What impacts matriculation decisions? A double-blind experiment via an AI-led chatbot trained with social media data

Ilker Cingillioglu

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Business School
The University of Sydney
2024

Thesis frontispiece

Thesis title

What impacts matriculation decisions? A double-blind experiment via an AI-led chatbot trained with social media data

Thesis abstract

This thesis explores students' matriculation decision factors via an AI-led chatbot trained with social media data. The primary aim of this thesis is to develop a series of methodologies through which rich qualitative and large quantitative data can be collected with a double-blind randomized controlled trial (RCT) run by the AI, and then be analyzed to make causal inferences delineating the factors that impact students' university choices. The novelty of this thesis resides in the following methodological approaches: Firstly, it employs data mining and text analytics techniques to explore the use of topic modelling and a systematic literature reviewing technique called algorithmic document sequencing to identify decision factors from social media to be integrated to the internal model of the AI through a methodological pluralist approach. Secondly, it introduces a chatbot design and strategy for an AI-led chat survey generating both unstructured qualitative and structured quantitative primary data. Finally, upon interviewing 1193 participants around the world, a double-blind true experiment was run seamlessly without human intervention by the AI testing hypotheses and determining the factors that impact students' university choices. In this automated experiment, I tested eight hypotheses for eight choice factors. The AI-experiment validated five of these hypotheses and rejected three factors previously acknowledged in the literature as influential in students' choices of universities. I showcased how AI can efficiently interview participants and collect their input, offering robust evidence through an RCT (Gold standard) to establish causal relationships between interventions and their outcomes. One significant contribution of the thesis lies in aiding higher education institutions in understanding the global factors influencing students' university choices and the role of electronic word-of-mouth on social media platforms. More importantly, the research enhances knowledge in identifying themes from social media and literature, facilitating the training of AI-augmented chatbots with these themes, and designing such chatbots to run large scale social RCTs. These developments may enable researchers from a wide range of fields to collect qualitative and quantitative data from large samples, run double-blind true experiments with the AI and produce statistically reproducible, reliable, and generalizable results.

Authorship attribution statement

Chapter 1 of this thesis is published as [Cingillioglu, I., Gal, U., & Prokhorov, A. (2021, November). An Analysis of Tweet Relevance, Twitter Activity and Student Preferences for Universities. In 2021 19th International Conference on Information Technology Based Higher Education and Training (ITHET) (pp. 01-10). IEEE.]

Chapter 2 of this thesis is published as [Cingillioglu, I., Gal, U., & Prokhorov, A. (2022, November). Facebook Engagement and Student Preferences for Universities. In *2022 20th International Conference on Information Technology Based Higher Education and Training (ITHET)* (pp. 1-10). IEEE.]

Chapter 5 of this thesis is published as [Cingillioglu, I., Gal, U., & Prokhorov, A. (2023). Social media marketing for student recruitment: an algorithmically sequenced literature review. *Journal of Marketing for Higher Education*, 1-23.]

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Supervisor Name, Signature, Date

Uri Gal

Artem Prokhorov

Table of Contents

Statement of originality

This is to certify that to the best of my knowledge; the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Signature

Name

Ilker Cingillioglu

Acknowledgements

I would like to express my sincere appreciation and deepest gratitude to all those who have supported and guided me throughout the journey of completing this thesis. Without their valuable contributions, this research would not have come to fruition.

First and foremost, I would like to express my gratitude to both of my thesis supervisors, Professor Uri Gal, and Professor Artem Prokhorov, for their unwavering support and guidance. Their insightful feedback and particularly constructive criticism have substantially shaped the direction and quality of my research. I am truly grateful for their mentorship and the invaluable research experience I have gained under their supervision.

I would like to extend my gratitude and appreciation to the faculty members of the Business Information Systems at the University of Sydney Business School, particularly Professor Dirk Hovorka for his precious guidance that truly helped me take the quality of this research to the next level.

I would also like to extend my gratitude to my dear wife Sema for her love, encouragement, and patience throughout my entire PhD candidature. Her support has always been my core strength, and I am grateful for everything she has done for me.

Last but not the least, I would like to acknowledge the researchers of the broader academic community whose earlier contributions have paved the way for this study.

Thank you all so much,

Ilker Cingillioglu

This thesis is made of 4 parts comprising 6 chapters in total. The first part provides a general overview of the thesis including its aims, scope, novelty, overall methodology and potential contributions. It also explains how its chapters link to one another. The second part contains 5 chapters building upon one another and leading to the third part which elaborates the process of the double-blind AI-based experiment and its results. The fourth (final) part contains discussion and implications of the results of the experiment, as well as its limitations, recommendations for further studies and conclusion of the thesis.

Part 1
Overview of the thesis

Introduction

In this thesis, students' matriculation decision factors are explored via an AI-augmented contextual chatbot trained with social media data. The overarching aims of this thesis are twofold: (1) to develop a methodology through which rich qualitative and large quantitative data can be collected and used in double-blind experiments through an AI-led interview-like survey, and (2) to make causal inferences delineating the factors that impact students' university choices.

Interviews are a common data collection method utilized mostly in qualitative studies where researchers record transcriptions as data and then analyse them to generate theory (Schultze & Avital, 2011) and gain insights into the question of "why" people behave, think or act in a particular way (Black, 1994; Rosenthal, 2016). Open-ended questions are usually asked in interviews and initial responses are probed with follow-up questions to collect thick and rich descriptions of respondents' opinions, lived experiences, and behaviour about a phenomenon that is needed to be explored in greater depth than quantitative methods (Leeson et al., 2019).

Qualitative interviews also have shortcomings. Firstly, the less structured they are and the more open-ended questions they incorporate, the more time-consuming it gets to interpret and analyze qualitative data. To extract relevant information from interview transcripts, researchers use a method called coding to identify text that match with key themes from which overarching dimensions of the inquired phenomena are developed (Linneberg & Korsgaard, 2019). Even if only a dozen respondents get interviewed, it may take weeks to complete this process (Rossetto, 2014). Secondly, since qualitative data are usually conducted by a team of researchers, as Diefenbach (2009) posited, dissent and inconsistencies in coding and interpreting the results may emerge as factors that require constant reconciliation. Thirdly, although the findings of most qualitative interviews may be transferable to another setting, since data are collected from a relatively small number of cases, the results cannot be generalized to a larger population (Anderson, 2010). To overcome this, researchers may opt to increase the number of cases, but this will generate larger volumes of data which will make the interpretation and analysis of data even more laborious and time-consuming. Furthermore, due to the low number of respondents, presence of researchers during data collection and the involvement of multiple analysts during data interpretation and analysis, other concerns may be raised such as that the qualitative study may be prone to bias and not be replicable and reliable (Turner, 2010).

Unlike traditional qualitative interviews, automated chat surveys conducted by artificial intelligence (AI) are neither constrained by time nor prone to observer bias or coding inconsistencies. Rather than relying on the perspectives of a small number of respondents, opinions and experiences of thousands of people can be collected and coded real-time through chatbots in a streamlined, efficient, and cohesive manner allowing researchers to statistically test their hypotheses supported by qualitative data.

In this research, a new design for an AI-augmented chat survey (i.e., interview-like chatbot survey) powered by the IBM's virtual chatbot agent, Watson Assistant, was built to conduct a double-blind experiment on humans. The chatbot was trained with the potential factors that may impact students' matriculation decisions. These factors were identified through a methodological pluralist approach incorporating the union of the output of a systematic literature review and topic models constructed with unstructured text data extracted from Facebook and Twitter.

Matriculation decision factors is a vital area of investigation, as it provides valuable insights into the decision-making process of prospective students, and helps universities and policymakers adapt their strategies to better meet the needs and expectations of students. There exists a notable research gap in this area, primarily due to the challenges of achieving generalizability on a global scale. While research on this topic is abundant, it often tends to be region- or institution-specific, making it challenging to draw comprehensive, globally applicable conclusions. The factors shaping university choices can vary significantly between countries, cultures, and socio-economic backgrounds, thus necessitating broader investigations to truly capture the diverse dynamics at play. What's more, these choice factors may constantly shift due to technologically induced trends (e.g., social media), creating a dynamic research environment that demands further exploration to guide institutions and policymakers in meeting the diverse needs of students on a global scale. Therefore, addressing this research gap with a more global perspective is essential to develop a holistic understanding of the universal and context-specific factors influencing students' matriculation decisions.

The novelty of this research are threefold: (1) The methodological approach incorporating the use of social media analysis techniques to collect unstructured secondary data from Facebook and Twitter; and transform them to 'triangulated' structured data which will then be the basis of (2) the architecture of a natural language processing AI-led chatbot developed to collect open-ended and quantitative data via an interview-like survey generating unstructured qualitative primary data and structured quantitative primary data and run (3) a double-blind experiment with the chatbot designed to test what factors influence students' university choices. To meet these objectives, the thesis proceeds as follows: During literature review, I gathered as many information as possible about how universities use social media to attract potential students and how students' matriculation decision factors vary around the world. During review, I narrowed my focus down on two of the most popular social media platforms (i.e., Facebook and Twitter) and the role of electronic word-of-mouth (eWOM) and engagement on these platforms. After the literature review, as initial stages of the study, I explored and established the relationship between student preferences for universities and public engagement on Twitter and Facebook. Once text data from both social media platforms about universities were collated, topic modelling techniques such as Latent Dirichlet Allocation (LDA) and Structural Topic Modelling (STM) and a systematic literature reviewing technique called Algorithmic Document Sequencing (ADS) were used to triangulate the data and common matriculation decision factors were identified as topics which were then fed into the internal model of an AI-led chatbot. This chatbot collected quantitative and qualitative data around these topics from participants and ran a double-blind experiment without the interference of human researchers. Finally, the results of the experiment as well as the insights gained from the performance of the chatbot were analysed and discussed.

Aims and potential contributions of the thesis

The primary purpose of this study is to develop a methodology where AI can be used to collect qualitative and quantitative data from human participants and run an experiment (randomized controlled trial (RCT)) autonomously where human researcher intervention is eliminated. Within this framework, the AI oversees participant interactions and data collection, following predefined procedures and algorithms. This approach warrants consistency, facilitates real-time collection of large data from diverse sources and locations, and eliminates researcher bias in the form of interviewer effect, intervention, and influence by removing human interaction with participants. Another key purpose of this research is to explore topic modelling and systematic literature reviewing techniques to identify potential decision factors from social media, train the AI with them and observe the AI's performance in running the RCT and producing relevant output.

Factors affecting students' matriculation decision were the main domain of data to be investigated via AI-led interview-like chatbot surveys. Hence, in higher education context the main contribution of this study is to help higher education institutions understand (a) the factors that impact students' university choices on a global scale, and (b) the role of electronic word-of-mouth on social media platforms whilst considering a factor. Therefore, this study can make a significant contribution to the broader field of business and marketing analytics for higher education sector, as the insights drawn from the analysed data and experiment output can be applied to the cases of higher education institutions operating in Australia and many other countries around the globe. Through this domain, I aim to advance the understanding of how to design and train AI-based chatbots, which can effectually conduct double-blind online social experiments, allowing researchers from all disciplines to collect qualitative and quantitative data from large samples in a rigorous, efficient, and ethical way and produce results that are statistically reproducible, reliable, and generalizable.

Thesis methodology: The overall framework

Initially two studies were conducted to explore and establish the relationship between student preferences for universities and public engagement on Facebook [Chapter 1] and Twitter [Chapter 2]. Subsequently, topics out of unstructured textual data pertaining to universities from their Facebook pages and Twitter were modelled by using Latent Dirichlet Allocation (LDA) [Chapter 3] and Structural Topic Modelling (STM) [Chapter 4]. The identified matriculation decision themes from these two topic models were combined through methodological pluralism with the relevant output of a systematic literature reviewing technique called Algorithmic Document Sequencing (ADS) [Chapter 5]. The output of LDA, STM and ADS were combined via a methodological pluralist approach [Chapter 5] yielding identified yet untested matriculation decision factors. These factors, as shown in Figure 1, were used to construct the dialogue architecture of an AI-led chatbot which was deployed to conduct a double-blind experiment on participants recruited from Prolific to prove their impact on students' university choice and extract new factors for future modelling [Chapter 6].

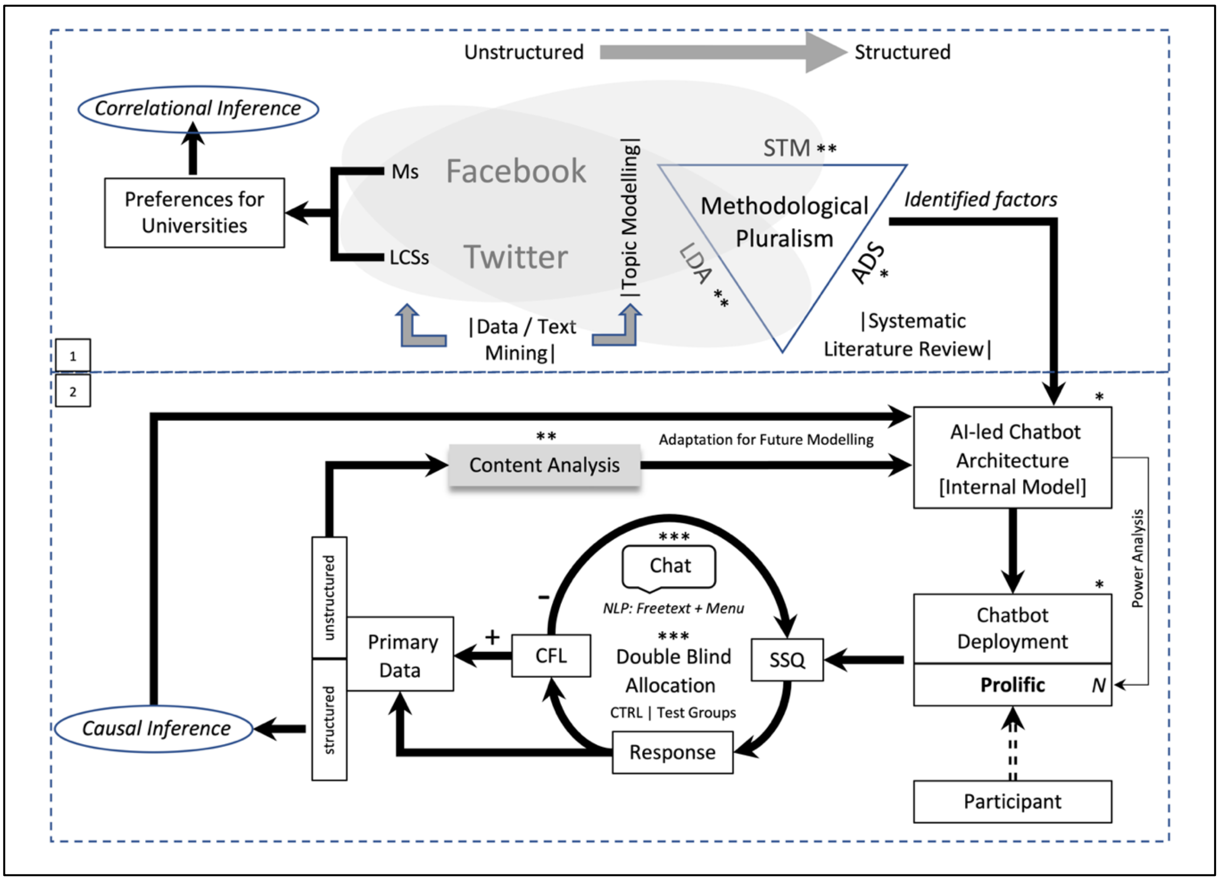

Figure 1. Flowchart of all processes including data collection from Facebook and Twitter, establishing correlational inference, topic identification via topic models (LDA and STM) and systematic literature review (ADS), initial training of AI-led chatbot (AILC) architecture, double-blind participant allocation, attaining causal inference through structured primary data and subsequent training of AILC with updated factors based on the initial experiment's structured and unstructured primary data.

Key to Figure 1. Ms: Mentions of university names on Twitter; LCSs: Likes, Comments and Shares of Facebook posts; LDA: Latent Dirichlet Allocation; STM: Structural Topic Modelling; ADS: Algorithmic Document Sequencing; CTRL: Control Group; SSQ: Semi-structured Questions; CFL: Confirmatory Feedback Loop; NLP: Natural Language Processing; *N*: Total number of participants being recruited from Prolific.
*: Supervised; **: Semi-supervised; ***: Unsupervised

This thesis is structured into four cohesive parts, encompassing a total of six chapters. The initial section (Part 1) offers a comprehensive overview of the thesis, delineating its objectives, scope, novelty, overarching methodology, and potential contributions. It also establishes the interconnection among its chapters. The second part comprises five chapters (Chapter 1-5) that progressively build upon each other, culminating in the third part (Chapter 6), which expounds on the benefits, novelty and features of a double-blind AI-based experiment and its related outcomes.

Chapters 1 through 4 represent published papers, each featuring its own literature review, methodology sections, and results. Chapter 5 consolidates the findings from the preceding chapters, employing the concept of Methodological Pluralism. The amalgamated results, specifically the factors influencing students' matriculation decisions, serve as the foundation for the AI-based chatbot utilized in the experimental framework presented in the sixth chapter. Accordingly, Figure 1 displays the overarching structure of the thesis. The upper section of the figure delineates Part 2, encapsulating chapters 1 through 5, while the lower segment, designated as Part 3, covers Chapter 6 elucidating the process of the double-blind experiment facilitated by an AI-led chatbot, which was trained using the combined output – namely, the identified decision factors – derived from Chapter 5.

The concluding part – the fourth and final section – engages in a comprehensive discussion and analysis of the experiment's results. This section also addresses the implications of the findings, acknowledges the limitations of the study, provides recommendations for future research endeavours, and ultimately concludes the thesis. The deliberate organization of these distinct parts ensures a logical and seamless progression of ideas throughout the entirety of the thesis.

Part 2

1. Establishing the relationship between social media data and preferences for universities, and topic identification from social media and literature

Chapter 1

1. The relationship between Twitter activity and student preferences for universities

1.1 Introduction

The competition for recruiting students has never been as technologically driven and severe as it is now. As universities have invested more in meeting and exceeding the expectations of students and communicated with them on digital platforms more effectively about their programs, student services, and graduate outcomes, the number of their student enrolments rose, as did their earnings from tuition and fees (Hisel & Pinion, 2020). They have become more student-centric as they adopted the "student is our client" philosophy and practices that include commercialization of the university brand.

Social media has reshaped the way organizations build long lasting relationships with their customers and provided marketers with convenient and affordable advertising and promotional opportunities (Wallace et al., 2011). Social media analysis has helped universities build their brand proactively, engage students with original conversations and create the content in which students are interested (Bolat & O'Sullivan, 2017). Therefore, understanding social media analytics, effectively connecting with prospective international (Bamberger et al., 2020) and domestic (Neagu et al., 2020) students on social media platforms, and using these platforms to attract potential students have become increasingly important, particularly for tuition-based universities around the world. Many universities around the globe have enjoyed the benefits of mass interaction with students (Veletsianos et al., 2017) as they have embraced social media platforms such as Twitter to commercialize and promote their offerings to new students (Shields, 2016). Since Twitter has become one of the most popular social media platforms, many organizations have been using it for their marketing and branding purposes. Although Twitter can be exploited to inflate a value offer and is considered less reliable than ordinary news blogs and newspapers, it still serves as a feedback mechanism which aggregates candid opinions of contributors about an organization, product, or service (Schmierbach & Oeldorf-Hirsch, 2012). Universities are no exception. Twitter allows for contributors through word-of-mouth to not only challenge, question and scrutinize the value of educational services offered by higher education intuitions but also promote them. This preliminary study builds upon the knowledge of strategic branding of universities in relation to whether their Twitter activity may be an indicator of performance in terms of student preferences and enrolment. Specifically, we aim to initially explore the association between Twitter activity and prospective student preferences for NSW and ACT Universities in Australia.

1.2 Background

1.2.1 Higher education and students' matriculation choice

Higher education is perceived by many as a vital, typically once-in-a-lifetime, distinct personal investment; a multifaceted and intangible service consumed mostly around 3-4 years (undergraduate degrees) (Walsh et al., 2015), and a choice probably to impact the life-long career of a student (Dunnett et al., 2012). Matriculation choice is defined as a relatively complicated decision process influenced by students, their environment and higher education institutions (HEIs) (Perry & Rumpf, 1984). It usually is a one-off and high credence purchase (Walsh et al., 2015) comprising a wide range of factors, particularly when, besides future concerns, current risks exist in relation but not limited to: time, psychological and social pressure, and, today arguably more than ever, intensifying financial burden on students and their families (Dunnett et al., 2012). Nurnberg et al. (2012) discuss that, although difficult to hypothesize and quantify, the investment utility of matriculation is one of the most profound antecedents of this decision process influenced predominantly by career opportunities and prospects – in terms of anticipated lifelong financial gain, work satisfaction, self-actualization, and status – accessible to a graduate of the HEI.

Dunnett et al. (2012) posited that the convenience of internet incorporating plethora of information (i.e., value propositions, features, statistics) about HEIs allows potential students to seek and evaluate multiple factors available at university comparison websites, forums, ranking tables, as well as at universities' own or affiliated websites and social media sites. Some of the decision factors used as proxies indicating the potential value of higher education offered by institutions are mainly their reputation, ranking, student satisfaction scores, facilities, WOM, tuition and fees (Dunnett et al., 2012). However, determining, weighing, and comparing the value propositions of HEIs is still a difficult task for many prospective students (Nurnberg et al., 2012) as the decision process encompasses a conclusion of the filtering of past and current quantitative and qualitative data pertaining to many features of HEIs through layers of predispositions impacted by individual priorities, family expectations, cultural norms and background, and subjective judgement (Briggs & Wilson, 2007).

Earlier studies devised various choice models exploring matriculation decision for students in terms of which HEI to enrol. One study, based upon Fishbein and Ajzen's Theory of Reasoned Action (TRA) (Fishbein & Ajzen, 1980) found that students usually follow their teachers' and parents' recommendations while recognizing the problem and evaluating alternatives during their university selection process (Moogan et al., 2002). Considering that higher education also entails an emotional and experiential decision, based upon Practice Theory (i.e., analysis of daily repetitive practices to understand cultural and social constructs), Allen (2002) introduced the "Fits-Like-A-Glove" (FLAG) framework suggesting that no matter how much external information exist, the matriculation decision will have come down to a "gut feeling" stemming from the holistic experience during students encounter with that HEI at its original place. Due to the additional pressure created by the intangibility, subjectivity, and complex nature of higher education and the uncertainty of what it might hold for the future, let alone a "correct" decision, prospective students go through a fairly difficult process to make a well-informed decision (Walsh et al., 2015).

Hemsley-Brown and Oplatka (2015) reviewed 75 studies that explored various factors linked to students' university choice. They argue that there is neither a "one-size-fits-all" approach nor a single pattern of choice behaviour for potential students whilst selecting a HEI. They posit that although some studies strive to find one deterministic list to explain the rational pattern of decision-making in this context, such an endeavour is difficult to actualize due to the highly varying individual circumstances of students (e.g., gender, race, social status, family income, proximity to home, school leaver vs mature, prior education and parental education, price sensitivity) and different institutional characteristics (e.g., prestige and reputation, fees, location, techno and sports facilities, admission process, student life on and off campus, course mode and flexibility, etc.).

After reviewing a two-decade literature on college choice in the USA, Hoyt and Brown (2003) found that the most prevalent and influential factors prospective students consider whilst selecting a university are its reputation, location, education quality, tuition and fees, scholarship availability and other financial aid, on campus employment opportunities, campus safety (particularly for female students), and flexibility in course delivery times and methods.

Another US-based study found the course and institutional reputation to be by far the most imperative features of HEIs shaping the college choice of prospective students regardless of their background and increasing tuition and fees (Dunnett et al., 2012). The study also found that in terms of establishing the overall utility linked to a college (HEI), course fees were relatively insignificant. Nevertheless, students whose parents had not received college education had less utility towards a college due to higher fees than the ones whose parents had attended college (Dunnett et al., 2012). Since parents' education status may indicate social class-based differences, Paulsen and John (2002) suggested that as a result of such differences, expectations and perceptions of students about tuition and fees may influence their college choice.

