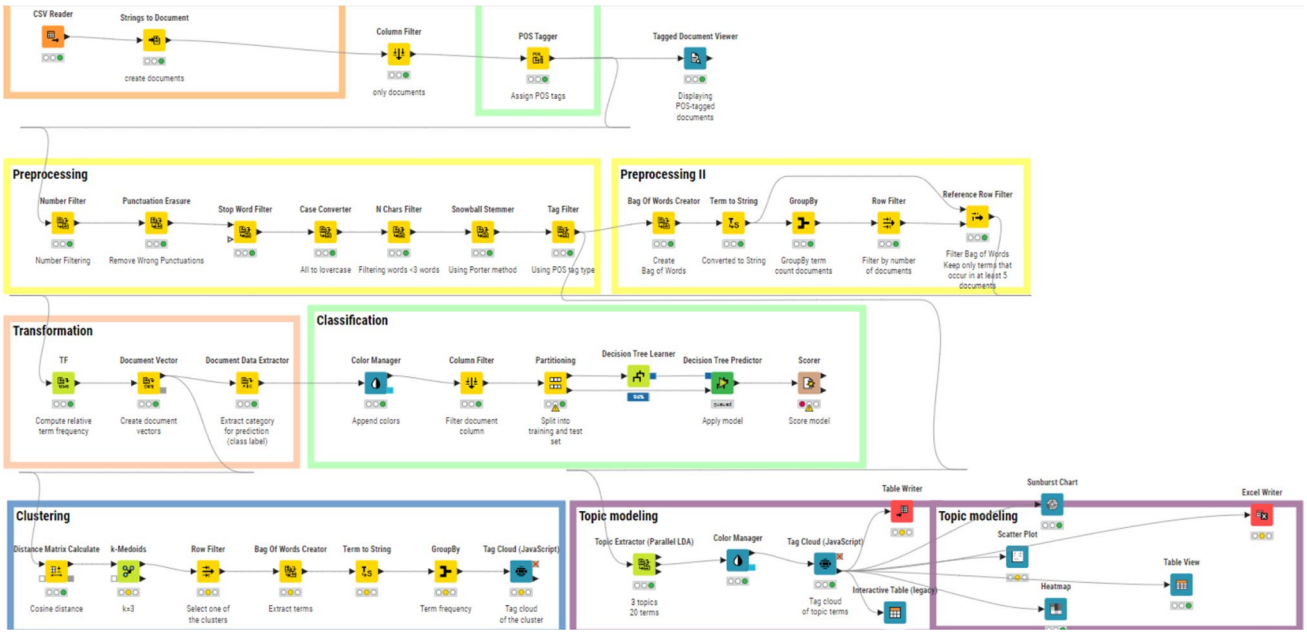


Fig. 19 Preprocessing and topic modelling workflow

References

1. Benchimol J, Kazinnik S, Saadon Y. Text mining methodologies with R: An application to central bank texts. Mach Learn Appl. 2022;8: 100286.
2. Hassani H, Beneki C, Unger S, Mazinani MT, Yeganegi MR. Text mining in big data analytics. Big Data Cogn Comput. 2020;4(1):1.
3. Namugera F, Wesonga R, Jehopio P. Text mining and determinants of sentiments: Twitter social media usage by traditional media houses in Uganda. Comput Soc Netw. 2019;6(3):1–21.
4. Ferreira-Mello R, André M, Pinheiro A, Costa E, Romero C. Text mining in education. WIREs Data Min Knowl Discov. 2019;9(6): e1332.
5. Batool S, Rashid J, Nisar MW, Kim J, Kwon HY, Hussain A. Educational data mining to predict students' academic performance: A survey study. Educ Inf Technol. 2023;28(1):905–71.
6. Arroyabe MF, Schumann M, Arranz CFA. Mapping the entrepreneurial university literature: a text mining approach. Stud High Educ. 2022;47(5):955–63.
7. Spada I, Chiarello F, Barandoni S, Ruggi G, Martini A, Fantoni G. Are universities ready to deliver digital skills and competences? A text mining-based case study of marketing courses in Italy. Technol Forecast Soc Chang. 2022;182: 121869.
8. Koytak HZ, Celik MH. A text mining approach to determinants of attitude towards Syrian immigration in the Turkish Twittersphere. Soc Sci Comput Rev. 2023;41(2):608–25.
9. Sharma A, Ghose U. Sentimental analysis of Twitter data with respect to general elections in India. Procedia Comput Sci. 2020;173:325–34.
10. Thilagaraj T, Sengottaiyan N. A review of educational data mining in higher education system. In: The Second International Conference on Research in Intelligent and Computing in Engineering, RICE. 2017. pp. 349–358.
11. Santos C, Rita P, Guerreiro J. Improving international attractiveness of higher education institutions based on text mining and sentiment analysis. Int J Educ Manag. 2018;32:00–00.
12. Qiu RG, Ravi RR, Qiu LL. Aggregating and visualizing public opinions and sentiment trends on the US higher education. In iiWAS '15: Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services (2015).
13. Abad-Segura E, González-Zamar MD, Infante-Moro JC, Ruipérez García G. Sustainable management of digital transformation in higher education: Global research trends. Sustainability. 2020;12(5):2107.
14. Castro R. Blended learning in higher education: Trends and capabilities. Educ Inf Technol. 2019;24(4):2523–46.
15. Adeshola I, Adepoju AP. The opportunities and challenges of ChatGPT in education. Interact Learn Environ. 2023. https://doi.org/10.1080/10494820.2023.2253858.
16. Fütterer T, Fischer C, Alekseeva A, Chen X, Tate T, Warschauer M, et al. ChatGPT in education: Global reactions to AI innovations. Sci Rep. 2023;13(1):15310.
17. Manca S. Snapping, pinning, liking or texting: Investigating social media in higher education beyond Facebook. Internet Higher Educ. 2020;44: 100707.
18. Agostino D, Arnaboldi M. Social media data used in the measurement of public services effectiveness: Empirical evidence from Twitter in higher education institutions. Public Policy Admin. 2017;32(4):296–322.
19. Malik S, Gupta SK. The importance of text mining for services management. Technoarete Trans Intell Data Min Knowl Discov. 2022;4(2):28–33.
20. Aldowah H, Al-Samarraie H, Fauzy WM. Educational data mining and learning analytics for 21st century higher education: A review and synthesis. Telematics Inform. 2019;37:13–49.