Walsh et al. (2015) emphasize that as frequently mentioned in extant literature, financial deliberations play an important role in students' college choices in the US, where students, however, are expected to pay for higher education with no lower or upper limit on tuition and fees enforced by the government. Ehrenberg (2020) posit that students who attempt to gain admission to high-cost selective HEIs believe that they make a reasonable economic decision; through which they gain long-term benefits in terms of higher probability of finding a better job with better pay after graduation and throughout their career. Bluntly put, Ehrenberg (2020) also argue that so long as more and more flocks of students keep applying to these selective HEIs, the highly decentralized US higher education system or any other market force will not put a cap on the rate of increase in fees.

In addition to tuition and fees, other US-based studies indicate that features of HEIs such as course and institution's academic reputation, modes of study, location, course information and content, online presence, and campus facilities (i.e., lecture halls and rooms, library, high-tech equipment, student clubs, sports venues), academic staff and faculty all play a role in students' college choice (Bergerson, 2009; Dunnett et al., 2012; Walsh et al., 2015).

One Scottish study found that WOM is more powerful shaping students' university choices than empirically-researched sources of information such as League Tables (Briggs & Wilson, 2007). The authors highlighted that HEIs need to invest in understanding the quality and quantity of information potential students seek and need to be able to make informed decisions. Adequately sourcing such information that mirrors the actual experience can help HEIs improve reputation in the higher education market and attain a competitive edge (Briggs & Wilson, 2007). In Italy, students' university choices were found to be influenced mainly by the job opportunities in HEI's region, proximity of HEI to home, reputation of the HEI and ease of access (i.e., satisfying admission requirements) (Azzone & Soncin, 2020). In a Portuguese study, geographical proximity (of the HEI to students' home city) stood out as the most important factor influencing prospective students' matriculation decision (Simoes & Soares, 2010). It was argued that students prefer to study at universities closer to their homes to reduce living expenses and maintain emotional attachment to family and friends. Academic reputation was found to be the second most important decision factor. The study also revealed that "former/current students of a university" and "university website" were the top two information sources for potential students whilst choosing a Portuguese university (Simoes & Soares, 2010). Chinese university admissions – subject to reforms in higher education system in the past two decades – have undergone a transition from a sequential to a parallel selection process allowing more students to get acceptance from highly ranked universities (Ashraf et al., 2017). In addition to the ranking of universities, whilst making their university choices, prospective students consider decision factors such as education quality, degree major, family expectations, and the city in which the university is located (Ashraf et al., 2017).

In order of importance, a Vietnamese study found that students choose universities based on (1) services and facilities (i.e., libraries, computer labs, lecture halls, health services, on-campus accommodation, etc.); (2) programme (i.e., course content, majors, credits); (3) price (i.e., tuition, fees, financial aid, scholarships, payment flexibility); (4) offline information (i.e., alumni contact, campus visits, recruitment advice and consultations); (5) opinions of family, teachers and friends; (6) online information (i.e., websites, social media and forums); (6) communications ways (i.e., e-mails, direct mail, phone calls); (7) premiums (i.e., class size, student diversity, availability of international student exchange programmes and distance-learning, etc.); and (8) mainstream advertising (i.e., TV, newspapers, magazines) (Mai Thi Ngoc & Thorpe, 2015). In Australia, one study found that the top four most imperative factors shaping Western Australian school-leavers' (i.e., high-school graduates') university selection were course type and suitability, employment prospects, academic reputation of the institution, and quality of the teaching staff (Soutar & Turner, 2002). Besides these four factors, authors of the study discussed that before enrolling a university, students also consider other factors such as distance of the institution from home, campus atmosphere, university type (modern or traditional), family advice and university choice of friends. Another Australian study found that prospective students seek information and consider eWOM on social media platforms about universities mostly in relation to reputation of the university and degree, job prospects, course difficulty, psychosocial life, and admission requirements (Le et al., 2019).

Although traditional WOM and eWOM, particularly generated by family and friends, are generally considered vital factors affecting students' decisions, Alfattal (2017) found that "recommendation from family" did not make to the top ten of the factors influencing domestic or international students' college choice in the USA. The top two factors for both student types identified in this study were (1) availability of major, and (2) affordability of costs. Academic reputation of the institution was the third and fifth most important factors for international and domestic students respectively (Alfattal, 2017).

A survey in New Zealand revealed that students' university choice was mostly affected by job prospects, course flexibility, degree quality, accommodation expenses and other costs associated with attending the university (Holdsworth & Nind, 2006). Likewise, other studies found that international students choose to study in host country universities to increase their chances of finding high-paying jobs upon graduation and having a rewarding career afterwards (Tas & Ergin, 2012). Yet, there is no consensus in literature as to what factors influence international students' decision more. Although many studies found that career is the strongest factor, one study revealed that 'personal growth' and developing 'English' as a second language were more important than 'career' as a decision factor for students who chose to study in US-based HEIs (Eder et al., 2010).

An Australian survey revealed that international students' choice of education destination is mostly influenced by the proximity of institution to students' home countries, the quality and variety of education, fees, and cost of living (Shanka et al., 2006). Although previous research indicated that a HEI's image including its beliefs, reputation and scholars have a strong impact on student's university choices (Erdmann, 1983; Bourke, 2000; Pampaloni, 2010), and bright international students are enticed to study in US-based universities because of their selectivity and global ranking (To et al., 2014), one study found that international students were driven neither by the selectivity of their HEI, nor by the HEI's rank or faculty's quality (Tan, 2015). Family and friend recommendations was the main reason, and the easiness of the acceptance and matriculation processes was the second most important reason for choosing their HEI. Furthermore, another study found that international students do not consider the accreditations such as AACSB whilst making a decision to study at a US-based institution (Daily et al., 2010).

A study that explored the study abroad decision factors of East Asian students reported that there are substantial differences in the purchasing pattern and reasons for studying abroad among different cohorts of students (Chen, 2008). The study demonstrated that whereas research graduate students are mostly encouraged by their professors to study abroad, non-research graduate students choose to study abroad to develop foreign language skills, get overseas work experience and improve their future career prospects. The study highlighted that the decisions for undergraduate students are predominantly shaped by their families.

A study comparing Germany to the UK in terms of international students' university choice factors found that 'education quality' was the prominent factor for students studying in both countries (Abbas et al., 2021). The study also demonstrated that international students studying in the UK perceived 'social engagement' as the second, whereas the ones having chosen the HEIs in Germany perceived 'job opportunities upon graduation' as the second and 'costs' as the third most important factor in their decision.

1.2.2 Impact of social media

Shields and Peruta (2019) stated that although students may not have reported in surveys that they visited social media platforms during their enrolment decisions, in interviews most of the students reported that they had used social media to gather more information about the HEIs under consideration. Nevertheless, in surveys students reported that social media had not influenced their decision-making.

Today, students have more alternatives in terms of higher education programs and providers than earlier generations (Retamosa et al., 2022). To entice potential students many HEIs strive to position themselves as education providers offering world-class education and academic excellence (Belanger et al., 2014). However, in a Canadian study, Steele (2008) found that students' selection of HEIs is based on more than just over-emphasized value propositions usually supported by metrics such as global rankings, but on emotional factors. In this study Canadian students classified HEIs as (1) Elite institutions with famous and successful alumni, (2) Outcome institutions with better internship and job opportunities after graduation, (3) Campus institutions with pretty campuses and better social/on-campus activities, (4) Nurturing institutions with smaller class sizes, and (5) Commodity institutions with unique value offerings as back-up.

1.2.3 Social media and social media marketing

In the early stages of the arrival of social media, companies were mostly sharing their content on social media platforms to increase the number of visitors to their websites and, particularly, sales (Jones et al., 2015). Marketing messages were generated by a relatively small number of companies to be received by a large cohort of potential customers (Alves et al., 2016). However, social media has evolved far beyond from being just a medium to share content. Evolution of social media goes back to the booming ease of convenience of the Internet and the advent of Web 2.0 (Power & Phillips-Wren, 2012). Progression of Web 2.0 enabled consumers to take an interactive part in the content creation and dissemination of websites allowing them to post various content such as text, audio, video, and simple reactions like “Like” (Jelassi & Martínez-López, 2020). Consumers started creating their own marketing messages - either positive, neutral or negative - in the form of reviews, blog or other social media channel comments, and mentions of brands, companies, products and services (Majid et al., 2019). In a way, social media has formed online communities made of users sharing not only content but opinions about content with one another over the worldwide web.

Today, many organizations use social media in a plethora of different ways. A company that intends to investigate what consumers think about its product or service would observe social media comments and engages with these people by responding to their comments (Buzeta et al., 2020). Companies use social media not only to engage customers but also to create their own content and publish them for promotional and advertorial purposes (Whiting & Deshpande, 2016). Through social media monitoring a company can also collect data about the expectations and preferences of its customers and acquire a multidimensional perspective of the market (Muninger et al., 2019) and a better understanding of current issues of concern (Cao & Sun, 2018). Accordingly, studies demonstrated that social media – as an informal source of information – present marketers a range of growth opportunities for their brands by helping them understand market trends, competitors' undertakings, feedbacks about customer satisfaction, as well as expectations, concerns, and desires of existing and potential customers (Arrigo et al., 2021). A company that is interested in understanding and gauging the level of its engagement on social media would evaluate performance metrics such as its reach, leads and sales by using a social media analytics tool (Orlandi et al., 2020). A company that aims to create brand awareness or deliver its value proposition to a particular set of audience would run a targeted social media advertising campaign (Dolega et al., 2021).

Social media marketing (SMM) is the use social media channels and social media networks to sell, advertise or promote a brand, product or service (Tuten & Solomon, 2017). SMM is a novel type of marketing that can significantly impact consumers at every phase of their purchasing process, from seeking information about brands to comparing their cost-benefit to post-purchase decisions (Di Pietro & Pantano, 2012). Therefore, SMM is imperative to marketing professionals who want to influence consumers at different stages of decision-making or want to steer consumers through the whole decision-making process (Tafesse & Wien, 2018b). Such engagement with customers can help marketers increase brand awareness, as well as listen to and understand the expectations, needs, desires and motivations of customers, allowing companies to make better products or improve upon their existing products or services (Hutter et al., 2013). This, in turn, can lead to higher sales, profits, and market capitalisation (Tuten & Solomon, 2017).

Consumers use social media mostly to communicate with others (Chen et al., 2017; Kwahk & Kim, 2017; Tafesse & Wien, 2018b). This communication is usually in the form entertainment, engagement (Chen et al., 2017) or learning (Prasad et al., 2019). Users of social media typically prefer to mingle with others of similar interest (Borges-Tiago et al., 2019). Since consumers put more weight to the ideas and comments coming from people in their online communities than company-generated messages including offers, promotions, warranties and even 100% money-back guarantees (Fang et al., 2021), peer-generated posts in social media usually in the form of advice and recommendations about a new or existing brand, product or service including pricing and customer support (e.g., post-purchase) have become a major source for people seeking information before and during their purchase decisions (Kwahk & Kim, 2017). Regardless of whether these posts are created by strangers or acquaintances (i.e., friends and family), social media users rely on them more than marketers (Tafesse & Wien, 2018b).

By freely expressing themselves, some social media users share their ideas, beliefs, past experiences, and behaviours influencing a much larger group of other users' ideas, beliefs, future experiences, and behaviours (Chen et al., 2017). The need to influence others through social media predicates primarily on three reasons: (1) To self-actualize themselves by giving something important or useful to the community (Kwahk & Kim, 2017), (2) To enhance self-esteem by improving social status and gaining popularity with the increasing number of subscriptions, positive comments and likes (Prasad et al., 2019), and (3) Financial gain by monetizing their personal social media channels (Tafesse & Wien, 2018a). As the number of social media users increase, particularly due to the convenience of mobile devices running social media applications, so does the average amount of time people spend on these platforms (Chen et al., 2017). As people spend more time on them, being engaged (i.e., connected) has become an essential part of their daily lives (Tafesse & Wien, 2018a). Companies are using SMM to manage their relationships with customers through a two-way channel. They often use social media platforms to deliver a message to customers, as well as to receive customer opinions and insights (Zhang et al., 2017). Customer relationship management (CRM) through social media may allow companies to identify matters of concern that may not have come to the attention of the management otherwise, alerting them to take action and meet the needs of customers (Harrigan et al., 2020). This, in turn, leads to satisfied, engaged and loyal customers sharing how their voices were heard by the company with other audience, willing to share even more about their experience with the brand, and organically creating more customer satisfaction, ownership and loyalty (Kim & Wang, 2019).

Since the primary motivation of companies for using social media is to generate leads and increase sales, their focus is on attracting new customers and retaining existing ones through social selling which includes little or no direct selling but allows marketers to create brand awareness, publicise the company and promote its offers, find new prospects and build long-lasting relationships with potential and existing customers (Terho et al., 2022). Accordingly, social selling cues such as posting the number of views a product has received and the number of customers who bought the product were found to impact customer purchase decisions (Das et al., 2021). Moreover, to amplify the positive outcomes of social selling, marketers often position themselves as product experts in pre-purchase stage (Ermer & Kleine, 2021).

SMM can help companies lower their cost of connecting and communicating with existing and potential customers (Lu & Miller, 2019). Customer engagement with companies and their brands can be positively impacted through social media and this interaction potentially has longer carryover effects than traditional offline marketing campaigns (Whiting & Deshpande, 2016). As consumers can engage with brands on social media in the long-term through a two-way channel which is inaccessible on traditional offline media, marketers are changing the way they convey their promotional messages on social media from how they were conventionally conveyed offline as 'pushing' – allowing consumers little or no opportunity to respond – to 'pulling' recognizing and valuing consumers as content creators, critics, and prosumers (Lu & Miller, 2019).

Besides plenty of opportunities, social media may present companies considerable threats. First and foremost, consumers may leave negative comments about a brand and if unhandled this can spread hurting brand image and future sales as it may impact the decisions of potential customers in an adverse way (Garcia-de los Salmones et al., 2021). Although some social media platforms allow companies to hide or even delete negative comments that appear on their sites, companies are recommended to handle such matters by constantly monitoring customer feedback on their feeds and take reasonable criticism seriously by responding to negative feedback with a public apology and provide a compromise resolution (Smith, 2022). Second, overt marketing and advertising content posted by marketers in SMM campaigns may be perceived by consumers as having no or little value, mobilizing merely a small portion of social media users to interact with such content and making most of them feel turned off and less connected to the brand (Alves et al., 2016). Finally, it may be difficult to appropriately assess the success of SMM and online promotional campaigns (Whiting & Deshpande, 2016).

Even if some social media platforms provide analytical information linking social media use to direct sales, this challenge still remains due to not knowing exactly why customer engagement and sales changed, as it may be as a result of such campaigns or an organic growth of the number of users (Alves et al., 2016). This uncertainty can significantly restrain budget allocations to SMM campaigns and even reduce the number of employees assigned to managing social media relationships with customers (Whiting & Deshpande, 2016).

Compared to traditional media marketing, since social media marketing is better geared towards reaching out target audience, it can allow marketers analyze consumer trends and behaviour more effectively about a product during the early stages of its life cycle (Schivinski & Dabrowski, 2016; Alalwan et al., 2017). Early monitoring and analysis of these trends can help marketers predict which products from their product range will have more demand so that they can develop a more comprehensive marketing campaign around them and coordinate with the procurement, manufacturing, and operations to meet future demand in time (Constantinides, 2014). This particularly is more important when the company operates in a fast-paced and competitive niche market (Rapp et al., 2013; Constantinides, 2014). When more resources are allocated to products predicted to sell more than other products, the company can gain a significant competitive advantage by launching and growing its product before the competitors do (Schultz, 2014; Alalwan et al., 2017).

1.2.3 Electronic Word-of-Mouth

In early literature, Westbrook (1987) defined word-of-mouth (WOM) as a type of communication informing other consumers about the ownership, features or usage of products or their pre- and post-purchasing experience with sellers. Research indicated that consumers consider WOM a more reliable source of information than traditional media such as radio, TV and print ads (Steffes & Burgee, 2009). Murray (1991) posited that consumers trust WOM to lower their perceived risk in their purchase decisions. Since consumers usually rely more on other consumers than sellers (Walsh & Mitchell, 2010), WOM can significantly impact the purchasing behaviour of buyers (Villanueva e al., 2008) and is regarded as one of the most powerful sources of information shaping the decision-making of consumers (Jalilvand & Samiei, 2012; Huete-Alcocer, 2017).

Internet facilitated online or electronic word-of-mouth (eWOM) is a form of WOM. Analogous to that of traditional (offline) WOM, the focus of eWOM (online) communication is the sharing of opinions – positive or negative – regarding consumers' past experiences with the usage of products or services (Steffes & Burgee, 2009). Although the online nature of eWOM in most settings reduces and, in some cases, completely abolishes the audience's ability to judge and ascertain the trustworthiness of the information providers and their comments, research repeatedly indicated that consumers rely heavily on eWOM in their decision making (Lopez & Sicilia, 2014; Yan et al., 2018).

1.2.4 Social media and eWOM

Social media technologies have enabled people to disseminate their opinions about products and experiences with services, generating substantial eWOM (Bilal et al., 2021) which has become a major source of information shaping the purchasing behaviour of online shoppers (Wu et al., 2018). Since consumers are putting their trust in eWOM generated through social media, marketers are striving to develop and maintain a three-way dialogue among social media contributors, their audience, and companies themselves (Alam & Khan, 2019). Studies revealed that purchasing decisions of consumers can be reshaped particularly by social media influencers and celebrities sharing content in the form of eWOM on social media and it is not unexpected that – more in the future than now – social media will keep influencing the decision-making processes of individuals and organizations that are connected through social networks (Cooley & Parks-Yancy, 2019). As the attractiveness of social media attains new heights, companies are recognizing that presence on social media platforms, monitoring eWOM and engagement with the public on these platforms have become crucial elements of their overall marketing strategy (Chae et al., 2020).

An individual's social media network is mostly founded on recruiting and retaining family members and friends (Ho & Ito, 2019). Hence, people communicating with one another through such networks already have an established trust within their internal community (Yahia et al., 2018) as the opinions of social media users can be impacted by the trustworthiness of the members of this community who share information regardless of the true source or nature of the information (Sterrett et al., 2019). Therefore, eWOM created from within social media are extensively acknowledged and bought-into by people who perceive the content mostly as reliable, bias-free and true (Choi et al., 2018). Nielsen market research revealed that 92% of consumers trust the suggestions of their family and friends over all sorts of advertising (Whitler, 2014). Research indicates that due to its persuasiveness, eWOM has more impact on consumer decision-making than conventional marketing forms, such as print ads (Yan & Wu, 2018). Discussions and comments posted about a brand by consumers on social media platforms have a significant effect on perceived brand image, and hence, should be monitored closely by marketing professionals to shape the current and future strategies for product, service and market development (Kubler et al., 2020).

Customers may speak positively or negatively about a brand, company, product, or service on social media. Company representatives monitor negative feedback and use that information to resolve the issue to build a 'caring' image and rapport with the customers and modify products or services for future use (Wardati & Mahendrawathi, 2019). Customers are more likely to leave a positive review or comment than a negative one (Ansari et al., 2019). Companies can also use such positive input in the form of eWOM to build upon their strengths and promote eWOM as testimonials integrated to their marketing campaigns (Chae et al., 2020). Since recommendations from users of a product are considered trustworthy for many new users, eWOM can have a substantial impact on purchase decisions (Choi et al., 2018; Ansari et al., 2019). Clark and Melancon (2013) found that when asked about the influence of social media on their purchasing decisions, 85% of the respondents admitted that they had bought a product based on the advice on blogs, and 60% of them decided to buy a product based on the recommendations on Facebook.

1.2.5 Social media engagement

Used by billions of people around the world, social media has become one of the vital technologies of the early 21st century (Appel et al., 2019). As the number of social media users increase, particularly due to the convenience of mobile devices running social media applications, so does the average amount of time people spend on these platforms (Chen et al., 2017). As people spend more time on them, being engaged (i.e., connected) has become an essential part of their daily lives (Tafesse & Wien, 2018b). Social media technologies have enabled people to disseminate their opinions about products and experiences with services, generating substantial electronic word-of-mouth (eWOM) (Bilal et al., 2021) which has become a major source of information shaping the purchasing behaviour of online shoppers (Wu et al., 2018).

The advent of social media triggered a paradigm change in online consumer behaviour, transforming the way consumers engage with brands, organizations, and each other (Nisar et al., 2019). The interactive features of social media platforms have changed consumers from silent readers of content to dynamic participants who now get involved in the creation of content and co-create value through two-way online communication channels (Van Asperen et al., 2018). This consumer-induced content and value creation can have a significant influence on improving brand awareness (Jamali & Khan, 2018), customer loyalty (Van Asperen et al., 2018), and sales (Garrido-Moreno et al., 2018). The impact of content and value creation may organically grow as more consumers exchange ideas, experiences, and emotions on social media through these interactions (Nisar et al., 2019).

Customer engagement on social media is defined as a behavioural process by which consumers create a sense of ownership as they approach becoming loyal to brand (Dolan et al., 2019) and is characterized by perceptive, responsive, and expressive states in their online interaction with the brand (Ajiboye et al., 2019). This interaction constitutes a form of behaviour reflection (De Vries & Carlson, 2014), and the focal point of this reflection is the brand itself (De Oliveira Santini, 2020). Such interactions reflecting consumer behaviour within social networks encompass creation of new content or contribution to existing content with a varying level of engagement (Ajiboye et al., 2019) from simply adding a quick remark such as “thumbs up” to writing a long and detailed comment or review about a brand or about someone else’s comment or review (Dolan et al., 2019).

Many companies have recognized that people have varying engagement levels with social media. Some social media users merely collect information about a brand before purchase, whereas others play a more active role as a contributor of eWOM in influencing others’ decisions (Moore & Lafreniere, 2020). Accordingly, Buzzetto-More (2013) defines six kinds of social media users that companies need to be aware of: (1) Creators: create and upload content, also known as influencers if capable of reaching out a large audience; (2) Critics: rate brands as well as post comments and reviews about them; (3) Collectors: read others’ input, collect information to make purchase decisions and allow others to see what purchases they made; (4) Joiners: keep a presence, but do little in terms of contributing to eWOM; (5) Spectators: occasionally monitor others' input and contribute to eWOM even less than Joiners; (6) In-actives: do not engage at all. Buzzetto-More (2013) argues that companies need to enable and empower Creators and Critics to post and share more influential material to drive sales and create more eWOM about the brand, support Collectors and Joiners and Spectators to become more active ambassadors of the brand, and finally encourage In-actives to increase their engagement levels.

Social media transformed many consumers from being mere listeners to active participants in the marketing process (Leong et al., 2021). This transformation created prosumers who are highly active and influential in steering the perception, behaviour, and opinion of other social media users towards not only brands, products, and services, but also political parties (Weeks et al., 2017). Since social media users are mostly bound within the confines of social media bubbles, they constantly receive messages from like-minded prosumers and feel pressured to conform to the feelings and opinions of the group (Sugihartati, 2020). Companies can identify and target these groups, develop emotional ties with the prosumers in them and organically turn them into brand evangelists (Anggraini, 2018).

Research has shown that when authentic, specific, and meaningful content is secretly shared with social media micro-influencers, stronger customer-brand relationships can be built with consumers as micro-influencers act as brand evangelists for companies (Pornsrimate & Khamwon, 2021). According to Savage (2012), brand evangelists are "believers" and "preachers" advocating what the brand stands for as they actively engage with consumers and incessantly disseminate positive eWOM across a wide range of social media platforms. Companies can decisively bolster this process in the backend by creating smart eWOM networks through which information is shared freely yet monitored closely and secretly and even manipulated by marketer as they create a meaningful cause for the consumers to believe in and let some of them become their next brand evangelist (Savage, 2012).

Companies are recommended to understand the motivating factors that make their customers visit their social media platforms (Bazi et al., 2020). Hence, analytically exploring the profiles and usage of their existing customers may help companies identify their needs, expectations, and motivations to build a strategic marketing plan around how to attract potential customers and sell through these mediums (Enginkaya and Yilmaz, 2014). It is recommended that companies should utilize customer-generated eWOM to drive sales instead of relying on company-generated content on social media (Leong et al., 2021). To attain this, as an SMM strategy, companies maintain close relationships with social media influencers and brand ambassadors, advocates or evangelists who provide advice and opinions about company's brands to their own network of people (Bazi et al., 2020). In addition to providing financial compensation to people with a large reach of potential customers, companies provide product discounts, gifts, and other forms of incentives to others who do not have a large reach but can still contribute to eWOM (Leong et al., 2021).

Since consumers are putting their trust in eWOM generated through social media, marketers are striving to develop and maintain a three-way dialogue among social media contributors, their audience and companies themselves (Alam & Khan, 2019). Studies revealed that purchasing decisions of consumers can be shaped particularly by social media influencers and celebrities sharing content in the form of eWOM on social media and it is expected that social media will keep influencing the decision-making processes of individuals and organizations that are connected through social networks (Cooley & Parks-Yancy, 2019).

Social media is transforming the way many companies build ongoing relationships with their customers as companies are recognizing the importance of social media presence, monitoring eWOM and engagement with the public on these platforms allowing their marketing teams to run effective and cost-efficient advertising campaigns (Chae et al., 2020). Tuition-based universities are no different as they treat students like clients (McDonald, 2019). As social media constitutes a productive and cost-efficient channel to reach large and targeted audience, it has become a vital component of the strategic marketing plan for many higher education institutions around the world (Brady, 2019). Universities have been utilizing social media analytics to identify relevant content for students and monitor engagement to build their brand. However, the integration of social media to increase student enrolments is not a straightforward process as higher education institutions have been battling with the urgency to align their rapidly changing SMM campaigns with their long-term conventional marketing strategies (Le et al., 2019). This urgency stems from the need to not only reach larger numbers of potential students but also provide cost-effective marketing at a time when controlling costs has become a matter of survival – particularly for universities that do not rely on the funding from government or donations (Paladan, 2018).

A global higher education-based digital marketing survey found that the level of engagement on Facebook between universities and prospective students was more than on any other social media platform (Leonard, 2018). The survey also revealed that international students in particular make their decisions solely based on their online experience with the institution, and most of them do not physically visit the campus before enrolling. The survey also pointed out that the higher education institutions that invest heavily in online marketing have higher numbers of new student enrolments. Nevertheless, merely 22% of the surveyed institutions have increased their investment in online marketing between 2016 and 2018. One study found that 89% of university admissions offices considered social media "somewhat important" and 55% "very important" to their future student recruitment strategy (Barnes & Mattson, 2008). Another study found that more than half of freshmen in US public universities reported that social media, and particularly Facebook, was conducive to their college decision making (Sandvig, 2016). Likewise, it was found that universities with considerably larger numbers of “Followers” and “Likes” on social media platforms such as Facebook and Twitter enjoy a higher student recruitment performance than the ones with lower numbers of Likes and Followers (Rutter et al., 2016).

1.2.6 Branding in higher education context

Literature that emphasized the commercialization of the university brand started to build up by the end of the 20th century when the higher education market has domestically and internationally become more and more competitive (Lomer et al., 2018). Research in this era was concentrated on concepts such as the commercial potential of academic research (Lowe, 1993), adopting new technologies to gain a competitive edge (Moran, 1998), and making funding arrangements with corporations (Rosenzweig, 1999). For the last two decades, however, researchers have been focusing on more specific aspects of commercializing the University brand as they explored the internationalization of higher education (Haigh, 2008; Guo & Chase, 2011; Knight & De Wit, 2018); brand identity construction (Lowrie, 2007), service quality issues (Vauterin et al., 2011), and ethical dilemma (Natale & Doran, 2012).

There are many similarities between how business entities and higher education institutions commercialize their brands (Mampaey et al., 2015). Since universities offer their students a service that comes with an intangible opportunity for a future career rather than a tangible consumer product, the real commercial value of their service is difficult if not impossible to measure. Many universities, however, attach this value to the elements that have accumulated over years such as their reputation, image, and the success of their alumni (Paul & Pradhan, 2019). Not unlike pure service-based businesses, by building upon these elements and creating a brand that stands out from the crowd in higher education, many universities have built a long-term competitive advantage over others (Dholakia & Acciardo, 2014; Chapleo, 2015).

In many countries such as the US, with the lessening federal and state funding, not only tuition-based private but many public higher education institutions (HEIs) are competing for recruiting new students (Matic, 2019). As a result, HEIs have become more student-centric treating students like a 'client' and implemented marketing strategies and practices that include commercialization of the university brand as they have become more involved in brand development activities to increase their recognizability and maintain a positive image in the minds of their existing and potential students. (Dennis et al., 2016) With an aim to elevate and sustain their market position in a rapidly evolving competitive environment, not unlike businesses, HEIs have attached various identifiable features such as a symbol, logo, tagline, and design to their value proposition so students can associate themselves with the institution brand (Hemsley-Brown et al., 2016). Using these features in advertisements have helped HEIs improve not only the recognizability but also the value of their education programs (Kaushal & Ali, 2020).

The commercialization of a university brand extends beyond national borders and can lead to substantial revenue gains. In Australia, for instance, higher education became the leading service export and the third largest export class overall, as it doubled in size from 17.5 billion AUD in 2013 to 37.6 billion AUD in 2019 (Hinton, 2020). As the Australian education minister highlighted, without the inclusion of fees and tuition earned from international students, Australian universities would have together incurred a loss of 7.7 billion AUD in 2019 (Hurst, 2020).

As the competition for tuition paying students intensifies and external funding becomes scarce, higher education institutions need to build and sustain an image that comes with a value proposition for prospective students (Adcroft et al., 2010). In addition, Ivy (2001) argued that universities are becoming progressively more determined in their marketing agenda to create a brand image and reputation that is favored and respected by the public. Due to the convenience and ease of distance learning the competition in global higher education market gets even more intense as more education providers invest in marketing activities and penetrate international markets (Armstrong, 2010). Researchers predicted student satisfaction with perceived brand image of higher education institution and revealed the strong association between institutional image and perceived value of the institution (Brown & Mazzarol, 2009).

As most public HEIs are considered not-for-profit, categorizing students as ‘clients’ for branding and marketing reasons has spawned much controversy in higher education sector (Naidoo, 2018). Nevertheless, most HEIs have embraced evolving market conditions and re-positioned themselves to develop a positive brand image to attain a competitive advantage (Panda et al., 2019). The main reasons behind HEIs’ branding endeavours were the recent and imminent decline in the demand (student enrolments and retention) and financial resources (Nguyen et al., 2019). The unchanged supply (current HEIs) and decreasing demand are regarded as the major reasons for the severe competition among HEIs (De Wit & Altbach, 2020). As studies indicated, particularly in North America, Europe and Australia, generation Z are less interested in studying a higher degree than their seniors, which has already reflected on the reduced number of enrolments (Schwieger & Ladwig, 2018). However, since HEIs with low acceptance rates (i.e., Ivy League) already have more demand than they can handle or prefer to handle (Musselin, 2018), and due to their global reputation and established brand, these universities were not impacted by the overall decline in the demand (Esteki & Kalati, 2021).

When started and supported from the inside out, marketing initiatives are more effective and sustainable in the long-term (Barros-Arrieta & Garcia-Cali, 2021). The concept of internal branding is crucial in higher education because it improves internal stakeholders’ (faculties, non-academic employees, managers, governing board, alumni, etc.) connection and identification with the HEI, empowering them to become brand champions and ambassadors who “live and represent the brand” (Clark et al., 2020). Thus, in an organizational culture supportive of internal branding, internal stakeholders can be more engaged in the co-creation of the meaning of the brand out of their experiences and build upon the existing value of the brand with their social interactions as they communicate the brand’s value proposition to customer at every encounter (Dean & Arroy-Gamez, 2016).

With an intention to use their brand identity to differentiate themselves from other institutions and develop a sense of ‘brand ownership’ (Leijerholt et al., 2019), some HEIs defined their positive attributes clearly and aligned them decisively with the expectations, preferences and challenges of their target audience, namely prospective students (Wu & Cheong, 2021). A study that investigated the university selection criteria of students in the US found that while strategic plans and initiatives of HEI branding need to be tailored to the type and circumstances of an institution, students from both and public and private universities consider reputation of the institution and cost of studying, and expect to receive a higher education experience that encompasses modern technology and a vibrant campus environment where they can participate in various community activities (Joseph et al., 2012).

An important aspect of the commercialization of a university is its ranking. Since the real value of a higher education program is intangible and therefore difficult to quantify and compare against others’, many students may refer to local and global ranking scores of universities for comparison before making an application. University rankings created and published by various media sources may shape the decisions of prospective higher education students as they tend to prefer a higher-ranked university to a lower-ranked one (Johnes, 2018). One study demonstrated that medical school applicants in Germany were more inclined to choose higher ranked institutions (Horstschräer, 2006). However, a Canadian survey (Drewes

& Michael, 2006) indicated that potential students would rather study at universities closer to their homes, and that offer more scholarship opportunities and a wider range of on- and off-campus student services.

Although many HEIs have integrated their institutional branding to their overall marketing strategy, they still face some challenges. One major challenge is that HEIs usually must overcome difficulties in conveying the necessity and urgency for institutional branding to internal stakeholders who may have developed some degree of resistance to change (Chapleo, 2015). If such resistance persists, often due to lack of top management support (Mampaey et al., 2020), the efficacy of marketing efforts put into achieving internal branding along with its associated benefits will fade (Kuoppakangas et al., 2020). To tackle this challenge, HEIs are recommended to get buy-in from and bring all internal stakeholders on board whilst building the brand and create a ‘ripple effect’ by reaching out as many members as possible with the ones who carry the flag forward along the way and support the agenda towards a more holistic internal branding campaign (Nguyen et al., 2019). Therefore, HEIs are recommended to create a sense of ‘ownership’ with their internal stakeholders and give them every opportunity to represent the university brand offline and online (Rutter et al., 2016). The notion of ownership aligns with the traditional branding approach adopted by many corporations (Leijerholt et al., 2019). Particularly, current students and alumni are considered ‘brand ambassadors’ who share HEIs’ key value propositions with target audiences through traditional WOM and eWOM (Belanger et al., 2014). Students enrolling in a course at a HEI is essentially commencing a lifelong relationship with that institution, as they will most likely associate that HEI’s name with their own (Chapleo, 2015).

Melewar et al. (2018) stated that the service focus of higher education makes institutional branding even more critical than for companies that sell tangible products. Belanger et al. (2014) argued that effective branding is essential for universities to create a differentiated value offer at organizational level rather than at the level of individual service categories, programs, or units many of which have different titles (e.g., Bachelor of Business, Bachelor of Commerce etc.) yet analogous offers. Besides at institutional level, higher education branding at national level is considered vital for countries such as the UK as higher education has transformed from a mere social service providing ‘public good’ to an industry that generates substantial profits for HEIs particularly from international students and significantly contributes to national employment (Lomer et al., 2018).

Altbach and de Wit (2020) posited that national systems and HEIs relying on the income generated by international students are suffering a substantial blow due to the Covid-19 outbreak causing mobility restrictions for international students on a global scale. As many HEIs quickly adapted to the new situation, they have upgraded their online course delivery infrastructure and established flexible study options incorporating a wider range of distance learning programs (Carnegie et al., 2021). However, regardless of how quickly the adaptation process was completed, due to economic disruptions and uncertainty in international students’ home countries, enrolment numbers, particularly for tuition-based HEIs, were negatively impacted (Watermeyer et al., 2021). Accordingly, these HEIs and national systems are recommended to reduce their dependability on one or two countries, such as India and China, and extend their institutional brand across borders covering more extensive geographies (Altbach & de Wit, 2020).

1.2.6 Social media marketing in higher education

Competition among higher education institutions in both local and global marketplaces to attract and recruit students has never been tougher (Calitz et al., 2022). Unlike a couple decades ago, prospective students today rarely obtain information about universities from brochures, letters or print media (Peruta & Shields, 2018). Although some universities may still rely on printing and distributing brochures for their overseas agents, throughout the past decade, and particularly since the Covid-19 outbreak, universities have realized the huge potential of social media to reach out wider audiences and deliver their value proposition to prospective students (Rutter et al., 2016). Indeed, many universities have been working on developing their SMM strategies – regularly updating their social media feeds with enticing content to build their brand through their internal and external audience (Peruta & Shields, 2018).

Social media have evolved into a global networking and information sharing tool with growing social and financial impact on communities (Berger et al., 2014). Particularly in marketing, social media analytics has provided companies with an insight to make better informed decisions about reaching out their target customer base (Misirlis & Vlachopoulou, 2018). Companies often use social media platforms to deliver a message to customers, as well as to receive customer opinions and insights (Zhang et al., 2017). Although prior research indicate that social media and social networks have become an essential domain in information systems (Stieglitz et al., 2014), and the effective use of social media analytics was positively associated with organizational performance (Nisat et al., 2019), rapidly changing technology has outpaced this line of research (Senadheera et al., 2017). Since a strong social media presence may lead to higher engagement with target audience, and brand loyalty, improving a higher education institution's social media presence may contribute to higher enrolment numbers (Constantinides & Stagno, 2012).

Over the past decade, as part of their overall marketing strategy, in addition to traditional avenues HEIs have been pursuing other modern marketing opportunities on borderless networking platforms, such as social media (Chugh & Ruhi, 2018). Many HEIs recognized the cost-effective and differentiating ways through such online avenues to build their brand (Belanger et al., 2014; Nguyen et al., 2019) and reach out both domestic (Le et al., 2019) and international markets (Lomer et al., 2018) for attracting and recruiting new students.

Whilst delivering news and other information about the university, HEIs use social media to create brand awareness, build up their institutional brand and promote their offerings (Nguyen et al., 2019). With the main objective of improving their market position in terms of student enrolments, donations, and grants, HEIs particularly in North America, Europe and Australia have been increasingly involved in social media networking (Belanger et al., 2014). HEIs also use social media to bolster peer-to-peer learning and interaction to enhance student experience and social life on and off-campus (Chugh & Ruhi, 2018). Some HEIs have also been increasingly supportive of connecting students to alumni to improve career outcomes particularly for their recent graduates (Nguyen et al., 2019).

Most universities have official accounts on popular social media platforms such as Facebook, Twitter, Instagram, and YouTube etc. Universities may implement two main strategies whilst using these platforms: (1) To provide news about the university, study programs, campus events, etc. and make announcements; (2) To increase interaction among the institution, internal stakeholders (e.g., students, parents, faculty) and external stakeholders (e.g., alumni, recruiters) (Belanger et al., 2014). Looking from an SMM perspective, researchers stated that content posted by HEIs under the first strategy are generated for and delivered to a broad audience expected to provide little or no input feedback in return, whereas the second strategy allows HEIs to target specific audience who may be willing to create positive eWOM and may even take up the role of brand ambassadors (Belanger et al., 2014).
As social media has become an integral intermediary for higher education institutions to communicate their messages with students, literature has also highlighted the importance of utilizing social media as a strategic tool to attract students. For instance, a theoretical social media-oriented communication model for universities was suggested by Zailskaite-Jakste and

Kuvykaite (2012) to promote their programs effectively on social media platforms. Another study by Alexa et al. (2012) demonstrated how universities can benefit from using social media to display an embracing and welcoming image as they communicate with their potential and current students via social media platforms. However, not only confusing but also untrue and misleading messages could also be communicated with students through social media by higher education institutions (Azmat et al., 2013).

Social media giants such as Facebook, Twitter, Instagram, and YouTube have reshaped the nature of marketing (Ashley & Tuten, 2014). They stated that organizations have started to take social media marketing more seriously than ever to achieve various marketing goals via their social media platforms with activities such as promotions, branding, customer service and relationship management, market research and data mining. A significant portion of internet users globally has been connected to social networks where they share their ideas and experiences through word-of-mouth and ultimately impact one another's perceptions about products, services, brands, organizations, and shape each other's purchasing behavior (Lund et al., 2018). Even though companies were aware of the potential benefits of social media marketing, due to its undetermined and uncertain return on investment outcomes (Hoffman & Fodor, 2010), they constantly had to review and reevaluate it on budget definition and allocation (Silva et al., 2020). Furthermore, research indicates the reluctance of many institutions to incorporate social media campaigns in their marketing mix, implement a consistent and compelling social media marketing strategy and allocate more budget for social media advertising (Dwivedi et al., 2015).

Among many reasons for which tuition-based HEIs are utilizing social media to communicate their brand, the main one is to compete for and attract potential students in local and international markets (Galan et al., 2015). Arguably one of the most challenging tasks SMM professionals pursuing this purpose in higher education is how to effectively communicate the university brand through social media with their target audience. Since prospective students are universities' focal point, it is vital for SMM professionals to not only communicate what students want to hear but also proactively post relevant information and respond to the comments and inquiries made by potential students (Pringle & Fritz, 2019). However, Belanger et al. (2014) argued that successfully developing a university brand on social media hinges on more than just 'being proactive', since posting unappealing, disengaging, irrelevant or too much sales-oriented content may be perceived as spam deterring potential students from engaging the content and creating positive eWOM.

In a highly competitive higher education market, institutions that regularly post more content than other HEIs are more likely to reach wider audience (Prabowo et al., 2020). Accordingly, one study found that the total number of Tweets posted about universities is strongly associated with the number of students that prefer to study at these institutions (Cingillioglu et al., 2021). However, it was argued that the quantity of posts may not necessarily drive engagement on social media and the quality of posts – in terms of relevance and significance – is considered a key factor affecting social media interaction (Clark et al., 2017). Nevertheless, determining the quality level in posts is not a straightforward process (Ajiboye et al., 2019). Furthermore, even if a post is commonly perceived as highly relevant and significant to promote the image and brand of the organization, it may not generate a high level of social engagement one would expect and vice versa (Guesalaga, 2016). For example, whereas a concise "All the best" message for New Year receives more than 400 likes, another post by the same HEI about a clean energy collaboration between the university and an industry pioneer may receive less than 20 likes on Facebook (Belanger et al., 2014). Hence, engaging a larger audience may depend on more than what the post content aims to convey, as particularly younger generations consider first and foremost their self-interest upon reading a post and may question "what's in it for me" and if they realize "not much", even if the post is about something that seems to make a vital difference to institutional branding, they may still say “so, what? Good for you” and do not engage (Chugh & Ruhi, 2018). Furthermore, Belanger et al. (2014) posited that just barely existing on social media “for the sake of it” can be as detrimental as having no social media presence at all. Therefore, universities need to allocate adequate resources to maintaining a strong social media presence and commit to providing an incessant two-way communication channel for their local and global target audience (Bamberger et al., 2020).

1.3 Research question of chapter 1

Many students are seeking information about the universities and their programs on social media before enrolling in them and share their content with others due to their past, current, or future association with these universities. Since student preferences can be a major indicator of student recruitment and enrolments, the dissemination of tweets mentioning a specific university can be an indicator of promoting the university brand to increase its student enrolments. Therefore, the primary research question that this preliminary study aims to answer is whether the quantity of Twitter content about a university relates to student preferences for that university.

To address this question, it is required to identify Twitter content that was relevant to promoting the university brand to attract students. Hence, I conducted a relevancy analysis on tweets and tested the accuracy of machine predicted classification outcome against the perceptive verdicts of human subjects. Finally, in accordance with the results of the Twitter relevancy analysis, I developed a moderated multiple linear regression (MMLR) model to provide an in-depth understanding of the variation in student preferences. At this final phase, the Global Ranking of universities (GloRank) and Total Tweets (Ttweets) were used as explanatory variables and universities' Group of Eight (Go8) status was included in the model as the moderating variable. Go8 member universities are the eight leading research-intensive public higher education institutions in Australia. The response variable in the model was Student Preferences (StPref) for universities.

1.4 Data collection

The data for the response variable – StPref – were collected from the open media release of the Universities Admissions Centre (UAC) which processes the admissions of most undergraduate programs offered by 13 higher education institutions located in the Australian Capital Territory (ACT) and New South Wales (NSW) in Australia. These universities were chosen because they accept students through the Universities Admissions Centre (UAC) of Australia, and every year UAC makes the statistics of student preferences for them publicly available. Full archives on the mentions about these universities on Twitter between 2017 and 2021 were obtained by using Twitter Developer API v.2 to conduct the Twitter relevancy analysis and subsequently determine the number of total tweets (Ttweets).

The university preference statistics provided by UAC include data from every year between 2017 and 2021 admissions for Commonwealth supported and domestic fee-paying courses for Australian students and exclude international applications. UAC has multiple offer-rounds each year as the admissions cycle is run year-round. Each year's data incorporate the number of student preferences recorded throughout the entire preceding year. The data contain the number of first and total preferences to all courses for each university, as well as the number of student enrolments out of the offers made by these higher education institutions. It should be noted that UAC allows applicants to select up to 5 course preferences. These courses may or may not be offered at the same university. For simplicity reasons, only the number of first preferences of students were considered for each university and used as the response variable (StPref).

Twitter data were collected using the Twitter Developer API v.2 which allows access to full historical archives that incorporate the mentions of the 13 universities. It should be noted that the outcomes of the student applications submitted in a particular year were recorded by UAC as the next year's student preferences for these universities. Therefore, the Twitter search spanned the years 2016-2020 to reflect the admission years 2017-2021. The search contained the full names of all NSW and ACT universities. Initially, we intended to include their acronyms to cover more mentions of a university, but soon we realized that since other universities around the world have been using the same 3-letter acronyms, to avoid overestimation issues, we included only the undisputed versions of a university's name such as "University of Canberra" and "Canberra University" for the University of Canberra and "Australian National University" for the Australian National University, but not UC, CU or ANU. While conducting the search, we realized that the mentions of the "University of Newcastle" in tweets were referring not only to the university in Australia but also to Newcastle University upon Tyne in England. Although these two institutions share the same name, they are entirely distinct from one another. To exclude the university in England, we slightly modified the keyword search for this university by adding "Australia" to its search syntax. We applied the same strategy to the search process of tweets for the University of New England. The other universities had no such issue. Keyword search was conducted for each university separately and the collected data were collated, stored, and then processed for further analysis.

1.5 Twitter analysis and tweet relevance

Since not all tweets about a higher education institution can be considered relevant to attracting students to increase enrolments, we intended to filter out the irrelevant tweets from the analysis. We did not search for the university names and keywords simultaneously in the tweets. Instead, we broke the whole process of analyzing tweet relevance down into three phases. First, we searched for and recorded the tweets that contained only the university names. Second, we built a lexicon of relevant terms from the university websites and third we assessed the appearance of lexicon terms in the tweets that we had collected.

After phase 1 (collecting and recording all tweets), we decided to define and locate string terms to differentiate between relevant and irrelevant tweets. Rather than manually identifying keywords, as was done in other research (Pringle & Fritz, 2019), we identified and counted the frequency of words used by universities for self-promotion on their own website. To determine the relevancy of each tweet that mentioned one of the 13 universities, we needed to identify relevant words and determine if each tweet contained these words. Relevant words would form a lexicon built with an aim to exclude irrelevant tweets. As we scraped the websites of all 13 universities, we examined webpages that had a direct link to either Homepage or About Us page. We collated the entire corpus of text from these webpages into a list and wrangled the list terms by converting them to lower case, stripping whitespace and removing stopwords, numbers, punctuation, as well as the names of the universities. We then further cleaned the list manually by removing some of the commonly used yet undifferentiating words such as city and state names.

We did not want one or two university website contents to dominate the term selection for our lexicon. Therefore, first we scraped the websites for each university separately and calculated the frequency of words used in them. The list for each university consisted of 500 words sorted in descending order (500 words were more than enough to cover all the relevant words used on these websites). We then collated them into one list by selecting the shared terms and removing the unshared ones. After the stemmed terms "stud" (study, studies, student, students) and "universit" (university, universities) the most frequently used term was "rank" which stemmed from "ranking", "ranked", and "rankings". We decided to omit words that we considered undifferentiating such as "university", "people", "part" and "need" even though they were frequently and commonly used on webpages and made it to the top 150 terms of the lexicon. Finally, we had 140 terms in total to be used as the differentiating tokens in our lexicon. Kimmons et al. (2021) counted common unigram and bigram occurrences of tokens to identify trends in Twitter text. Similarly, in our lexicon, we have included unigram string terms that were used frequently by universities to build upon and advertise their brand and promote their programs with an aim to attract students.