21. Lieharyani D, Ambarwati R. Visualisasi data tweet di sektor pendidikan tinggi pada saat masa pandemi. Build Inform Technol Sci. 2022;4(1):116–23.
22. Brandon D. Data mining twitter for COVID-19 sentiments concerning college online education. Future Bus J. 2023;9(1):104.
23. Kaurav RPS, Narula S, Baber R, Tiwari P. Theoretical extension of the new education policy 2020 using twitter mining. J Content Community Commun. 2021;13(1):16–26.
24. Apostolidis C, Devine A, Jabbar A. From chalk to clicks – The impact of (rapid) technology adoption on employee emotions in the higher education sector. Technol Forecast Soc Chang. 2022;182: 121860.
25. Blei DM. Probabilistic topic models. Commun ACM. 2012;55(4):77–84.
26. Blei DM, Ng A, Jordan M. Latent Dirichlet allocation. J Mach Learn Res. 2003;3:993–1022.
27. Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure [Internet]. arXiv; 2020. https://arxiv.org/abs/2203.05794
28. World University Rankings. World University Rankings. 2024. https://www.timeshighereducation.com/world-university-rankings/2024/world-ranking.
29. Mouronte-López ML, Ceres JS, Columbrans AM. Analysing the sentiments about the education system through Twitter. Educ Inf Technol. 2023;28(9):10965–94.
30. Samuels R. Why public higher education should be free: How to decrease costs and increase quality at American universities. New Brunswick: Rutgers University Press; 2013.
31. Arday J, Branchu C, Boliver V. What do we know about black and minority ethnic (BAME) participation in UK higher education? Soc Policy Soc. 2022;21(1):12–25.
32. Lomotey K, Smith WA. The racial crisis in American higher education. New York: State University of New York Press; 2023.
33. Brenan, M. Americans' confidence in higher education down sharply. Gallup [Internet]. 2023 Jul 11; Available from: https://news.gallup.com/poll/508352/americans-confidence-higher-education-down-sharply.aspx
34. Kelderman E. What the public really thinks about higher education. The Chronicle of Higher Education [Internet]. 2023 Sep 5; https://www.chronicle.com/article/what-the-public-really-thinks-about-higher-education
35. Biddle N. Top marks: Aussies show trust in education institutions. ANU Reporter [Internet]. 2023 Aug 9; https://reporter.anu.edu.au/all-stories/top-marks-aussies-show-trust-in-education-institutions
36. Arkan A, Göçen A, Bulut MA. Artificial intelligence in higher education: Applications and suggestions. İstanbul: Ibn Haldun Yayınevi; 2023.
37. Google Scholar. "ChatGPT" "education. 2023 Oct 22; https://scholar.google.com.tr/
38. SocialPilot. What is the Best Time to Post on Twitter in 2023?. 2023. https://www.socialpilot.co/blog/best-time-to-post-on-twitter [Internet].
39. SproutSocial. Best times to post on Twitter in 2023. https://sproutsocial.com/insights/best-times-to-post-on-twitter/, 2023.
40. Garritzmann JL. The political economy of higher education finance: The politics of tuition fees and subsidies in OECD countries, 1945–2015. Cham: Palgrave Macmillan; 2016.
41. Cattaneo M, Civera A, Meoli M, Paleari S. Analysing policies to increase graduate population: do tuition fees matter? Eur J Higher Educ. 2020;10(1):10–27.
42. Payne SSC. Equalization or reproduction? "Some college" and the social function of higher education. Sociol Educ. 2023;96(2):104–28.
43. Jongbloed B, Enders J, Salerno C. Higher education and its communities: Interconnections, interdependencies and a research agenda. High Educ. 2008;56(3):303–24.
44. Takei M, Porter SR, Umbach PD, Nakano J. Understanding themes in postsecondary research using topic modeling and journal abstracts. Res High Educ. 2024;65(3):510–51.
45. Skavronskaya L, Hadinejad A, Cotterell D. Reversing the threat of artificial intelligence to opportunity: A discussion of ChatGPT in tourism education. J Teach Travel Tour. 2023;23(2):253–8.
46. Rasul T, Nair S, Kalendra D, Robin M, Santini F, Ladeira W, et al. The role of ChatGPT in higher education: Benefits, challenges, and future research directions. J Appl Learn Teach. 2023. https://doi.org/10.37074/jalt.2023.6.1.29.
47. Statista. Leading countries based on number of X (formerly Twitter) users as of April 2024. https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/, 2024.