Since not only relevant but also irrelevant tweets might contain one of the 140 keywords, we used a technique called the Linear Term Counting (LTC): First, we developed a Gold Standard by manually selecting 10 relevant and 10 irrelevant tweets from each university list (260 tweets in total). Then we linearly counted the keyword terms that occurred in each tweet. We conducted 16 individual LTC tests in total to observe the impact of the changing number of keywords, as well as various loading of keywords on the matching accuracy with the Gold Standard. In the first LTC test (LTC1), we included a moderately low number of 76 terms and assigned a load between 3 and 0 to each term based on their rank with equal increments of 0.04. So, for example the top term "stud" had a loading of 3, the second term "rank" had 2.96 and the 76th term "gain" had 0 loading. For this initial test for relevancy identification, we established the cut-off threshold at 5 (as an initial step for further experimentation). After counting the occurrence of lexicon terms and factoring in their loadings, a tweet would be considered relevant if its total sum of term loadings was greater than 5. For example, if a tweet contained "develop" (ranked 17, load=2.36) and "international" (ranked 29; load=1.88), it would not be considered relevant. However, if it contained "develop" and "employ" (ranked 7; load=2.76), it would be considered relevant because the sum of the loadings of its two terms was greater than 5 (2.36+2.76). We could achieve 60% accuracy against the Gold Standard with LTC1 this way, however we achieved higher accuracies when term loadings and cut-off thresholds were altered. We achieved the highest accuracy of 66.5% with LTC15 when 140 terms were used with loadings of 3-0 (increments of 3/140=0.0214) at a cut-off point of 2.

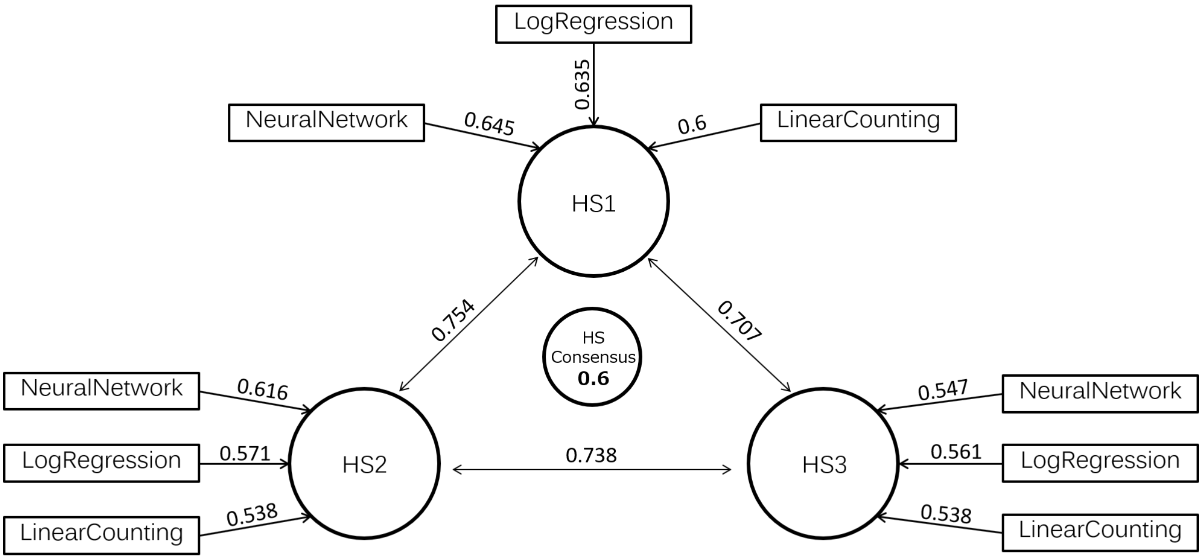

Figure 2. Classification accuracy scores between all venues of determination (3 Human Subjects: HS1, HS2, and HS3, & Linear Term Counting Technique and 2 Machine learning algorithms: Neural Network and Logistic Regression)

When we manually investigated the comparison outcome of the 260 tweets against the classification verdict of the LTC tests, we realized that numerous tweets were difficult to label as relevant or irrelevant because their messages could be perceived by some people as relevant but for others as irrelevant in terms of building-up the brand of a university to attract new students. For example, if a tweet mentions the discovery of a unique bacteria in the bark of an Australian tree by scientists of one of the universities, should this tweet be considered relevant or not? Another example would be the mention of a university team winning the national championship. Likewise, there are many tweets like "I played cricket on campus last week, it was so green and nice." Since some potential students may find it enticing for the campus to be green and nice-looking, whereas others will not mind, should such numerous mentions about a university be considered relevant or not? Even for a human being making the distinction between relevant and irrelevant in the context of attracting new students can be quite complex, subjective, and contradictory.

Since the Gold Standard was based on our own perception, it may be prone to observer bias. Therefore, we acquired two more Gold Standards by asking two other human subjects about whether they think each one of the 260 tweets was relevant to promoting or building up the name and brand of a university to attract potential students. Both human subjects were first-year university students studying at Australian higher education institutions. We addressed the relevancy classification verdict of these two human subjects as Human Subject 2 (HS2) and Human Subject 3 (HS3) and labelled our initial Gold Standard as Human Subject 1 (HS1).

When we applied the LTC tests on HS2 and HS3, we recorded a maximum accuracy of 53.8% for both. As suspected, this was a substantial drop from the maximum accuracy of HS1 (66.5%). Therefore, rather than linearly counting the number of keywords in each tweet, we decided to implement machine learning algorithms to predict the relevancy status of each tweet based on our lexicon terms. So, after a 70/30 data partitioning and 10-fold Monte Carlo cross-validation, we built Neural Network and Logistic Regression models that achieved an average accuracy of 0.602 and 0.589 respectively on predicting the classification verdicts of three human subjects. Although such accuracy scores seem low, when we measured the extent of overlap among the classification output of all human subjects between one another, as seen in Figure 2, we noticed similarly low average pairwise accuracy of 0.733 and a three-way human consensus of 0.6.

It was our initial intention to hypothesize that the number of relevant tweets that mention a university and the number of relevant words in these tweets have a positive relationship with student preferences for that university. Although substantial effort was put in the task of differentiating relevant tweets from irrelevant ones, we have empirically demonstrated through our Twitter relevance analysis including human subject classification comparisons that we could not determine with confidence whether a tweet was relevant to promoting and building up the university brand for attracting students. Therefore, there was no point in using ML or LCT to identify the relevant tweets or the number of relevant words in tweets. However, this does not mean that our original hypothesis is false. Nor does it mean that it is true. It simply means that due to human subjectivity, no matter how we train ML algorithms or recalibrate LCT, we have not been able to distinguish relevant tweets from irrelevant ones in a coherent and objective manner.

1.6 The model

As a result of the Twitter relevancy analysis, we opted not to attempt to identify relevant tweets due to the low accuracy of cohesion regarding what a relevant tweet is in this context. Instead, we counted the number of all tweets mentioning each university and recorded the sum as Total Tweets (Ttweets).

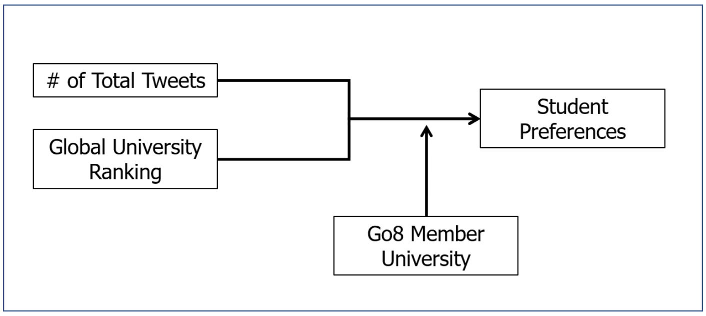

Figure 3. The Moderated Multiple Linear Regression (MMLR) model

Initially, we intended to use Global University Rankings (GloRank) and Go8 Membership as interactive variables between Ttweets and Student Preferences (StPref). However, we realized that a model in which GloRank was an explanatory variable, rather than a moderator variable between Ttweets and StPref, provided us with a better goodness of fit (in terms of R-squared) to explain the variation in StPref. Therefore, we introduced GloRank as an explanatory variable into our model with the objective of better understanding and predicting the variation in StPref. The inclusion of GloRank significantly enhanced the model's performance, as evidenced by a notable improvement in the goodness of fit. The fact that the model incorporating GloRank as an explanatory factor exhibiting a superior fit compared to a model without it, implies that GloRank may carry valuable information or may capture essential patterns that contribute to a more accurate explanation of the variations observed in StPref. Therefore, as shown in Figure 3, we have finally developed a Moderated Multiple Linear Regression (MMLR) model by including Ttweets and GloRank as explanatory variables, Go8 Membership as the moderator and StPref as the response variable. By including Go8 membership as a moderator, the model acknowledges that the explanatory or predictive influence of Ttweets and the GloRank variable on student preferences may vary depending on whether a university is a Go8 member or not. The interaction between Go8 membership and the explanatory variables (Ttweets and GloRank) allows for a more nuanced analysis taking into account whether such impact of Twitter activity and global ranking on student preferences in the context of Go8 membership is increased or diminished. This also facilitates the exploration of heterogeneity in the model as it recognizes that factors explaining or predicting student preferences may not be uniform across all universities and helps identify whether Go8 status introduces a significant moderating effect or not. With this model, in other words, we intend to explore to what extent the student preferences for universities can be predicted by Ttweets and GloRank moderated by whether a university is a member of Go8.

GloRank was determined based on the global university rankings published by the Times Higher Education. Go8 Universities in Australia are the University of Melbourne, University of Sydney (UniSyd), University of New South Wales (UNSW), Australian National University

(ANU), University of Queensland, University of Western Australia, Monash University, and the University of Adelaide. Since this study explores only the universities based in NSW and ACT, out of the 13 institutions, UniSyd, UNSW, and ANU were included as Go8 Universities and the remaining 10 institutions were recorded as non-Go8 Universities.

1.7 Results and discussion of chapter 1

After collecting and counting more than half a million tweets that mention the names of the ACT and NSW universities over a 5-year period between 2017 and 2021, as shown in Table 1, we have found a moderately strong positive relationship (r=0.594) between Student Preferences (StPref) and Total Tweets (Ttweets). The lowest recorded correlation, r=0.553, between the two was in 2018 and the highest, r=0.629, was in 2019 and 2021. Similarly, the correlations between the moderating variables, Group of Eight Membership (Go8) and Global Ranking (GloRank) and the outcome variable, StPref, were moderately strong. Overall, StPref had a positive moderate relationship (r=0.453) with Go8 and a negative moderate relationship (r=-0.64) with GloRank (lower numbers indicate higher ranking in the list of universities).

		StPref	*Go8*	*GloRank*	*Ttweets*
2017-2021	***StPref***	1			
	Go8	0.453	1		
	GloRank	-0.640	-0.552	1	
	Ttweets	0.594	0.812	-0.495	1
2021	***StPref***	1			
	Go8	0.437	1		
	GloRank	-0.543	-0.514	1	
	Ttweets	0.629	0.806	-0.467	1
2020	***StPref***	1			
	Go8	0.403	1		
	GloRank	-0.540	-0.643	1	
	Ttweets	0.574	0.814	-0.539	1
2019	***StPref***	1			
	Go8	0.523	1		
	GloRank	-0.644	-0.758	1	
	Ttweets	0.629	0.867	-0.671	1
2018	***StPref***	1			
	Go8	0.489	1		
	GloRank	-0.596	-0.819	1	
	Ttweets	0.553	0.771	-0.629	1
2017	***StPref***	1			
	Go8	0.435	1		
	GloRank	-0.581	-0.731	1	
	Ttweets	0.592	0.828	-0.601	1

Table 1: Correlations between all variables

The only strongly (>0.7) correlated variables were Go8 and Ttweets. Furthermore, Go8 and GloRank had a strong negative (<-0.7) association in 2017 (-0.731), 2018 (-0.819) and 2019 (-0.758). Finally, we have noticed no substantial discrepancies or fluctuations in the correlations between any variables over the 5-year period.

A vital discovery of the study was, as shown in Figure 4, that there was a strong positive relationship between Ttweets and StPref only for the universities that were globally ranked in the top 47-184 and are a member of Go8. For non-Go8 member universities, we noticed a strong positive association between Total Tweets and StPref only for the ones ranked between 275-450, and for other non-Go8 members ranked below 275 or above 450 we either noticed a weak positive association or a negative one. However, this negative association can be ignored due to its small size.

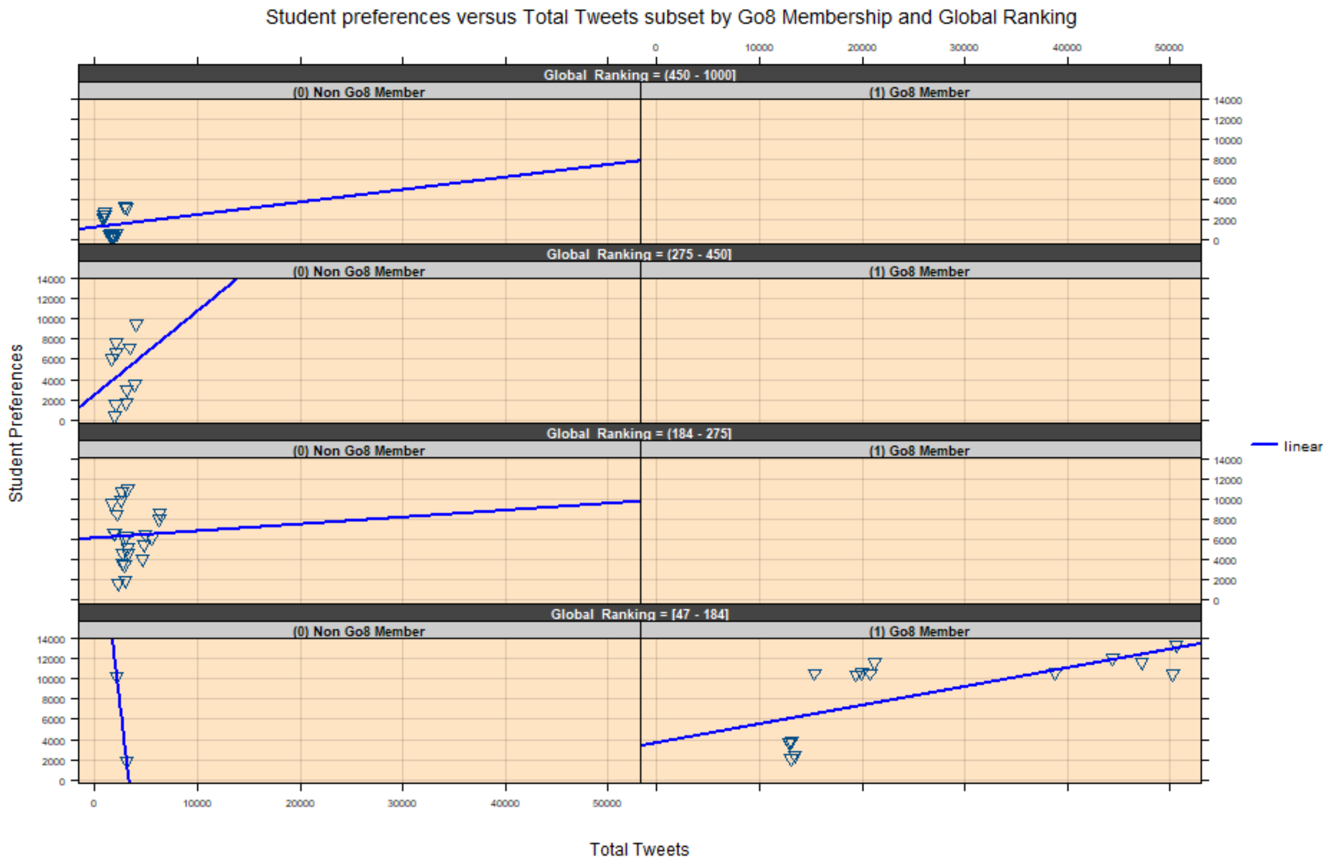

Figure 4: Student Preferences (StPref) versus Total Tweets (Ttweets) subset by Go8 Membership and Global Ranking (GloRank)

We adopted a six-step differential validity and prediction approach to perform the MMLR analysis. First, through simple linear regression we obtained evidence (AIC: 1233.7, Adj R-squared: 0.3425, p-value: 1.831e-07) of criteria related validity that Ttweets explains or predicts StPref. Second, we intercepted differences via multiple linear regression by including GloRank as a second explanatory variable in addition to Ttweets and found out that the model was a good fit (AIC: 1217.4, Adj R-squared: 0.4952, p-value: 2.335e-10) and the negative coefficient of GloRank (-5.63) was significantly different from 0, since GloRank had a p-value of 3.29e-05. Accordingly, we inferred that universities with smaller values of GloRank (higher ranked) were preferred more as first choice by students. Third, when we measured the slope differences in the model by using GloRank as the interactive term with Ttweets and performed a Simple Slopes Analysis, we noticed that when GloRank was at 62.62 (-1 SD), the moderation was statistically significant (<0.01). However, when it was at 373.66 (Mean), and 684.7 (+1 SD) GloRank could not be deemed statistically significant due to the p-values of 0.2 and 0.27 respectively. Therefore, we may infer that GloRank moderated the association between StPref and Ttweets only for the universities that had a low number of Global Ranking (highly ranked). Although the model was a good fit overall with a p-value of less than 0.01 and an adjusted R-squared value of 0.4944, we cannot infer that GloRank moderated the association between StPref and Ttweets because as an interactive term it had a p-value of 0.348. So, based on the analysis we have inferred from step 2 and 3, we realised the additive rather than the multiplicative contribution of GloRank to explaining StPref as it proved more useful as an explanatory or predictor variable along with Ttweets. Fourth, we intercepted differences in the multiple linear regression model but this time by replacing GloRank with Go8 which was used as an explanatory variable along with Ttweets. Although the model may be considered a moderately good fit (AIC: 1235.4, Adj R-squared: 0.3346, p-value: 1.224e-06), this was due to Ttweets not Go8 because the p-value of Go8 was 0.618, and the p-value of Ttweets was less than 0.001. In the fifth step, we estimated the slope differences in the model by using Go8 as the interactive term with Ttweets. The model yielded good fit results (AIC: 1230.9, Adj R-squared: 0.3879, p-value: 3.026e-07) and since Go8 had a p-value of 0.014, we noticed that rather than an explanatory or predictor variable for StPref, Go8 functions better as a moderator variable that moderates the association between StPref and Ttweets.

Finally, in the sixth step we tested the goodness of fit for the model where Ttweets and GloRank were used as explanatory variables and Go8 was used as the moderator between Ttweets, GloRank and StPref. Since the model yielded the highest adjusted R-squared value of 0.593 with a p-value of 1.758e-11 and a p-value of 0 for the interaction coefficient of Go8 on Ttweets and GloRank, we conclude that, as shown in Figure 4, Go8 strengthens the statistically significant association that Ttweets and GloRank together have with StPref.

Social networking sites such as Twitter produce huge amount of data reaching petabytes daily (Khan et al., 2017) and researchers are allowed for the time being (2021) to drill down into Twitter data and make inquiries about the patterns, sentiments, and trends. Therefore, gaining insights about social phenomena through big data is no longer out of reach. However, let alone for AI even for humans there is limited consistency in determining whether a message conveyed in a tweet contributes to promoting and building up a university brand to attract students. After comparing the classification verdict of relevant tweet identification performed through machine learning and LCT with that of human subjects, we realized where humans could not have consensus, it would be unwise to expect a machine to produce a valid classification outcome that could be generalized. Therefore, as initially hypothesized our strategy of using the number of relevant tweets and the number of relevant words in tweets for each university as explanatory variables would not be pertinent to be associated with student preferences for universities.

In the literature, besides other social media platforms, Twitter was used as an explanatory or predictor variable to explore the extent of student engagement in class and online, and performance in terms of grades. For example, an experimental study by Junco, Heiberger and Loken (2010) revealed that activities such as social networking and micro-blogging on Twitter had an impact on student engagement and grades in higher education. However, no previous study had such an experimental design that explored the causal relationship between social media interaction and student enrolments. Even if we used the most sophisticated statistical models, due to the lack of an experimental design and the fact that student preferences and the level of social media activity may have profound unobserved confounders, we could not explore the causal but only correlational relationship between Twitter activity and student preferences for universities.

Although we cannot suggest that social media impacts student preferences, this does not mean that we cannot predict student preferences for these universities with the level of Twitter content mentioning these institutions, since we have found a strong enough association between student preferences and Twitter activity. Being able to predict future student enrolment numbers this way and being aware of this association may help higher education institutions not only take the necessary actions and implement the strategies that will increase the extent of their interaction with prospective students on social media platforms such as Twitter with an aim to increase student enrolments, but also better prepare themselves in terms of budgeting and scheduling for the upcoming student intakes.

1.8 Conclusion of chapter 1

Our results lead to several vital theoretical and practical implications. Firstly, after having compared the classification verdicts of three human subjects with one another and against that of machine learning algorithms, we could not make a probabilistically high accurate distinction between relevant and irrelevant tweets due to human subjectivity. Therefore, we conclude that it was not possible to determine whether the users (readers, posters) of Tweets were interacting with the institution. This outcome significantly contradicts previous research (Kim & Ko, 2012) that claimed Twitter interaction increases attention and affection towards brands. Secondly, as a result of not being able to determine the real meaning or value of a tweet that contributes to the brand promotion of an institution, we used total number of tweets that mention universities and found out that there was a strong positive association between the number of total tweets and student preferences particularly for universities that were highly ranked. This outcome opposes previous research (Rutter et al., 2016) which stated that “a large number of tweets is not a predictor of performance”. In fact, it is a predictor of performance in terms of student recruitment, however we cannot and should not say that it affects student preferences or enrolments. Thirdly, our findings indicate that global ranking has a statistically significant positive relationship with student preferences and this relationship is strengthened by the Group of Eight membership status of a university in Australia. Finally, as opposed to previous research (Kietzmann et al., 2011) which posited that an institution would increase student recruitment by being more active on social media than another one with similar ranking, prestige and reputation, we conclude that the universities that are not highly ranked in the world should not rely on Twitter activity for attracting new students, but universities may use the level of their Twitter activity, global ranking and prestige to predict the number of students that will choose to study their higher education programs.

1.9 Limitations of chapter 1 study

Arguably the greatest limitation of this preliminary study was not knowing whether the tweets have been read by prospective students. Any single tweet may potentially have the power to build up the university brand name to attract new students, but we cannot know for sure if it was actually shared or read by prospective students, because we do not have any information regarding any identification details of Twitter users. Simply put, just because there are many tweets written about a university does not mean that they reach target audience.

Another limitation of this study was that we could not include all mentions of universities due to not being able to include the acronyms for them. For example, if we used "Usyd" in our search, we would have covered many more mentions for the University of Sydney. However, we opted not to do that because for keywords such as ACU, ANU, and UTS, the data may have nothing to do with the corresponding university since the same acronyms have been used by and for other entities. "CSU" acronym for example is used not only for Charles Sturt University, but also for California State University in the USA.

Chapter 2

2. Facebook engagement and student preferences for universities

2.1 Introduction

To strengthen the previous study, in this chapter I aim to establish the relationship between university preferences and social media data but this time by focusing on another popular social media platform – Facebook. Social media marketing on Facebook is considered imperative to marketing professionals who aspire to shape consumers' behaviour at different stages of decision-making or want to steer consumers through the whole decision-making process (Tafesse & Wien, 2018a). Increasing customer engagement on Facebook can help marketers enhance brand awareness, as well as listen to and understand the expectations, needs, desires and motivations of customers, allowing companies to make better products or improve upon their existing products or services (Hutter et al., 2013). Moreover, social selling cues such as posting the number of views a product had received and the number of customers who had bought the product were found to impact customer purchase decisions (Das et al., 2021). In higher education context, while earlier studies have shown the use of social media platforms such as Facebook for improving student engagement in online (Hoi et al., 2021) and face-to-face educational settings (Junco, 2012; Dyson et al., 2015; Datu et al., 2018) no studies have investigated the association between universities' level of Facebook engagement with student enrolments. To address this gap, based on the number of posts' Likes, Comments and Shares, we evaluate the level of engagement on universities' official Facebook sites and explore the relationship between Facebook engagement and student preferences for these universities. We found out that the Facebook posts that receive a large number of Shares and Comments are a strong indicator of students' university choices. When total Likes, Comments and Shares of Facebook posts determine Facebook engagement, and either Global Ranking or Group of Eight membership status of universities are included in models along with Facebook engagement, the number of student preferences for universities can be better predicted. We finally discuss the theoretical and practical implications for universities to analytically measuring Facebook engagement and predicting future student enrolments.

2.2 Facebook Engagement

Bonding and identity-based brand attachment drives customers to engage with brands on an organization's Facebook page (Hinson et al., 2019). Accordingly, Peruta and Shields (2018) posited that by building strong attachment with internal and external communities, higher education institutions can create a reputable image through higher Facebook engagement. This, in turn, can increase student enrolments and retention. When Facebook users develop an attachment with a higher education institution (HEI), they tend to create content about the institution on their own (Lund, 2019). Such user-generated content (UGC) is considered more reliable by Gen-Z audience than the content posted by the brand's internal stakeholders (Goldring & Azab, 2021).

It was argued in extant literature that there is not enough evidence to determine that the decisions of prospective students have been affected by the content posted on the official social media sites of universities (Nyangau & Bado, 2012). Similarly, Fuciu and Gorski (2013) studied Romanian high school students and found that although all students that participated in their survey had Facebook accounts, they showed little interest in using Facebook to search for information. However, students were influenced by advertisements on Facebook rather than by electronic word-of-mouth. Similarly, one study found that Turkish undergraduate students rarely use Facebook for educational purposes (Baglione et al., 2012). Furthermore, another study revealed that most Facebook users between the ages of 18 and 25 were disinterested in creating any form of connection with advertisers on Facebook and that almost no advertising worked on this demographic (Sashittal et al., 2012). However, a decade later, company statements of Facebook Inc. show that Facebook's annual revenue soared from $3.7 billion in 2011 to $117.9 billion in 2021 (Rodriguez, 2021) with the vast majority of Facebook's annual revenue generated by its advertising stream (Isaac, 2021).

Universities create and post Facebook content on their official or affiliated sites by adopting different strategies. One is that HEIs post events as they happen or periodically in a news format, while another HEIs deliberately use a prearranged activity schedule upon which they disseminate strategic and time-sensitive messages to their audience (Peruta & Shields, 2018). Many HEIs strategically tailor their content to deliver key messages to different segments of their target audience such as current students (for retention), prospective students (for recruitment), alumni (for word-of-mouth) and other external stakeholders (Le et al., 2019). Many students and alumni identify with their HEIs and take pride in accomplishment stories associated with the school and share that content with their social network (Kumar & Nanda, 2019). Therefore, university marketing teams often look for, develop, and post such stories to boost engagement on Facebook (Lund, 2019).

Although previous studies demonstrated the value of creating and maintaining an active Facebook site for image and brand building, and customer retention and attraction, no research hitherto has examined the link between Facebook Engagement (based on posts' Likes, Comments and Shares) and student preferences for universities.

2.3 Research questions of chapter 2

Studies have found a strong link between purchase intent and social media engagement (Coursaris et al., 2016; Labrecque et al., 2020). For Facebook in particular, Yoon et al. (2018) revealed that the number of comments about a company on Facebook is positively associated with company revenue and Hutter et al. (2013) found that consumer engagement on a brand's Facebook page positively impacts purchase intention, whereas over-exposure to direct marketing and advertising material negatively affects eWOM and consumer engagement in general. It was also revealed that measuring consumer engagement on social media can help companies predict consumers' purchase decisions and future sales (Garg et al., 2020) (purchase decisions and future sales are analogous to students' university preferences and future enrolments respectively in higher education context). Although consumer engagement on social media has been explored and measured through analytical data and insights drawn from the number of "Likes" (Coursaris et al., 2016), "Comments" (Yoon et al., 2018) and "Shares" (Malhotra et al., 2013), no study has investigated the link between a combination of these consumer engagement indicators with the institution choices of students in higher education sector. Many prospective students seek information about the HEIs and their value propositions on social media before submitting their applications. Since student preferences for one university over another can be indicative of future enrolment numbers, the extent of public engagement on a university's Facebook site can also be a determinant of promoting university image to targeted audiences to increase enrolments. However, existing research has not addressed which publicly available quantitative engagement indicators (i.e., Likes, Comments and Shares), or a combination of these indicators play a more effective role in determining student preferences. Hence, building upon what earlier studies have already uncovered, we ask the following research questions (RQs) to address this gap in extant literature:

RQ1: Can social media engagement for a university brand on a popular platform such as Facebook be determined by the number of “Likes”, “Comments” and “Shares” posted on that university’s Facebook site?

RQ2: Which of the indicators of Facebook engagement or a combination of them perform better in terms of explaining the variation in and therefore predicting student preferences for universities?

RQ3: How would different predictive models incorporating these engagement indicators perform along with the Ranking and the Group of Eight membership status of universities in terms of estimating the number of student preferences each year?

To address these questions, we collected all Facebook posts on universities’ official sites and built Structural Equation Models to better understand the factors leading to and associated with Facebook engagement and student preferences for universities. From a theoretical perspective the findings of this study may improve our understanding of how social media engagement can be quantitatively evaluated for the purpose of predicting the university choices of prospective students. From a practical perspective this may help universities not only improve their income by attracting more students to their academic programs but also develop an appropriate financial plan for budgeting and resource allocation in advance. Moreover, besides the brand managers in higher education sector, this study may potentially help the marketing professionals in other sectors help understand how social media analytics can be utilized to predict customer trends, purchase intent and organizational performance.

2.4 Methodology of chapter 2

2.4.1 Data collection

We used R software version 3.5.3 along with Octoparse to scrape and extract public posts data belonging to each university from Facebook Social Graph. We customized XPath settings of Octoparse on Scroll Page Loop Mode and extracted the text, and the number of Likes, Comments and Shares (LCS) of all posts available on the official Facebook sites of 13 universities[1] in New South Wales (NSW) and Australian Capital Territory (ACT). We then filtered and grouped the entire data by year for each university. LCS were used as observed exogenous variables loading on and leading to the latent exogenous (independent) variable: Facebook Engagement.

The data for the observed endogenous (dependent) variables – First Preferences and Total Preferences – were gathered from the publicly available Universities Admissions Centre (UAC) media release [https://www.uac.edu.au/media-centre/media-releases]. These higher education institutions were selected in this study because UAC officially administers the pre-enrolment process of students on behalf of all these institutions located only in NSW and ACT and shares student preference statistics in terms of "First Preferences" and "Total Preferences" every year between 2017 and 2022 for domestic fee paying and Commonwealth supported places excluding international admissions. Every year each new applicant is allowed to select maximum 5 courses which may or may not be offered by the same higher education institution. The results of student application rounds made in a year were represented with a cumulative number in the following year's first and total student preferences for these institutions. Hence,

[1] Australian Catholic University, Australian National University, Charles Sturt University, Macquarie University, Southern Cross University, University of Canberra, University of New England, University of Newcastle, University of Sydney, University of Technology Sydney, University of Wollongong, University of New South Wales, Western Sydney University.

we scraped all Facebook posts spanning 2016-2021 to account for the UAC student admissions between 2017 and 2022.

The data for the observed exogenous (independent) variable – Global Rankings of universities (GloRank) – were recorded in accordance with the official rankings established by Times Higher Education [https://www.timeshighereducation.com/world-university-rankings]. The other observed exogenous variable – Go8 members[2] – consist of the 8 research-intensive and some of the largest and oldest public universities in Australia. Since this study investigates the HEIs based in NSW and ACT, out of the 13 universities, University of Sydney, University of New South Wales, and Australian National University were labeled as Go8 members and the other ten universities were included as non-Go8 members.

2.4.2 The models

Structural equation modelling

We used Structural Equation Modelling (SEM) to uncover latent (unobserved) variables out of measured (observed) variables and construct a structural model that explains the connections between measured and latent independent (exogenous) and latent dependent (endogenous) variables. Accordingly, we specified our model with two components: (1) measurement model through which we established causal relationships between measured and latent variables, and (2) structural model within which we demonstrated regression-based correlational relationships between exogenous and endogenous variables. To draw the path diagram for the default model, as shown in Figure 5, we used SPSS Statistics v.28 and AMOS v.26.

[2] University of Adelaide, University of New South Wales, University of Queensland, University of Melbourne, University of Sydney, Monash University, Australian National University, and University of Western Australia.

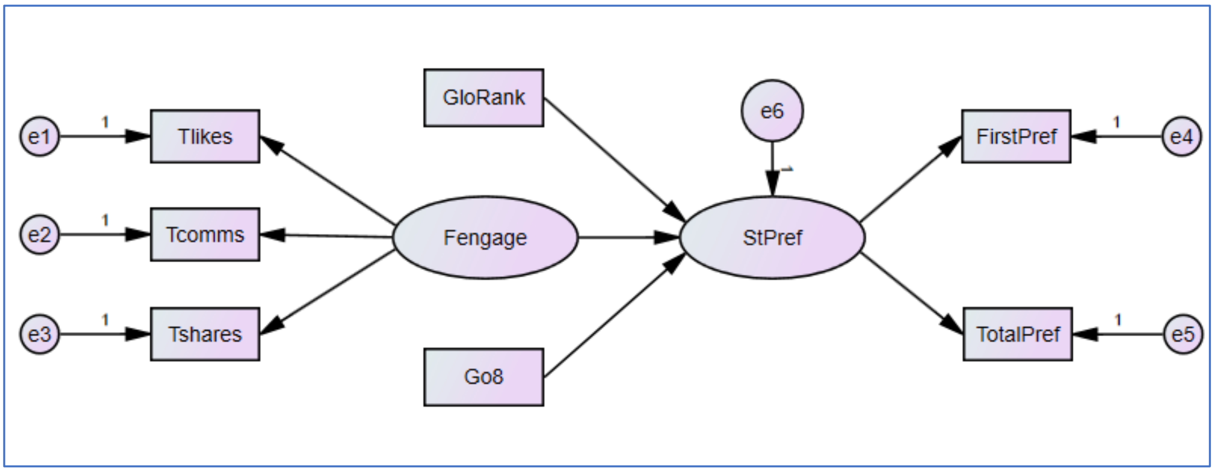

Figure 5. The default Structural Equation Model.

Three observed exogenous variables – Tlikes (Total Likes) & Tcomms (Total Comments) & Tshares (Total Shares) – are loaded onto the latent exogenous variable: Facebook Engagement (Fengage). Additionally, two observed endogenous variables – FirstPref (First Preferences) & TotalPref (Total Preferences) – are loaded onto the latent endogenous variable: StPref (Student Preferences). Furthermore, two observed exogenous variables – GloRank (Global Ranking) and Go8 (Group of Eight membership) – are included in the model along with Fengage to provide a better explanation of the variation in StPref. Accordingly, Fengage has been conceptualized as a formative construct rather than a reflective one (Coltman et al. 2008), where likes, comments, and shares don't just measure, but in fact constitute the essence of engagement.

Nested Models

In SEM, to analyze the nested models based on the default model, since the data were normally distributed per Shapiro-Wilks normality parametric test, we used Maximum Likelihood (ML) estimation as we did not directly analyze the raw data but the variance/covariance matrix of the observed variables. Therefore, we summarized the var/covar matrix of the observed variables by specifying a simpler underlying structure which embodies the structural equation models. By employing lavaan package on R, we generated 27 models in total using a combination of endogenous and exogenous variables in the dataset and subsequently conducted diagnostic estimations as to which constructs of each model generated inconsistency between the measurement model and the data.

As the output we derived from lavaan included var/covar matrices of the estimated associations among variables, we statistically analyzed how analogous the predicted data were to the var/covar matrices incorporating the true associations among the variables in the actual dataset. Accordingly, we used comparative measures such as CFI, RMSEA (Root Mean Square Error of Approximation) and SRMR (Standardized Root Mean Square Residual) to analyze the discrepancies between the actual data and hypothesized models, allowing us to determine the goodness of fit for each of the 27 models. Finally, we used Chi-squared difference test, BIC (Bayesian Information Criterion), and AIC (Akaike Information Criterion) measures within ANOVA to evaluate, compare and sort the models in terms of fit.

2.5 Analysis of chapter 2

2.5.1 Initial analysis

As a preliminary analysis, we retrieved the number of total Likes and Followers directly from each university's Facebook page in around mid-December 2021. We noticed that the correlation between 2021 Rankings of universities and their corresponding Likes and Followers was moderately strong (≈-0.55). Likewise, the correlation between the Go8 membership of universities and their corresponding Likes and Followers was also moderately strong (≈ 0.49). More importantly, as seen in Table 2, there was a strong correlation (≈ 0.7) between Student Preferences in 2022 admissions and the total number of both Likes and Followers for these higher education institutions.

	Go8	*2021Ranking*	*Likes*	*Followers*	*StPref_2022*
Go8	1				
2021Ranking	-0.5145	1			
Likes	0.4873	-0.5492	1		
Followers	0.4969	-0.5561	0.9999	1	
StPref_2022	0.4489	-0.5654	0.6913	0.6945	1

Table 2: Preliminary analysis results: Correlations between total Likes, Followers, 2021 Ranking, Group of Eight membership of universities and Student Preferences for them.

2.5.2 Main analysis

In total, we recorded approximately 29 million Likes, 1.5 million Comments and 663,000 Shares from the posts on the official Facebook sites of 13 universities between 2016 and 2021. After conducting a relevancy analysis using machine learning (ML) models and manually investigating the randomly selected 260 Facebook posts, we realized that determining the relevancy of many Facebook posts was a challenging task because the messages conveyed by these posts could be perceived by some as relevant, but others may consider them irrelevant in terms of building-up a university's image and reputation to attract and recruit potential students. For instance, if a Facebook post mentions that a university "sits on the ancestral lands of Australia's First Peoples, where we come together as one Sydney, and many peoples, to continue to share knowledge", should this post be considered relevant? Furthermore, there are many Facebook posts about the landmarks and natural beauties (sky, trees, animals, etc.) on university campuses. Whereas some prospective students may find such features desirable, others will not mind. Thus, differentiating between relevant and irrelevant Facebook posts in terms of enticing new students can be relatively complicated and subjective. Due to human subjectivity, regardless of what ML models we implemented, relevant Facebook posts could not be distinguished from irrelevant ones. As a result, we used the entire Facebook posts in the upcoming stages of analysis.

Rather than discovering causal relationships, an SEM should be based upon the researcher's causal assumptions as the reliability of its results depends on causal assumptions among variables (Bollen & Pearl, 2013). Therefore, we should note that although some applications of SEM are entirely aimed at explaining causal connections, due to our research design, we can partially infer causation throughout the model. For example, since the reliability of our results were bolstered by the causal assumption between LCS and "Facebook engagement", we could infer a causal relationship between them. Likewise, observed variables such as "Total preferences" and "First preferences" also impact their direct latent variable, "Student preferences". However, due to a lack of scientific or observational causal assumption, other observed variables such as "Global ranking", "Go8 membership" and the latent variable, "Facebook engagement", may not impact "Student preferences" no matter how good a model fit turns out.

To control for outliers and extreme values in the data, we reran the analysis by using the mean values of the total number of LCS each year for each university rather than the raw total numbers. We found out that the new model explained the response variable Student Preferences (StPref) merely less than 1% better. The insignificance of the difference between two models – raw total vs mean – was due to the considerably large population size of the collected Facebook posts. We also noticed that grouping data by year and university, and then randomly sampling 2730 Facebook posts (35 posts per university for each year: 35*13*6) from the entire population did not significantly change the results from the original model. However, the original model outperformed the sampled data model by approximately 5% in terms of explaining the endogenous variable StPref. Furthermore, when we inspected the outlying content that attracted extreme high numbers of Likes, we noticed that it was neither due to the number of Comments and Shares on that post nor the type of created content. For example, the average number of Likes, Comments and Shares of all Macquarie University posts in 2017 were 7501, 76 and 43 respectively. It would make sense when a photo of a student flying a kite on the campus of Macquarie University receives 25 Comments and 11 Shares, however when the very same post was liked by more than 28000 people, there could only be two explanations: 1) Some of the 11 people and/or their network with whom the post was shared had considerable audience, or 2) The Likes were procured inorganically. What's more, same year for the same university a post about a Fish Laboratory – where researchers study learning, memory, cerebral lateralization and personality – received 352 Comments and 637 Shares, but merely around 4600 people liked the post. Whereas another news post about a research team – being awarded a total of $780,000 by the Australian Research Council – received only 5 Comments, 5 Shares and 230 Likes. Furthermore, we noticed similar discrepancies with other universities as well. For example, a virtual open house invitation post by the Southern Cross University (SCU) in 2021 received no Comments and just 1 Share, but more than 3000 Likes, whereas the average number of Likes for all other posts in the same year for SCU was merely 19.7 and the second highest number of Likes (after 3K) was only 89.

Linear Regression

	Corr	*p*	*Adj R-sqr*
Total Likes	0.462	<0.001	0.203
Total Comments	0.668	<0.001	0.439
Total Shares	0.738	<0.001	0.539
Mean Likes	0.437	<0.001	0.181
Mean Comments	0.499	<0.001	0.239
Mean Shares	0.497	<0.001	0.237

Table 3. Summary Statistics of the relationship between observed endogenous variables (Likes, Comments and Shares) and the endogenous outcome variable Student Preferences (StPref).

We found that the variation in the endogenous outcome variable StPref was explained by Total Likes (0.203) significantly less than it was by Total Comments (0.439) and Total Shares (0.539) (Table 3). Likewise, the adjusted *R-squared* of the average number of Likes (Mean Likes) was less than that of Mean Comments and Mean Shares. Therefore, we could infer that Shares and Comments were better indicators of StPref than Likes. We also found a strong positive correlation between Total Shares (0.738) and StPref. However, we do not observe a similarly strong correlation between Mean Shares (0.497) and StPref. Moreover, StPref was explained substantially better by Total Shares (0.539) than it was by Mean Shares (0.237). These findings indicate that where Total Shares and Total Comments had a strong positive relationship with StPref, the average number of Shares and Comments only had a moderately strong positive relationship with StPref. Therefore, it can be inferred that more Facebook posts that receive some number of shares and comments are a better indicator of a higher StPref than few Facebook posts that receive *on average* a larger number of shares and comments. In other words, student preferences can be predicted more accurately by the number of *total* shares and comments that posts attract on a university's official Facebook site.

	StPref		Fengage					Fit Measures			ANOVA		
Model	TotalPref	FirstPref	Tlikes	Tcomms	Tshares	GloRank	Go8	CFI	RMSEA	SRMR	AIC	BIC	Chisq
A	X	X	X			X	X	0.986	0.183	0.024	−154.7	−139.2	6.5
B	X	X		X		X	X	0.992	0.148	0.026	−202.6	−187	4.9
C	X	X			X	X	X	0.985	0.19	0.022	−171.3	−155.7	6.9
D	X	X	X	X		X	X	0.952	0.214	0.24	29.4	53.8	28.8
E	X	X	X		X	X	X	0.922	0.282	0.303	−1.2	23.1	44.8
F	X	X		X	X	X	X	0.917	0.293	0.27	−21.7	2.6	47.8
G	X	X	X	X	X	X	X	0.903	0.269	0.277	40.8	69.7	71.1
H		X	X		X	X	X	0.862	0.353	0.331	159	176.7	37.9
I		X		X	X	X	X	0.855	0.366	0.301	141.5	159.3	40.4
J		X	X	X		X	X	0.917	0.259	0.265	193.5	211.2	22.1
K	X		X		X	X	X	0.856	0.353	0.334	159.3	177	37.9
L	X			X	X	X	X	0.843	0.366	0.304	152.2	169.9	40.4
M	X		X	X		X	X	0.909	0.259	0.265	204.2	222	22.1
N	X	X	X			X		0.982	0.285	0.022	−146	−132.7	6.51
O	X	X		X		X		0.989	0.24	0.023	−188.4	−175.1	4.9
P	X	X			X	X		0.982	0.295	0.02	−160.1	−146.8	6.9
Q	X	X	X	X		X		0.987	0.183	0.019	−187.3	−171.8	6.5
R	X	X	X		X	X		0.984	0.191	0.017	−158.9	−143.4	6.9
S	X	X		X	X	X		0.985	0.198	0.018	−189.3	−173.8	7.3
T	X	X	X	X	X	X		0.987	0.148	0.016	−187.4	−169.6	7.4
U	X	X	X				X	0.991	0.211	0.012	−167.3	−153.1	4.4
V	X	X		X			X	0.994	0.171	0.016	−201.1	−187	3.2
W	X	X			X		X	0.991	0.21	0.012	−189.7	−175.5	4.4
X	X	X	X	X			X	0.992	0.147	0.012	−205.9	−189.4	5.3
Y	X	X	X		X		X	0.994	0.127	0.01	−189.1	−172.6	4.5
Z	X	X		X	X		X	0.991	0.152	0.012	−213.8	−197.3	5.6
Δ	X	X	X	X	X		X	0.993	0.111	0.01	−211.8	−192.9	5.9

Table 4. All Model inclusions, fit measures, and ANOVA results of chapter 2 study

Overall, as seen in Table 4, we noticed better goodness of fit statistics in terms of CFI, RMSEA, SRMR, BIC, AIC and Chisq when the model included StPref (latent endogenous variable) onto which TotalPref and FirstPref were loaded in comparison to models including either TotalPref or FirstPref separately. Excluding Go8 increased the goodness of fit particularly in terms of SRMR (≈88%) for the models overall. When GloRank was included in the models and both TotalPref and FirstPref were loading onto StPref, for all combinations of Tlikes, Tcomms and Tshares included in the models, the average SRMR was 0.166 (subpar) when Go8 was included, and it substantially dropped to 0.019 (good) when Go8 was excluded.

When we excluded GloRank but included Go8 in the models, we attained even better fit (≈92% for SRMR and ≈28% for RMSEA) compared to including both GloRank and Go8. When Go8 was included in the models and both TotalPref and FirstPref were loading onto StPref, for all combinations of Tlikes, Tcomms and Tshares included in the models, the average SRMR and RMSEA were 0.012 and 0.16, respectively. Finally, we observed the best model fit results (CFI=0.993, RMSEA=0.11, SRMR=0.01) in ModelΔ which included all endogenous and exogenous variables but GloRank. Whereas Model V - including Total & FirstPref, Tcomms and Go8 - provided the best fit results in terms of Chi squared difference (3.2) and CFI (0.994). Furthermore, we noticed a trend indicating that the models provided worse fit when only one of the observed exogenous variables (Total or First Preferences) was included in the model to capture the latent variable StPref. We also noticed that when both Go8 and GloRank were integrated in models, including only one of the three Facebook engagement indicators (LCS) particularly Tcomms or Tshares provided a better fit than the models including a combination of two or all of them in general.

2.6 Findings of chapter 2 study

One key finding of the study was that Facebook posts that receive a large number of Shares and Comments are a strong indicator of the choices made by prospective students. Furthermore, when total Likes, Comments and Shares of Facebook posts were loaded significantly onto a latent variable such as Facebook Engagement, Student Preferences could be better predicted by Facebook Engagement along with either the Go8 membership status or Global ranking of universities than other models that included both Go8 membership status and Global ranking of universities. Additionally, to our surprise, we achieved the best model fit when we included all variables but the Global Ranking of universities to predict student preferences for them.

Another key finding of this study was that we observed consistently better model fit results when a model included the number of Student preferences which was derived from the number of total and first preferences for universities in comparison to models including either total or first preferences for universities. This implies that when predicting for next year's enrolments, a university should be taking into account the total number of students whose not only first choice but also one of the 5 choices was to study there.

Earlier studies indicated that prospective university students resist social media marketing (Bal et al., 2015; Hou, 2018). However, our findings show otherwise. Firstly, we observed a strong relationship between student preferences in officially reported admissions and the total number of both Likes and Followers for these universities. Secondly, our findings demonstrated that some of the indicators of Facebook engagement such as Total Shares and Total Comments about the university posts have a strong positive relationship with student preferences for these universities. Thirdly, we could build a robust model for estimating and predicting the number of student preferences each year between 2016 and 2021 with all hypothesized Facebook engagement indicators – Total Likes, Total Shares, Total Comments – in tandem with either Global ranking of universities or the Group of Eight membership status of universities. This has a vital practical implication in advertising domain. It was found that social media users will more likely buy a product or service featured in an advertisement (ad) when they are more engaged with the brand as a result of likes, comments and shares (Lee & Hong, 2016). Moreover, social media users perceive ads as more relevant and useful when ad contents include more information related to their past interaction with the brand (Alalwan, 2018). This, in turn, enhances their purchase intent after being exposed to such ads. After all, prospective university students are not immune to the effects of eWOM or to the campaigns run by HEIs as part of their SMM strategies.

2.7 Implications and conclusions of chapter 2 study

As the competition for recruiting new students amplifies, universities are urged to create and maintain a brand image and reputation linked to a strong value offer (Adcroft et al., 2010) and respected by their target audience (Ivy, 2001). It was argued that Facebook can be utilized as a marketing tool to influence the choice of potential students by promoting the value propositions and messages of a university and spreading them virally throughout a large number of targeted audiences (Khan, 2013). Although this statement may or may not be true, without asking students who have already decided to study at a HEI about whether or not their decisions were affected by what they saw on Facebook, it should not be claimed that Facebook impacts prospective students' choice-making. However, based on our findings we demonstrated that key indicators of Facebook engagement such as the number of likes, comments, and shares on posts can be utilized to predict the number of students who prefer to study in HEIs. This is in line with previous studies that focused on other social media platforms such as Twitter demonstrating a strong positive relationship between the number of tweets and student preferences for universities particularly for the ones that have a high academic ranking (Cingillioglu et al., 2021).

We all play our roles as a social entity in the society, and we all have multiple role identities. In mid-20th century, people exhibited their identity salience and hierarchy by what they wore at an event, whereas a century later we do so by what we share on social media. Sometimes this could be the latest news about a scientific breakthrough, or a sunset photo taken from the university campus. However, it is difficult to say which one will be shared more or lead to improved brand image, reputation, and trust, and resulting in higher student retention and recruitment. When more students follow a university and share its content, this may be an indicator that more students will be choosing to study there next year. Our results indicate a moderately strong positive relationship between Facebook Engagement and student preferences for universities. Since correlation is not causation, we cannot claim that Facebook engagement influences students' decisions. Based on our findings, however, we could claim that 1) Likes, Comments and Shares impact Facebook Engagement, and 2) Total Preferences and First Preferences impact Student Preferences.

The key theoretical contribution of this study is that it can potentially help universities build upon their analytical competencies through indicators related to their Facebook interaction with the public so that they will have a better understanding of not only how the decision-making process of new students evolve over time but also predict upcoming years' student enrolments. From a marketing and financial perspective, this may give universities the competitive edge to increase their income by attracting and recruiting more students and devise a more pertinent strategic plan in terms of budgeting and resource allocation in advance. Last but not the least, it should be noted that the methodological and conceptual takeaways and lessons learned from this study may not be exclusive to only the higher education sector. Indeed, in addition to the extant literature, this study may serve as a basis for future studies with an aim to use social media analytics to understand customer trends & decision-making and predict organizational performance in a wide range of fields.

2.8 Limitations of chapter 2 study and recommendations for future research

One limitation of this study was due to the recent amendments in Facebook's policy regarding site scraping that the data could be collected merely in a quite unstructured form. Data wrangling process was laborious and time consuming because we had to filter, group, and then analyse all Facebook posts pertaining to each year between 2016 and 2021. We hope that Facebook Inc. will provide in the future ease of access to its full historical data via Facebook Social Graph for researchers using authentication tokens as other social media giants do like Twitter.

Another limitation was the inaccessibility of extensive data pertaining to student preferences for universities. We could only use data involving 13 universities in NSW and ACT because UAC is the only provider of raw open data containing such student preference statistics for universities, and yet it covers merely 13 of them. For future studies, we recommend researchers to try to gain access to national data repositories incorporating student enrolment statistics for all universities in their country.

Chapter 3

3. Topic identification via topic modelling from social media

Topic Modelling is a suite of techniques to identify latent topics (themes) in a group of documents (e.g., articles, news feeds, reports). Some researchers like to think of topic modelling as a tool for "amplified reading", considering it a technique for seeing themes out of large groups of texts that humans could never possibly read themselves. Despite the intentions of the original developers of topic modelling, this technique has transformed to be commonly used for identifying features and variables in textual data.

Probabilistic topic models have been developed by machine learning researchers to discover and annotate large and otherwise unstructured collection of documents comprising thematic information (Blei, 2012). Topic modelling algorithms are statistical methods that can be applied to large collection of documents to analyze the words in original texts (without the need for document labelling or prior annotations) to uncover the topics that permeate through them, how these topics are linked to one another, how they evolve over time, and we can summarize and collate electronic collections of documents via topic modelling at a scale and efficiency level that would be insurmountable to achieve via human annotation (Blei, 2012).

A topic model does not tell us how many topics there are in our corpus or their names. The output of a topic model is a list of words associated with each topic with high probability which reflects the grouping capability within the corpus since documents with an analogous topic probability distribution can be clustered together (Liu et al., 2016). In generative approaches, we have a story about how the data came to be and this story is told in terms of probability using tiny building blocks comprised of many different distributions. In this story, there are some missing pieces (i.e., latent variables) that we aim to uncover using a process called probabilistic inference. As a multivariate version of Beta distribution, Dirichlet

Distribution "provides a convenient conjugate prior for Bayesian analyses involving multinomial proportions" (Lange, 1995). Therefore, using a multivariate continuous probability distribution such as Dirichlet Distribution, each topic from a corpus can be generated from a multinomial distribution over terms (i.e., words) that are to some extent related to one another. The documents also have a Dirichlet distribution over topics. For every document, the Dirichlet distribution over all possible topics selects to which topics the documents are allocated.

3.1 Latent Dirichlet Allocation

3.1.1 Conceptual background

In Bayesian generative probabilistic modelling, data are treated as occurring from a generative process comprising latent variables. This process delineates a joint probability distribution over latent and observed random variables. Analysts and researchers conduct data analysis by utilising this joint probability distribution to calculate the conditional distribution of the latent variables given observed variables. This conditional distribution is referred to as the posterior distribution (Blei, 2012).

As a generative probabilistic statistical model used broadly in natural language processing, Latent Dirichlet Allocation (LDA) is a topic modelling technique for extracting topics (themes) from a given corpora. LDA has become a building block that facilitates a plethora of applications (in addition to text, it can be applied to issues encompassing data collections such as collaboratively filtered data domains, bioinformatics, and content-based image retrieval (Blei et al., 2003). Since organising and findings patterns in text is a vital task in a wide range of fields, industry and culture (Blei, 2012), researchers have made algorithmic improvements to fit models to big data. In LDA, where the words in the documents are the observed variables; the topic structure is established by the latent variables. To infer the latent topic structure from documents, the conditional distribution of the latent variables given the documents (posterior distribution) must be computed (Blei, 2012).

In their original paper, Blei et al. (2003) used the language of text collections as they referred to entities such as "words", "documents" and "corpus". These key terms were defined as:

1) Word: *S*imple unit of discrete data, defined within a set of vocabulary $\{1, \ldots, V\}$. Words were represented with a unit-based vector having a single component of 1, whereas all other components were 0. Accordingly, the vth word in the vocabulary was denoted by a vector of V: w such that $w^v = 1$ and for all other components: $w^u = 0$ where $v \neq u$.
2) Document: Sequence of *N* words in a given text where w_n is the n^{th}word in the sequence denoted by $\boldsymbol{w} = (w_1, w_2, w_3, \ldots, w_N)$.
3) Corpus: Collection of *M* documents denoted by $D = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3, \ldots, \boldsymbol{w}_M\}$.

Blei et al. (2003) aimed at finding a probabilistic model of a corpus that assigns high probability to the components of the corpus, as well as to other "similar" documents. Accordingly, they developed LDA upon the basic idea of representing documents as random combinations over latent topics where every topic is categorized by a multinomial distribution over words. For each document **w** in a corpus D, LDA assumes (Blei et al., 2003):

1) Choose $N \sim$ Poisson(ξ),
2) Choose $\theta \sim$ Dir(α), and
3) For each of the *N* words w_n:
    a. Choose a topic $z_n \sim$ Multinomial (θ).
    b. Choose a word w_n from $p(w_n \mid z_n, \beta)$ a multinomial probability conditioned on the topic z_n.

In the basic model, Blei et al. (2003) made several underlying assumptions. The first assumption is that the dimensionality of the Dirichlet distribution k (number of topics for variable z) is predetermined and fixed. The Dirichlet random variable θ with k dimensions take values in the $(k-1)$-simplex. The second assumption is that the probabilities of words are treated as a fixed quantity parameterized by a $k \times V$ matrix β where $\beta_{ij} = p(w^j = 1 \mid z^i = 1)$.

Furthermore, distributions for more feasible document lengths can be used as required since ξ assumption is uncritical to anything that follows. Finally, of variables θ and z that generate other data, *N* is independent hence considered an ancillary variable and its randomness in the further development of the model is mostly ignored.

Blei defined LDA later formally by using the following notation (2012): Where each β_k is a distribution over the vocabulary, the topics are represented by $\beta_{1:K}$. Where $\theta_{d,k}$ is the topic proportion for topic k in document *d*, the topic proportions for the dth document are represented by θ_d. The topic allocations for the dth document are represented by z_d, where $z_{d,n}$ is the topic allocation for the nth word in document *d*. The observed words for document *d* are w_d, where $w_{d,n}$ is the nth word in document *d*. Using this notation in the generative process of LDA, Blei (2012) formulated the joint distribution of latent and observed variables as follows:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n})$$

There are key dependencies that define the LDA. These dependencies as specified in the distribution formula above are encoded behind the generative process in the statistical assumptions (Blei, 2012). For instance, the topic allocation $z_{d,n}$ depends on the per-document topic proportions θ_d. Another example is that the observed word $w_{d,n}$ depends on the topic allocation $z_{d,n}$ and *all* of the topics $\beta_{1:K}$. Using the same notation, the posterior is calculated by dividing the joint distribution of all random variables by the marginal probability of observations (probability of observing the corpus under any topic model) as shown below:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D} \mid w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}$$

We can also represent the LDA model with a graphical-plate illustration as shown below (Figure 6), where w_{ij} denotes the specific observed word*, M* is the number of documents, *N* represents the number of words in a document, α is the proportions parameter, β is the topic parameter, θ_i denotes the topic proportion for document i, φ_k represents V-dimensional vectors containing distribution parameters of the Dirichlet-allocated topic-words (V represents the number of words in the vocabulary and K is the number of topics), and z_{ij} is the topic for the jth word in document i. Where nodes represent random variables, plates and edges indicate replication and dependence, respectively. The only observed variables (shaded nodes) are the specific words (i.e., w_{ij}) whereas the rest of the variables are latent variables (unshaded nodes).

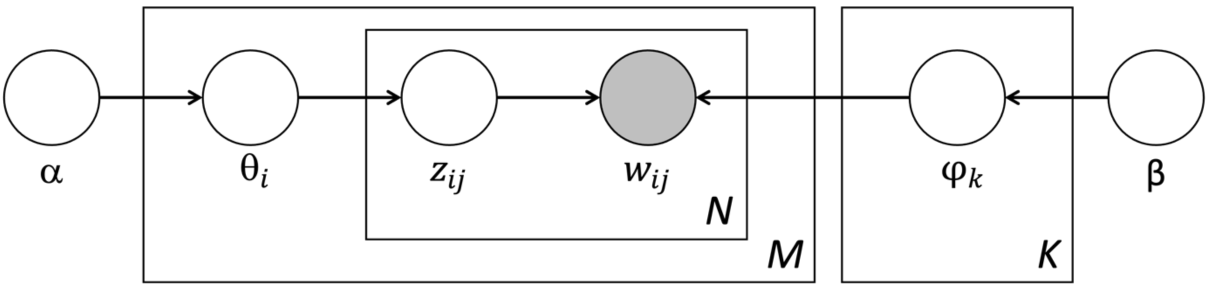

Figure 6: LDA as a graphical model. Adapted from Blei (2012).

As a probabilistic graphical model illustrated in Figure 6, the LDA model comprises three levels of variables (Blei et al. 2003). First, α and β are corpus-level parameters and in corpus generation process they are assumed to be sampled only once. Second, sampled once per document, θ_i are document level variables. Third and last, sampled once for every word in every document, z_{ij} and w_{ij} are word level variables. Through these three levels of variables, LDA allows for topic node to be sampled repeatedly within documents which can be linked to multiple topics. The graphical model is useful because, as stated by Blei (2012), it defines a factorization of the joint probability distribution of the hidden and observed variables (the joint defines a posterior, $p(\theta, z, \beta \mid w)$), encodes independence assumptions about these variables and joins the algorithms for computing inferences from data such as: Per-word topic assignment z_{ij}, per-document topic proportions θ, and per-corpus topic distributions β. Finally, the model allows us to take posterior expectations of inference to perform the task at hand from topic exploration to document similarity, to information retrieval, to building a navigator around collection, and others.

In principle LDA trades off two goals: (1) In every document, words are allocated to only few topics, (2) In each topic, high probability is assigned to only few terms. These goals are in conflict because putting very few terms in each topic makes #1 difficult to attain (to cover a document's terms, many topics must be assigned to the document) and putting a document in a single topic makes #2 difficult to attain (all the document terms must have probability under that topic). As a result of this trade-off, tightly co-occurring terms emerge (Blei, 2012).

Developing upon LDA, supervised LDA (sLDA) was introduced by Blei and McAuliffe (2007) as a statistical model of labelled documents to improve the prediction performance of document categories allowing for better topic detection and demarcation. The advantage of sLDA is that a variety of response types could be accommodated by the model as the unsupervised topic detection process strives not only to optimize the consistency of the topics underlying the data, but to maximize the model's capability to predict related values (Resnik et al., 2015). Ashktorab et al. (2014), for example, utilized sLDA to extract information about crisis events in North America from Twitter, annotated each tweet and made this information available to first responders. After procuring sets of seed words from existing dictionary sets in literature and Amazon Mechanical Turk, Toubia et al. (2019) extracted the features of movies from their descriptions via sLDA to provide a recommendation classifier based on psychological topics. Resnik et al. (2015) provided promising results utilizing variations of supervised LDA models such as Blei and McAuliffe's sLDA (2007), Nguyen et al.'s supervised anchor topic models (2015) and supervised nested LDA (SNLDA) model. It was found that these supervised models uncover latent structure better than LDA alone and as weekly aggregation of tweets improved precision, other types of aggregation may also produce better outcome (Resnik et al. 2015).

Like all other topic modelling techniques, LDA is not without limitations. First, since correlations among topics are not captured via Dirichlet distribution and sentence structure is not considered in LDA, due to its complete unsupervised nature, LDA does not allow for drawing inferences from more advanced NLP techniques such as sentiment or semantic analyses (Koltcov et al., 2014). However, there are some LDA extensions that slacken the model's relative assumptions. For example, the pachinko allocation machine (Li & McCallum, 2006) and the correlated topic model (Blei and Lafferty, 2007) permit topics to establish correlation among one another (e.g., a document about *pharmacology* is more likely to be correlated with *biology* than it is to be about *management*). Furthermore, the spherical topic model developed by (Reisinger et al, 2010) identifies the unlikelihood of words to be included in a topic (e.g., "screwdriver" will be quite unlikely in a topic about *cattle*); by considering the "burstiness" in topic models (Doyle & Elkan, 2009) a more feasible model of word frequencies can be attained; and sparse topic models (Wang & Blei, 2009) apply additional structure to the distribution of topics (Blei, 2012).

Second, the assumption of document exchangeability is heavily criticised in literature because the evolution of topics over time is not considered a changing but a static factor in LDA (Du et al., 2012; Cinbis et al., 2015; Jelodar et al., 2019). Dynamic topic modelling was introduced by Blei et al. (2006) to address the issue of not considering topic evolution in LDA over time. A dynamic topic model in principle considers the sequence of the documents and provides a posterior topical structure that is richer than LDA. Since a topic becomes a part of the order of distributions over words, rather than a single distribution over words, Blei et al. could discover an underlying topic of the collection and monitor how it has transformed over time (2006). However, Blei (2012) stated that LDA is not suitable for sophisticated NLP goals such as language generation mainly because it does not consider the order of the words (bag of words: BOW) in a document. As improvement extensions to LDA, a topic model developed by Griffiths et al. (2004) shifts from LDA to a standard Hidden Markov Model and another model developed by Wallach (2006) relaxes the BOW assumption by considering the conditionality of consecutive words in topic generation. Although substantially expanding the parameter space, these models provide a better language modelling performance than the standard LDA (Blei, 2012).

Finally, another commonly criticised aspect of LDA is that since the number of topics K must be predetermined at the onset and be fixed, one study explored how K impacted results and found that the performance of LDA varies greatly in accordance with the encoded K (Tian et al., 2009). A solution was proposed by introducing Bayesian nonparametric topic models (I et al. 2006) which find a tree structure – inferred from the data – of topics and shift from more general to more specific. Through Bayesian nonparametric topic modelling where K is determined during posterior inference by the collection, new documents can help reveal uncovered topics. (Blei, 2012).

For implementing basic models of LDA, Blei et al. (2003) adopted empirical parametric Bayes approaches for estimating model parameters such as α and β. For more complex models, as a building block enabling many applications, with a capability to unravel K number of hidden topics in documents through posterior inference, LDA can be used as a powerful NLP technique in tandem with probabilistic Latent Semantic Indexing (LSI) and Principal Component Analysis (PCA), matrix factorization for information retrieval (Perkio et al., 2004) and Collaborative Topic Models (CTMs) that connect content to consumption (Blei, 2012). For instance, when scientists share their research libraries, CTMs can help readers discover new and old documents, categorize readers in terms of their topical preferences so that the documents in which they are more likely to be interested can be recommended to them, and after ingesting a big collection of articles, impactful and interdisciplinary documents can be identified and recommended to interested readers.

Since possible topic structures are exponentially large and the sum is difficult to compute (because it is over all probable ways of allocating every observed word to each topic), probabilistic topic modelling algorithms have been developed to approximate formulas like the one shown above. These algorithms adapt an alternative distribution over the hidden topic structure closing in on the true posterior (Blei, 2012). One way to achieve this is via sampling-based algorithms and *Gibbs sampling* is the most commonly used sampling-based topic modelling algorithm (Blei, 2012). In Gibbs sampling, a Monte Carlo Markov Chain is built upon the latent topic variables for a specific corpus. The modelling algorithm runs the Markov chain repeatedly, collecting samples from the restrictive distribution, and estimates the distribution with these samples (Blei, 2012). In simpler terms, Gibbs sampling along LDA returns a bag of words which is taken from each document in the collection, and in each document the words may characterise the relevant potential topic (Montenegro et al., 2018).
Besides common inferential techniques such as Gibbs sampling and variational Bayesian inference for fitting LDA models, more advanced inferential LDA methods have been proposed addressing not only unimodal latent topic distributions but also multimodal ones. These models can help researchers and data miners acquire “unbiased estimates under flexible modelling for heterogeneous text corpora via partial collapse method and Dirichlet process mixtures” (Park

et al., 2019). As an advantage over probabilistic LDA models, these models can automatically derive optimal hyperparameters from the data and through partial collapse the sampler provides feasible inferences and makes impartial parameter approximation for extremely multimodal hidden topic distributions with rapid convergence (Park et al., 2019).

Since the discrete distribution of topics over words in LDA creates words in documents, as a mixed membership model of grouped data, LDA allows for each group of data to be linked with multiple components (topic) in different proportions rather than a single component (Blei, 2012). This becomes an advantage of LDA as the options for data distribution and topic parameter can be adapted to other areas of observation by making minor modifications to the relevant inference algorithms. For example, LDA-based models have been adapted to numerous sorts of data, including social media data, audio and visual data, codes, logs, user preferences and survey data (Blei, 2012).

3.1.2 LDA models on Social Media Data

LDA is commonly used in social media analysis due to its rigour, simplicity, and ease of interpretability (Chen & Ren, 2017). There are in extant literature a plethora of studies applying LDA-based probabilistic generative models to social media data. Most of them so far have focussed on some the most popular social media platforms such as Twitter and Facebook.

3.1.2.1 LDA models on Twitter

LDA is a powerful exploratory topic modelling technique when deployed on a large Twitter corpus as it can help researchers uncover latent themes in documents and makes it easier to analyze, categorize and summarize big text data (Yang & Zhang, 2018). For example, using LDA, Xue et al. (2020) categorized almost 2 million tweets about coronavirus into ten topics and calculated the intertopic distance to identify related themes. Brzustewicz and Singh (2021) posited that Twitter is a rich data source for researchers to explore the behaviors, unbiased opinions, and true feelings of public. Therefore, they used LDA to identify topics from Twitter related to sustainable consumption vital during Covid-19 pandemic. To understand more about public awareness through Twitter for social challenges that minorities and other social groups of people face, Tong et al. (2022) utilized LDA to detect top high-level words and categorized themes related to emerging online social movements such as Black Lives Matter and Stop Asian Hate.

Latent Semantic Analysis (LSA) is a commonly used NLP technique in tandem with LDA. In addition to LSA, Dikiyanti et al. (2021) used LDA to identify and classify latent themes in tweets containing a single keyword. However, since LSA does not consider intertopic relationships like LDA, it was found that LDA had provided better outcomes than LSA alone (Qomariyah et al., 2019). Besides LSA, studies used other supervised methods to complement LDA for analysing sentiment in documents. Montenegro et al. (2018), for instance, used LDA to identify topics and topic clusters on Tweets about Dumaguete City and conducted a sentiment analysis by employing machine learning algorithms such as Support Vector Machine (SVM) to detect the sentiment of each topic cluster. Similarly, Jamal et al. (2020) used LDA to extract topics from a large number of Tweets and *k*-Nearest Neighbor (KNN) algorithm to detect and predict sentiment from the extracted topics. In another study, researchers used an *n*-stage LDA-based machine learning classifier to determine one of the five sentiments (happy, sad, angry, scared and surprised) for each tweet and achieved the highest accuracy of 76.4% with a 3-stage classifier (Güven et al., 2018).

Yang and Rim (2014) used a temporal trend sensitive LDA-based model to identify latent topics in tweets upon which they scored the attractiveness of each tweet. They weighed topics by considering their illustrative words and analysing the probabilities of their spatial and temporal variation. Ostrowski (2015) assessed the performance of LDA for topic modelling to gain a deeper understanding of social media trends particularly on Twitter. The researchers applied LDA as an unsupervised model to a classification task on a filtered collection of Tweet corpus. Although the performance of LDA did not exceed that of pure Bayesian-based supervised models, the results indicated that LDA could be used as a complementary method to back up large-scale corpora classification, derive information about trends and help researchers identify new trends and noteworthy themes from social media data (Ostrowski, 2015). Utilizing probabilistic modelling based on LDA, Kim and Shim (2014) developed a recommendation system on Twitter that identifies and recommends a maximum number of other users to follow and a maximum number of tweets for a user to read. They managed to enhance the performance of recommendations by using friend networks among users as well as tweet content. Building upon LDA modelling, their model exemplifies a genuine process of tweeting and establishing friend connections by using matrix factorization. To approximate the maximum likelihood function and learn about the model parameters, they utilized a variational Expectation-Maximization (EM) algorithm and then calibrated the ranking algorithms in accordance with the projected model parameters to recommend a user the maximum number of tweets to read and the maximum number of other users to follow (Kim & Shim, 2014).

To make better informed weather forecasts, understand disaster trends, and classify other climate related information posted by the official Twitter account of a meteorology institution, Hidayatullah et al. (2019) applied LDA to tweets and ranked the extracted topics based on their relevance. Similarly, Zhou et al. (2021) implemented an LDA model on the pre-processed Tweets that have been posted during the disastrous event of Hurricane Laura. Automatically created topics were monitored by the researchers and category labels were manually assigned to them. Topics were easier to identify and interpret through a supervised approach determining a classification outline into which Tweets were classified improving the likelihood that the categories better represent the possible topics that reflect societal concerns and matters instigated by the hurricane (Zhou et al., 2021). Since in noisy Twitter stream, manually detecting disaster topics can be time consuming, to allow for a quicker disaster response Ferner et al. (2020) developed a fully automated LDA-based topic identification system that detects a relevant disaster topic in accordance with a set of seed words initializing the topic model. The researchers generated the seed words automatically from earlier tweets posted in the same geospatial area and confirmed that the geographical distribution of tweets related to 2014 Napa Valley earthquake and 2017 Hurricane Harvey match the official release of these disaster events' footprints (Ferner et al., 2020).

Any topic detected and labelled as "bursty" on Twitter is a topic that initiates a stream of related tweets in a short time, which usually echoes main occurrences of mass interest (Jelodar et al., 2019). Hence, leveraging large-scale Twitter data to identify bursty topics has been a research issue with valuable practical implications. Xie et al. (2016), for instance, introduced a technique called TopicSketch that detects bursty topics on Twitter real-time. Through this technique, they also demonstrated that huge numbers of tweets reaching hundreds of millions can be processed on a single machine real-time and bursty topics can be identified concurrently in fine granularity.

Recent modelling applications of LDA on Twitter data incorporated a wide range of areas for various research purposes. Singh and Glińska-Neweś (2022) modelled public attitude from Twitter toward organic foods. Pardo et al., (2022) used LDA to extract topics and analyse gaps between the vocabulary and themes identified in the corpus of tweets and retweets belonging to B2B companies. Lossio-Ventura et al. (2021) compared the performance of LDA-based topic models on health-related tweets. Furthermore, variations of LDA were used as either a semi- or unsupervised approach for modelling topics from Twitter on racial discrimination (Balakrishnan et al. 2022), Covid-19 (Gupta & Katarya, 2021; Gourisaria et al., 2022), urban governance (Alswedani et al., 2022), surgery patients' personal goals (Li et al., 2019), genderless consumer fashion trends (Kim et al., 2022), and many more.

3.1.2.2 LDA models on Facebook

LDA has widely been used for identifying topics from Facebook posts. Qian et al. (2014) used a supervised version of LDA to extract social events as topics from large social media data posted on popular platforms such as Facebook and classified these events so that they can be searched, explored, and monitored by users and governments. Other researchers such as Abinaya et al. (2014) mostly used unsupervised versions of LDA for event identification. Furthermore, it was used to drill down into more specific events such as social movements. For example, to identify the kind of topics discussed through user comments on Facebook, Smith and Graham (2019) used LDA and analyzed discourse on Facebook pages to explore and map antivaccination movements mobilising public for antivaccination practices. Furthermore, to help social media content providers of healthcare understand autism from the standpoints of users seeking emotional support, motivation, and advice (Newman et al., 2011), improve patient-family-carer communication and thus deliver autism-affected families better service, Zhao et al. (2019) utilized LDA to detect topics from relevant user-generated content posted on Facebook's support groups for autism. In a prospective cohort study examining a sample of Facebook posts, Smith et al. (2017) used LDA to delineate topic variations in the patterns of postings across validated patient health conditions. In another study, Tai et al. (2015) used LDA to detect mental disorder and predict depressiveness by the content users posted on social media.

LDA was also used to establish correlations among identified topics. For example, to detect and monitor the progress of the topics conversed in free-text format on a cancer institution's Facebook page, Tang et al. (2017) used LDA and identified ten topics from some of which they revealed interesting trends such as a negative relationship between greetings and blessings but a strong positive association between greetings and other family members/friends, as well as a positive association between the cancer institution and blessings. Similarly, Marengo et al. (2019) used LDA to establish an association between text features and alcohol use in young adults.

LDA was also used to explore various forms of interaction among Facebook users and their demographics. Cakmak and Eroglu (2020), for instance, used LDA to conduct content analysis on the posts generated on the Facebook pages of public libraries and found that user interaction is usually in the form of liking and content were created mostly to attract users from pre-school ages to young adults. Besides topic modelling, LDA was used for dimensionality reduction of large corpora. However, using Facebook data, Schetgen et al. (2021) found that singular value decomposition (SVD) had a better performance than LDA when first-time donation was predicted for the non-profit sector. Moreover, through LDA Zuorba et al. (2017) identified important topics related to excessive sadness expressed by students in the form of text on Twitter and Facebook. Besides these popular platforms, similar psychological studies incorporated data obtained from other social media platforms. For example, Hwang et al. (2020) used LDA to analyse behavioral data collected from Reddit on emotional eating. Besides from Facebook, LDA was used to analyze data from other social media platforms in other contexts. Qiang et al. (2017), for instance used LDA on Weibo data to discover topics indicating different geographical locations.

Other researchers extracted text from Facebook and used LDA to identify and classify latent topics in a variety of different contexts on various domains such as student learning (Zarra et al., 2016), identity theft (Funcion, 2017), strategies of U.S. presidential candidates (Ryoo and Bendle, 2017), eating disorders (Moessner et al., 2018), and more recently on electric vehicles (Debnath et al., 2021), opioid epidemic (Stokes et al., 2021), scholarly articles on measles (Wawrzuta et al., 2021), concerns of Thalassemia patients, carriers, and caregivers (Phang et al., 2021), substance use disorder (Liu et al., 2022) and COVID-19 halal vaccination (Feizollah et al., 2022), Vietnamese traditional medicine conversations (Nguyen, 2021), follower engagement (Chun et al., 2021), cultural government projects (Silva et al., 2021), sponsored content (Martins et al., 2022) and prediction of users' mental state (Kotenko et al., 2021), personality (Sagadevan et al., 2022), emotions (Khan et al., 2021), donation pattern and behaviour (Schetgen et al., 2021), and excessive alcohol use (Jose et al., 2022).

3.1.3 Topic models in Higher Education

As already discussed, researchers have implemented various topic models in a wide range of fields and domains. However, from a marketing perspective of higher education institutions, the extent of this research was quite limited. Wijenayake et al. (2017), for example, developed a neural network model to identify brand personalities for higher education institutions from social media data. However, let alone some of the most popular topic models such as LDA and STM (Structural Topic Modelling), no topic model has hitherto been used to extract decision factors for students' higher education choice from social media.

3.2 Data analysis via LDA

Tweets mentioning 13 universities located in NSW and ACT between 2017 and 2021 were collected through Twitter Developer API v.2 using R software version 3.5.1. Facebook posts created by each of these universities in the same period were extracted from Facebook Social Graph via R software version 3.5.3 and Octoparse. The lists of tweets and Facebook posts were collated and all duplicates (i.e., retweets, reposts) were removed from the dataset. Hence, each unique Facebook post or tweet was considered a document. After removing duplicates there were 8091 unique documents left to be analysed further.

The data were pre-processed (i.e., transformed to lower letters, numbers & punctuation & stopwords & whitespace were removed) and a document term matrix was developed using *tm* package on R. To calculate the optimal number of topics for the LDA model, the LDA model parameters were tuned, and model results were scored using *ldatuning* package.

One of the most important inputs in parametric topic models to be determined by topic modellers is arguably the number of topics. While there is no accepted 'best solution' to finding the most converging or fitting number of topics in a model, more detailed and fine-grained representations of the data can be provided by many topics, yet at the cost of being less precisely estimated than by fewer topics. To determine the optimal number of topics in our model, we tuned the LDA model parameters and scored the model results by using *ldatuning* package (Nikita & Chaney, 2022) on R. It should be noted that even though we have a large corpus comprising many documents, we need to take the length of each document into account since particularly tweets are made of quite short text. Hence, we decided to use Gibbs sampling with four metrics namely "Griffiths2004", "CaoJuan2009", "Arun2010", and "Deveaud2014"[3] to determine the number of topics (k) to be between 2 and 15 inclusive ($2 \leq k \leq 15$). Upon implementing its scoring algorithm, each metric produces a scalar LDA model score. The most preferable number of topics for the LDA model is estimated by running the scoring algorithms of all four metrics in a data frame. According to these four metrics, the most desirable number of topics can be established through the LDA model parameters that minimize "Arun2010" and "CaoJuan2009", but at the same time maximize "Deveaud2014" and "Griffiths2004."

[3] Detailed information about these metrics can be found in their corresponding articles: Griffiths et al. (2004), Juan et al. (2009); Arun et al. (2010); and Deveaud et al. (2014).

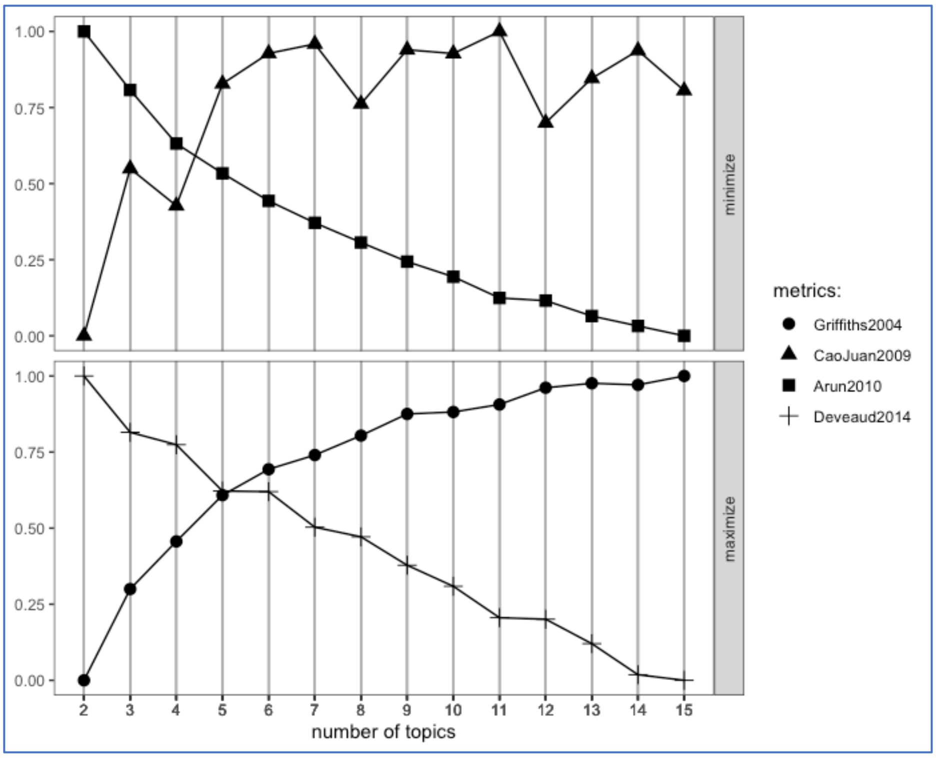

Figure 7: Topic number determination metrics' results for the LDA model

To analyse optimal topic number, a visual support function is provided within the ldatuning package. Upon plotting this function, as shown in Figure 7, we notice a fanning out after 5 or 6 number of topics for both pairs of metrics. As a result, we opted to include 6 topics in our LDA model.

After we developed an LDA model using Gibbs sampling (seed:5555) with 6 topics, we modelled every unique document (N=8901) as a mixture of topics and estimated per-document-per-topic probabilities (gamma) in tidy function of R with the argument: matrix= "gamma". In other words, we estimated topic probability per document using "gamma" matrix and created a data frame with gamma results (Figure 8).

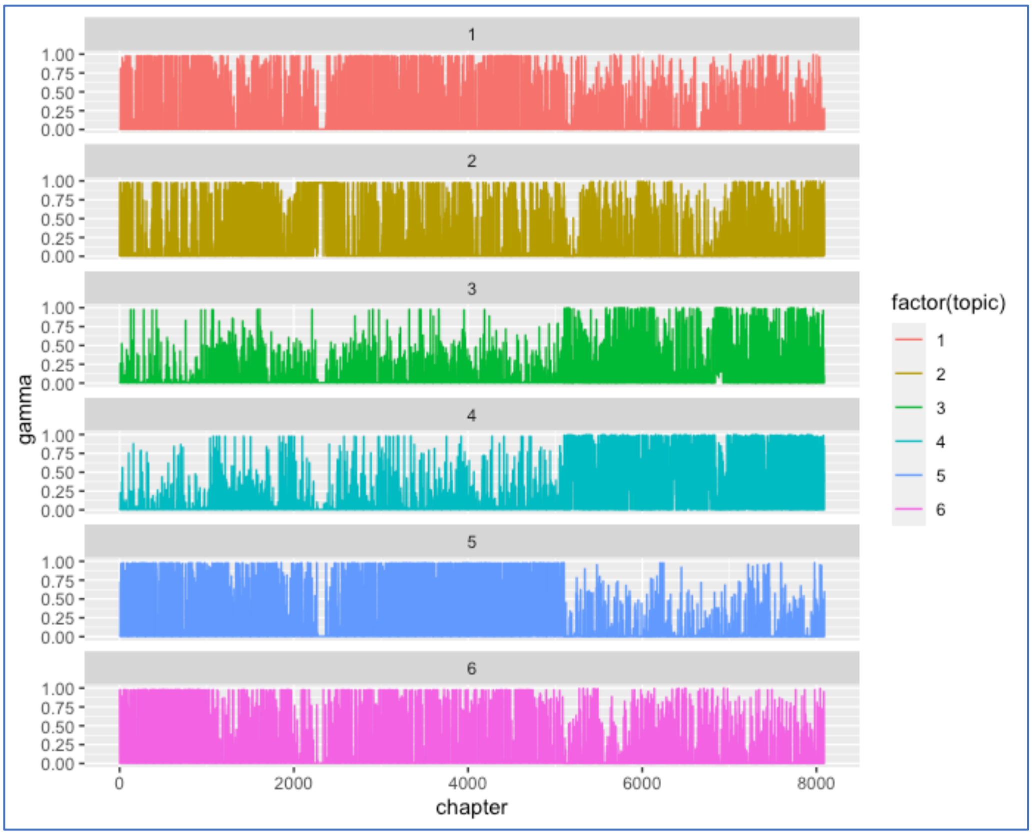

Figure 8: Topic probability per document estimations using "gamma" matrix via LDA modelling

Each gamma value is an approximated proportion of words from the corresponding document generated from that topic. Although there is no discernible trend, in terms of per-document-per-topic probabilities we notice a shifting point after mid 4000-6000 range as there is a conspicuous similarity pattern between topic 3 and 4, whereas topic 5 has a reverse distribution compared to topics 3 and 4.

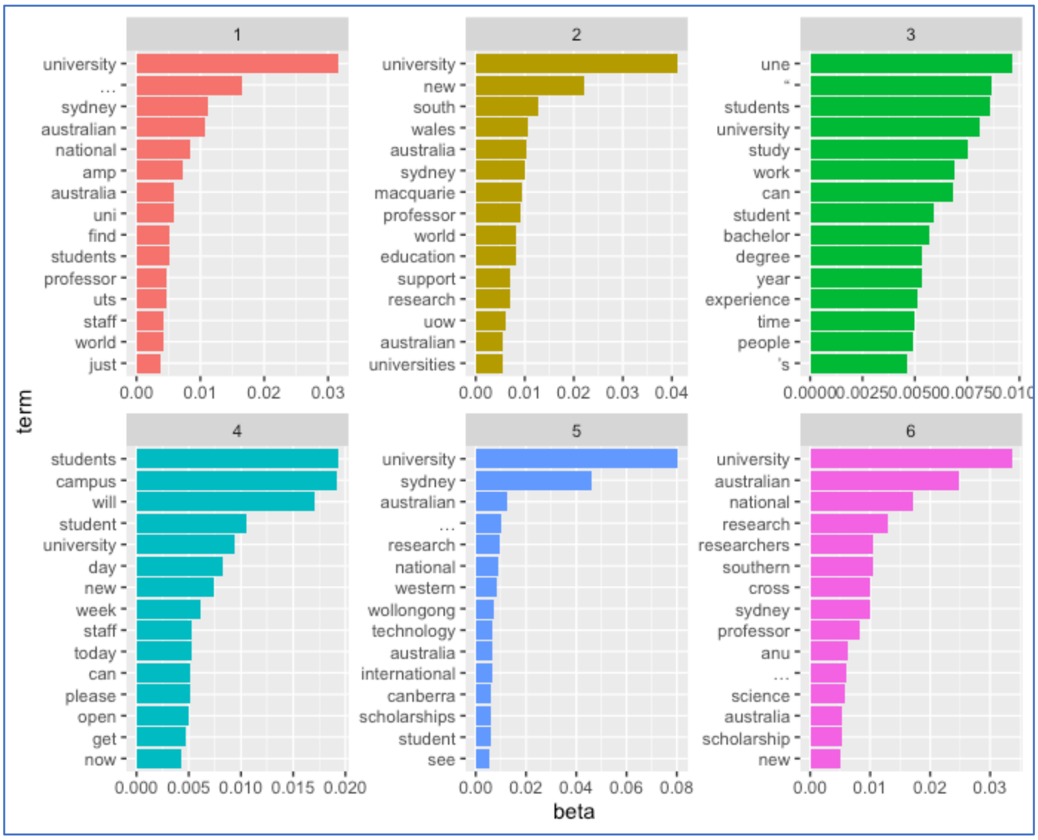

Figure 9: Word probability per topic estimations using “beta” matrix via LDA modelling

To identify and label distinct topics, we constructed a beta matrix where word probabilities per each one of the six topics are estimated (Figure 9). In line with the gamma matrix output, the most useful differentiating terms appear for topics 3 and 4 in the beta matrix. Conversely, the other four topics mostly incorporate quite similar hence undifferentiating terms such as “professor”, “94odelling”, “research”, “university” as well as the names, locations, and abbreviations of some of the universities located in NSW and ACT. We notice many identical or similar terms pertaining to various topics generated by the LDA model making it difficult to identify topics. This is not due to the unsupervised or non-automated nature of the modelling, but due to the short text usually mentioning quite similar themes in Facebook posts and particularly in tweets. Although a fully automated LDA-based topic identification system such as the one developed by Ferner et al. (2020) or a supervised model with a classification outline

(Zhou et al., 2021) might work for cross validating topics extracted from Twitter, due to the short text in each document, the best results for differentiating topics from one another can only be attained manually (until AI becomes capable) upon the output generated by the beta matrix. In a meta study exploring how topic modelling was utilized in software engineering research, Silva et al. (2021) found that out of 111 papers, most articles (n=75: 67.6%) did not even mention how topics were labelled. More importantly, it was revealed that most papers which explained how they assigned labels to topics (27 out of 36) used a manual approach as they deducted names based on the frequency of words in a topic and relied on human (researchers') interpretation of words in clusters.

In our case, once the LDA model returns the top 10 words (signature words) that have the highest probability of being generated by a corresponding topic, the key themes in documents need to be identified by clustering these words in a meaningful way. However, this is not a straightforward process. Since there are many overlapping words to be clustered with other words across 6 topics, instead of computational techniques (e.g., k-means clustering, hierarchical clustering), we preferred to use a manual approach to explore how the signature words relate to one another so that a cohesive theme they represent can be identified. In many of David Blei's papers, he and his colleagues also used a similar manual approach to assign names to the topics discovered by the LDA algorithm. To interpret and communicate the results to a wider audience, they typically examined the most representative words in each topic providing a high-level summary of the main themes discovered in the data and finally came up with a label (i.e., topic name) that best describes the overarching themes.

It is important to note that identifying key themes out of these words in LDA requires human judgment and discretion. To strengthen researchers' interpretation of these words as well as to validate and refine the topics, a common approach is to review the literature based on the initial analysis and its output (i.e., signature words). Therefore, to identify students' university choice themes, based on the signature words, we conducted a 3-stage systematic literature review (SLR). In the first stage, we carried out a broad search for articles published in Business Source Ultimate (via Ebsco) and ERIC: Educational Resources Information Center (via Proquest) Databases using each topic's signature words as well as essential bigram combinations of domain specific keyword terms such as "university", "college", "choice/s", "decision/s", and "enrolment/s" (e.g., "college choices"). To capture a comprehensive list, we used the logical disjunction operator "or" between the signature words and between the bigram terms, whereas we used the logical conjunction operator "and" between the set of signature words and the set of bigram terms in the search. As a result, we recorded a total of 1266 papers for all 6 topics (average number (μ) of 211 papers per topic). In the second stage, we identified the relevant papers and filtered out the irrelevant ones by qualitatively assessing their titles, abstracts and key findings. The strategy for detecting relevant papers and excluding irrelevant ones was based on whether one or more of the key findings of a paper was about a university choice factor in relation to at least one of the signature words for each topic. At the end of this filtering process, we were down to 33 distinct relevant papers for all topics (μ =5.5). At this stage, we clustered the relevant papers together under each topic in accordance with their thematical representativeness to each corresponding topic's signature words. In the final stage of the SLR, we manually synthesized the key findings of these papers with the signature words of each topic and recorded a coherent decision theme for each topic (Table 5). Based on the signature words, we managed to derive a university choice theme from each single topic except for the first two topics which we merged to generate a semantically representative theme made from the signature words as well as supported by the key findings of the relevant literature. This involved a thorough examination of the literature to extract essential insights related to university choice factors, ensuring alignment with the themes initially identified through the LDA analysis. Papers were clustered under each topic based on their thematic relevance to the corresponding signature words, facilitating a focused analysis of the literature's contributions. This synthesis process demanded critical thinking and attention to detail, as we integrated information from multiple sources to construct meaningful decision themes. By aligning the literature's key findings with the signature words, the final output provided several representations of university choice factors. In the end, out of the signature words pertaining to 6 topics, 5 university choice themes were identified (Table 5): (1) Living and study costs, availability of scholarships and access to technology, research and facilities (buildings, libraries, science labs, etc.); (2) Ease of admission, entrance requirements and open communication with admissions staff; (3) Word-of-mouth in form of guidance, advice, suggestions, and influence (by family, friends and communities including current students and graduates); (4) Reputation, image and global ranking; and (5) Work and internship placements during study, job opportunities and potential work-related benefits after graduation.

We should note that due to the proximity of the signature words among some of the topics, same articles could at times be used to validate and refine the topics. However, we could derive a different choice factor for each topic due to the variety of insights being provided by these articles. We should also note that although cross-referencing signature words to literature was a vital step in validating the results of LDA and ensuring that the identified topics were meaningful and relevant to the original data, we consider it a laborious and iterative process that required meticulous interpretation and synthesis of information from multiple sources.

Topic #	Signature Words	Stage 1 # of all papers	Stage 2 Relevant papers	Stage 3 Key findings (factors that play an essential role in students' university choice)	University choice theme
1	university, national, scholarship, research, see, technology, using, library, support	228	Price et al. (2003) Najimudinova et al. (2022) Lombard (2012) Maringe (2006) Choi et al. (2019)	Learning, research, and other facilities Tuition, fees, scholarships Academic libraries Living and study costs Overall value/cost	Living and study costs, availability of scholarships and access to technology, research and facilities (buildings, libraries, science labs, etc.)
2	university, professor, world, building, national, first, one, science	221	Dunnett et al. (2012) Veloutsou et al. (2004) Imenda et al. (2004) Callender & Jackson (2008)	Fees and other costs Facilities and infrastructure Cost and availability of financial aid/scholarships Financial constraints	
3	university, day, campus, please, staff, student, online, health, week, open	218	Szekeres (2010) Briggs (2006) Brown et al. (2009) Imenda et al. (2004) Pasternak (2005) Retamosa et al. (2022)	Relationship building and open communication Ease of entry Communication with front-line staff Efficient enrolment Ease of admission and convenience Communication with admissions and teaching staff	Ease of admission, entrance requirements and open communication with admissions staff
4	university, research, people, times, video, according, likely, click, help	186	Broekemier & Seshadri (2000) Eldegwy et al. (2022) Mazzarol & Soutar (2002) Hemsley-Brown & Oplatka (2015) Bonnema & Van der Waldt (2008) Wut et al. (2022) Le et al. (2019) Le et al. (2020) Brown et al. (2009) Cho et al. (2008)	Parents' advice and guidance Similar-education parents' guidance Word-of-mouth referral Friends' suggestion and influence Social and community influence Electronic word-of-mouth and peer influence Electronic word-of-mouth from social media Parents' influence Communication with student ambassadors Parents' influence	Word-of-mouth in form of guidance, advice, suggestions, and influence (by family, friends and communities including current students and graduates)
5	university, professor, research, world, read, health, top, rankings, higher, global	202	Ackerman et al. (2022) Han (2014) Mazzarol & Soutar (2002) Najimudinova et al. (2022) Delavande & Zafar (2019) Siegfried & Getz (2006) Horstschräer (2012)	University prestige & fulfilment of students' self-image Branding and university image Reputation of the institution Quality of education and academic staff Nonpecuniary outcomes enjoyed at university Reputation and academic ranking Ranking and excellence status	Reputation, image and global ranking
6	university, study, year, bachelor, work, degree, like, time, now	211	Veloutsou et al. (2004) Hemsley-Brown & Oplatka (2015) Calderon & Sidhu (2013) Azzone & Soncin (2020) Lindahl & Regnér (2005) Holdsworth & Nind (2006) Ehrenberg (2020)	Part-time employment opportunities Part-time employment prospects Graduates' job status Job opportunities in the university region Graduates' earnings Job prospects Higher chances of finding a better job with better pay	Work and internship placements during study, job opportunities and potential work-related benefits after graduation

Table 5. Students' university choice themes (i.e., factors) based on the signature words identified via LDA from Facebook and Twitter, and their cross-referenced relevant literature.

3.3 Conclusion of LDA analysis

A topic modelling method such as Latent Dirichlet Allocation does not tell us how many topics there are in our corpus or their names. In this study, we used an LDA tuning algorithm and its associated metrics to provide a solution to determining an optimal number of topics for an LDA model. Due to the shortness of textual data in each document that we retrieved from Twitter and Facebook, we encountered highly similar or overlapping hence undifferentiating terms for topic identification. Lack of differentiating terms makes it quite difficult for researchers to demarcate topics from one another. Although there is no straightforward solution to this challenge, upon developing a beta matrix, we strived to handle this matter by conducting a 3-stage systematic literature review based on the signature words derived by the LDA model and cross-referencing these words to relevant literature. We found that albeit being time-consuming and iterative, in order to ensure cohesiveness and relevancy this was a necessary process. Finally, upon identifying five university choice factors, we found that topic identification out of short text from social media using LDA requires human judgment and discretion.

3.4 Limitations of LDA analysis and recommendations

One limitation of the LDA analysis is that since correlations among topics are not captured via Dirichlet distribution and sentence structure is not considered in LDA, due to its complete unsupervised nature, LDA does not allow for drawing inferences from more advanced NLP techniques such as sentiment or semantic analyses. Therefore, supervised Latent Dirichlet Allocation (sLDA) may be used to address this gap to some extent. The advantage of sLDA is that a variety of response types could be accommodated by the model as the unsupervised topic detection process strives not only to optimize the consistency of the topics underlying the data, but to maximize the model's capability to predict related values (Resnik et al., 2015). Ashktorab et al. (2014), for example, utilized sLDA to extract information about crisis events in North

America from Twitter, annotated each tweet and made this information available to first responders. After procuring sets of seed words from existing dictionary sets in literature and Amazon Mechanical Turk, Toubia et al. (2019) extracted the features of movies from their descriptions via sLDA to provide a recommendation classifier based on psychological topics. Although these sLDA models can potentially uncover latent structure better than standard LDA models, we should note that uncovering any latent structure is not a pressing concern in our case.

Another limitation is that since we assume document exchangeability, the evolution of topics over time is not considered a changing but a static factor in LDA. Dynamic topic modelling could be used to address this issue. Dynamic topic modelling was introduced by Blei and Lafferty (2006) to address the issue of not considering topic evolution in LDA over time. A dynamic topic model in principle considers the sequence of the documents and provides a posterior topical structure that is richer than LDA. Since a topic becomes a part of the order of distributions over words, rather than a single distribution over words, Blei and Lafferty (2006) could discover an underlying topic of the collection and monitor how it has transformed over time. Later, however, Blei (2012) stated that LDA is not suitable for sophisticated NLP goals such as language generation mainly because it does not consider the order of the words (bag of words: BOW) in a document. As improvement extensions to LDA, a topic model developed by Griffiths et al. (2004) shifts from LDA to a standard Hidden Markov Model and another model developed by Wallach (2006) relaxes the BOW assumption by considering the conditionality of consecutive words in topic generation. Although substantially expanding the parameter space, these models provide a better language modelling performance than the standard LDA (Blei, 2012). However, since the main goal of this study is to identify topics from short text rather than find variables and estimate relationships among variables, we believe that dynamic LDA models would not make any difference in our results.

Chapter 4

4. Topic identification via structural topic modelling

4.1 Background

4.1.1 Structural topic modelling

An alternative topic modelling approach to LDA is Structural Topic Modelling (STM). Through variational approximation, STM allows researchers to determine a topic model that contains document-level metadata in a fast and flexible way (Roberts et al., 2019). Building off LDA and its extensions, STM's key novelty is that it allows modellers to integrate arbitrary metadata, containing information about every document, into the topic model. Fresneda et al. (2021) found that inclusion of metadata by utilizing STM provides better text-grouping outcomes and reinforces richer segment profiles than can be obtained through traditional topic modelling techniques. STM can be used in a wide range of contexts and domains. Through STM, researchers modelled open-ended survey responses in political science (Roberts et al., 2014), explored perceived service quality attributes in Airbnb accommodation (Ding et al., 2020), predicted user sentiment towards chatbots (Sánchez-Franco et al., 2021), and analyzed free-text data on compliance indicators of Covid-19 guidelines (Wright et al., 2022).

As a semi-supervised statistical modelling technique, STM is different from unsupervised techniques such as LDA. Unlike LDA, for instance, STM incorporates covariates of interest in the prior distributions for topic-word allocations and document-topic proportions (Roberts et al., 2014). While collecting open-ended responses from people, researchers can include these covariates over which variance is expected rather than assume that topical content (i.e., words indicating a topic) and prevalence (i.e., frequency of topics being discussed) are constant across all respondents. Roberts et al. (2014) further summarized three key differences between the STM and LDA-based topic models. First, unlike in LDA, topic proportions (θ) for documents in STM can be correlated, second the topics' prevalence in STM may be impacted by a set of covariates X through a logistic regression model, therefore unlike in LDA where each document shares. global mean, in STM each document has its own prior distribution over topics, demarcated by covariate X. Third, since word distribution may contain a second set of covariate U, word use may vary through this covariate. These extra covariates may provide a way of re-shaping the prior distributions in the topic model, inserting valuable information into the inference process (Roberts et al., 2014).

The STM offers quick, flexible, visible, and replicable analyses that entail few a priori assumptions on the corpora (Roberts et al., 2014). However, it is a supervised machine learning method, and the analyst is a key component of interpreting the texts. The analyst's effort for understanding the text is directed by the model and the semantics of texts. However, as Grimmer and King (2011) posited, to discover insightful conceptualizations the STM can relieve the researcher of the burden of attempting to create a classification system from scratch and undertake the monotonous work of linking documents to their associated categories (Roberts et al., 2014).

A key advantage of STM for open-ended text analysis is the diversity of identifiable measures of interest beyond what LDA can offer (Roberts et al., 2014). In all topic models, for each document word proportions to each topic are estimated by the researcher, establishing a quantity of topic prevalence. The words that are likely to be associated with each topic are also measured by the model establishing topical content. Yet, since data collection in traditional LDA is unstructured, researchers must assume that each document is generated through the same data-creation process regardless of any extra information the researcher might have (Roberts et al., 2014). Unlike LDA, STM integrates metadata pertaining to the document and/or its author into the evaluation method. This enables analysts to estimate structured deviations in topical content (determined by U covariates) and topical prevalence (determined by X covariates) over the circumstances in an experiment as analysts may minimize uncertainty in approximating topic proportions through variational estimation of the posterior distribution (Roberts et al., 2014). It should be noted, though, that a structural topic model allows using covariates $X, U,$ both or neither. When there are no covariates and for β point estimates, the structural topic model shrinks to a (rapid) application of Blei and Lafferty's (2007) Correlated Topic Model (Roberts et al., 2019).

In STM, researchers have the option to pick covariates to integrate to the model. The selected covariates determine either the topical content or the topic prevalence in form of hidden variables with measured text data which may be provided by an open-ended survey participant. The researcher has the option to incorporate a covariate in the topical content (defined by U covariates) when he thinks that the measured covariate will impact the words (i.e., content) to be used in a respondent's discourse about a specific topic. Alternatively, the researcher may prefer to incorporate a covariate in the topical prevalence (defined by X covariates) portion of the model if he thinks the measured covariate will impact the extent (i.e., scale and scope) of the discourse about a specific topic (Roberts et al., 2014).

Researchers often prefer to use structural topic models to uncover topics from text and evaluate relationships between these topics and document metadata as model outputs can be used for hypothesis testing. (Roberts et al., 2019). For example, Rehs (2020) used STM and integrated paired cosine similarities to a linear regression framework to test a hypothesis. He also argued that due to the complex and dynamic nature of language embedded within text, topic models should not be aimed at understanding topics nor labelling them but discover meaningful relationships among topics and documents. Furthermore, upon LDA, Genovese (2015) used STM to estimate correlations among topics (i.e., religious, spiritual, and political) and external covariates at document level. Similarly, yet in the domain of the United States

Law, Law (2016) used STM to estimate to which extent constitutional preambles draw upon one of the three constitutional archetypes, namely liberal, statist, and universalist.

STM on social media data

Since 2014, numerous studies have been published using STM in a wide range of fields and domains to analyse social media data for various purposes.

In a Massive Open Online Course (MOOC) setting, Reich et al. (2014) used STM to identify topics in student discussion forums, as well as to map self-reported motivations of students and discover feedback patterns in course evaluations. Two years later, the researchers used STM to measure student engagement on similar forums and found correlations between students' engagement preferences during MOOCs and students' political beliefs (Reich et al., 2016). To build knowledge and understanding of dementia in aged care personnel through a free Understanding Dementia MOOC (UDMOOC), Doherty et al. (2020) used STM to identify themes from participants' open-ended responses and explored motivation for involvement and effectiveness of this course to meet the needs of carers and nurses in health sector.

Yang and Han (2021) used STM to identify topics from Twitter discussions about the challenges, fears and reactions about the Covid-19 pandemic in the hospitality industry. Likewise using STM, Han et al. (2021) identified and categorized topics generated by British news organizations and general public on Twitter. They found a significant agenda difference between the views of general public and news media's response to crises such as Covid-19 pandemic. Similarly, yet based on public opinions collected through an online questionnaire shared mostly on Facebook, Enria et al. (2021) used STM to perform a thematic analysis of the topics on the perceptions for the UK government's response to Covid-19. STM was also used to make time-series comparisons to gauge public opinions in relation to Covid-19. For example, Janmohamed et al. (2020) obtained a large number of documents from social media, forums and online blogs, then used STM to explore the evolution of semantic structures within topics about vaping and their word prevalence before and after the emergence of Covid-19 as it was first reported to WHO at the end of December 2019.

In a gender study, Garcia-Rudolph et al. (2019) identified and extracted main topics/themes from Twitter and classified them by gender. Researchers, then, allocated happiness scores to all the words representing the recognized themes. They finally compared them by gender and found that women's topics indicated higher levels of happiness scores. Mertens et al. (2019) used STM to explore gender bias in digital communication with politicians and found systematic gender disparities in tweets directed at politicians.

Mishler et al. (2015) used STM to identify and cluster Twitter users who had different political views (i.e., users sympathetic to Ukraine versus Russia). In a Norwegian study, Tvinnereim and Flottum (2015) used STM to explore public views on climate change. Their analysis revealed four different topics, namely Weather/Ice, Future/Impact, Money/Consumption and Attribution. Bail (2016) collected Facebook posts created by organizations advocating organ donation and used STM to classify Facebook users, as well as these organizations based on their strategies, resources, and wider external factors. Heft et al. (2022) used STM to analyze radical right parties' campaign agendas in their Facebook communication and identified a set of shared topics such as blaming elites and referring to a broad range of national actors, such as the media, banks, government, and other parties. Having extracted Facebook posts, Thorson et al. (2020) used STM to identify clusters in them about politics and policy issues in the community. Their analysis allowed them to discover patterns of post-topic allocations across and within the types of organizations they were investigating. Carrascosa et al. (2018) extracted text, comments and reactions from news stories, Facebook posts, tweets, as well as comments and reactions on European Union's cohesion policy, then used STM and sentiment analysis to understand and compare opposing opinions on this policy. Using STM, Ravenda et al. (2022)

found that on their official Facebook pages, municipalities in Italy usually post on 5 different topics and the prevalence of each topic is positively correlated with the most relevant municipal expense per capita. I et al. (2022) combined different text mining techniques one of which was an extension of LDA-STM to detect common topics from Twitter on plastic pollution and considered topic correlation besides topic coherence and prevalence. Parkinson et al. (2022) used STM to explore the thematic structure of online everyday talk mostly from Facebook and other online sources with UK domains over 27 months on the Scottish independence debate of 2012–2014.

Researchers have applied STM upon text extracted from other online platforms besides Facebook and Twitter. For example, Karkhanis et al. (2022) collected text from Glassdoor.com where employees rate and comment on their current or former employers. They used STM to compare employer branding parameters and to identify overarching dimensions across business cycles. Cripps et al. (2020) used STM to analyze tweets collected from interviewees Twitter accounts and identified 20 topics on innovation and technological developments in Europe. The topics demonstrated how technically progressive areas of Internet of Things (IoT) and big data have been prevalent among small and medium-sized businesses seeking innovation. Exploring online extremism in Japan, Zeyu (2019) used STM to profile and categorize online discussion pertaining to different ideological groups on Twitter and found that merely a small number extremists demonstrate a significant inclination to participate in discussions related to social or political issues.

Furthermore, upon social media data, STM was applied to identify and compare trending topics in various domains such as alternative and mainstream Lithuanian media (Mandravickaitė et al., 2020), alt-right and white supremacist movements on YouTube (Van der Vegt et al., 2021), students' online learning behaviour during the COVID-19 Pandemic (Lim & Lee, 2021), online reviews of wildlife tourism (Shang & Luo, 2022), national climate strategies and climate actions (Hsu et al., 2020), parties' campaign communication in European Parliament election (Heft et al., 2022), fast fashion brand-related corporate social responsibility agendas (Mickelsson et al., 2022), agricultural economics development (Cei et al., 2022), progress towards benchmarks in adaptation to climate change (Sietsma et al., 2021), and public expectations of the impact of Covid-19 on climate (Savin et al., 2022).

STM in higher education

Researchers have so far implemented various topic models including structural topic models in a wide range of fields and domains. Yet, from a marketing perspective of higher education institutions, the extent of this research was quite limited. For instance, Wijenayake et al. (2017) developed a neural network model to identify brand personalities for higher education institutions from social media data. However, no studies have hitherto used STM in the field of higher education marketing to analyze the content of college websites and social media posts.

4.2 Analysis via STM

4.2.1 Data mining, pre-processing, and coding

We used the same dataset (i.e., Tweets and Facebook posts from 13 universities), pre-processed the text (i.e., removed stopwords and numbers, transformed to lower letters) and calculated the optimal number of topics (k=6) as we did in LDA modelling explained in "Data analysis via LDA".

For structural topic modelling, we used the *stm* package on R because the package already has powerful built-in functionalities allowing users to explore topics in rich ways, determine uncertainty, and visualize quantities of interest (Roberts et al., 2019). In other words, *stm* package provides features that facilitate the workflow effectively in relation to topic discovery and statistical analysis of text data as it allows users to have a wide range of options to process raw text data, analyze the data from multiple aspects, and present results through a variety of informative graphing tools.

A typical workflow is displayed in Figure 10 with a heuristic description of the *stm* package. Different functions of the package that fulfil each task are listed for each step. Analysts first ingest (i.e., read and pre-process) text data and prepare (i.e., associate text with metadata) them for analysis. Then a structural topic model is computed. Finally, the results are evaluated (i.e., model selection and search). The capability of the package to evaluate the model swiftly allows for the estimation, interpretation, and visualization of the outcomes (Roberts et al., 2019).

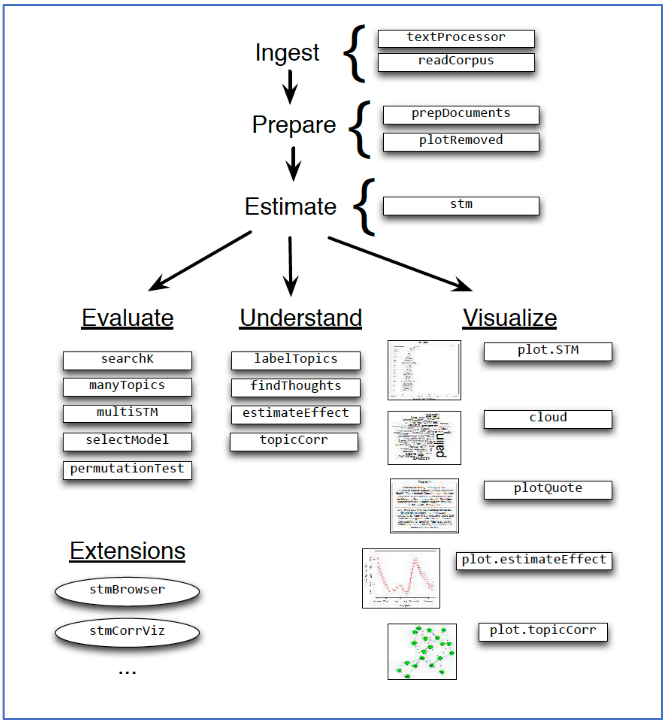

Figure 10: Heuristic description of the *stm* package workflow. Adapted from (Roberts et al., 2019).

As the STM allows topic models to be estimated with covariates at document-level, the *stm* package on R incorporates necessary tools for model identification, plotting, and topic-covariate regression estimation. The package has the functionalities to 1) ingest and process text data; 2) estimate Structural Topic Models; 3) evaluate covariate impacts on hidden topics with uncertainty; 4) build a plot of topic correlations; 5) estimate diagnostics for the structural topic model and summary metrics; 6) develop the graphs that Roberts et al. (2019) used in their various papers.

4.2.2 Text analysis via STM

Similar to LDA Analysis, textual data were collected from both Twitter and Facebook. We used Twitter Developer API v.2 on R v.3.5.1 to extract tweets comprising the mentions of 13 higher education institutions located in NSW and ACT between 2017 and 2022. For the same period, we extracted all the posts created by these institutions on their Facebook pages via Octoparse from Facebook Social Graph on R v.3.5.3. We collated all Facebook posts and tweets as lists and then removed all duplicate text (i.e., retweets and reposts) from the dataset. Eventually, each unique post or tweet is considered a document in further analyses. Therefore, upon removing duplicate text we have 8091 documents to be processed and analysed. We processed the textual data by removing punctuation, numbers, stopwords and whitespace. We also transformed the entire text to lower letters for consistency and developed a document-term matrix (DTM) on R with the *stm* (structural topic modelling) package.

Before developing a structural topic model, we labelled each document (i.e., Tweet or Facebook post) created by a university or mentioning a university with the name of that university. Therefore, the names of the 13 universities in NSW and ACT were the metadata that we associated with the textual data. Next, we looked into how important or in other words how 'distinguishing' a term (i.e., word) is in its corpus for each university. To do that, we created a tf-idf matrix vectorizing each term by multiplying the term's unit frequency by its document frequency. A term is considered important when it receives a higher tf-idf score than other terms meaning that it exists frequently in a document, but rarely in other documents.

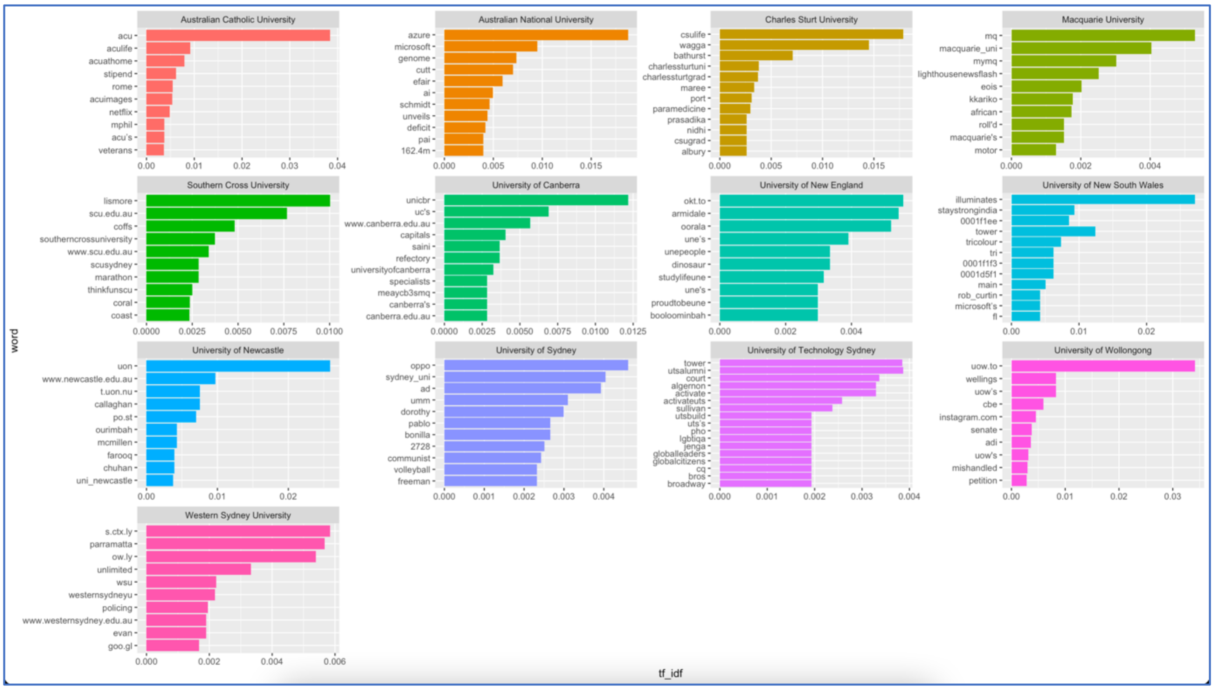

Figure 11: TF-IDF scores of the most important/distinguishing words for each university

As shown in Figure 11, we notice that the most important thus distinguishing words for each university are directly associated with the name of that university, including their acronyms (e.g., uts, unsw, une) and their social media or web extensions. We also notice some important words such as “oppo” under the University of Sydney and “azure” under the Australian National University. These words may indicate the use of technology by these universities in forthcoming analyses. Although based on their tf-idf scores identifying these important words may not be conducive to identifying main topics across the documents, upon building the structural topic models and analyzing results, we can look back into these words and discuss which institutions contribute more to the identified topics.

We notice that some of the most common words across all documents have little or no distinguishing impact (i.e., explanatory power) on identifying students’ university choice factors (Fig. 1). These words are usually made of the names and acronyms of the institutions (e.g., unsw, uts), or cities where these institutions are located (e.g., Canberra, Sydney, Wollongong) or generic terms such as “student”, “111odelling”, and “university”. Upon detecting these terms, we removed them altogether otherwise they would hold the places of other terms that were less common yet had more distinguishing effect for determining matriculation decision factors.

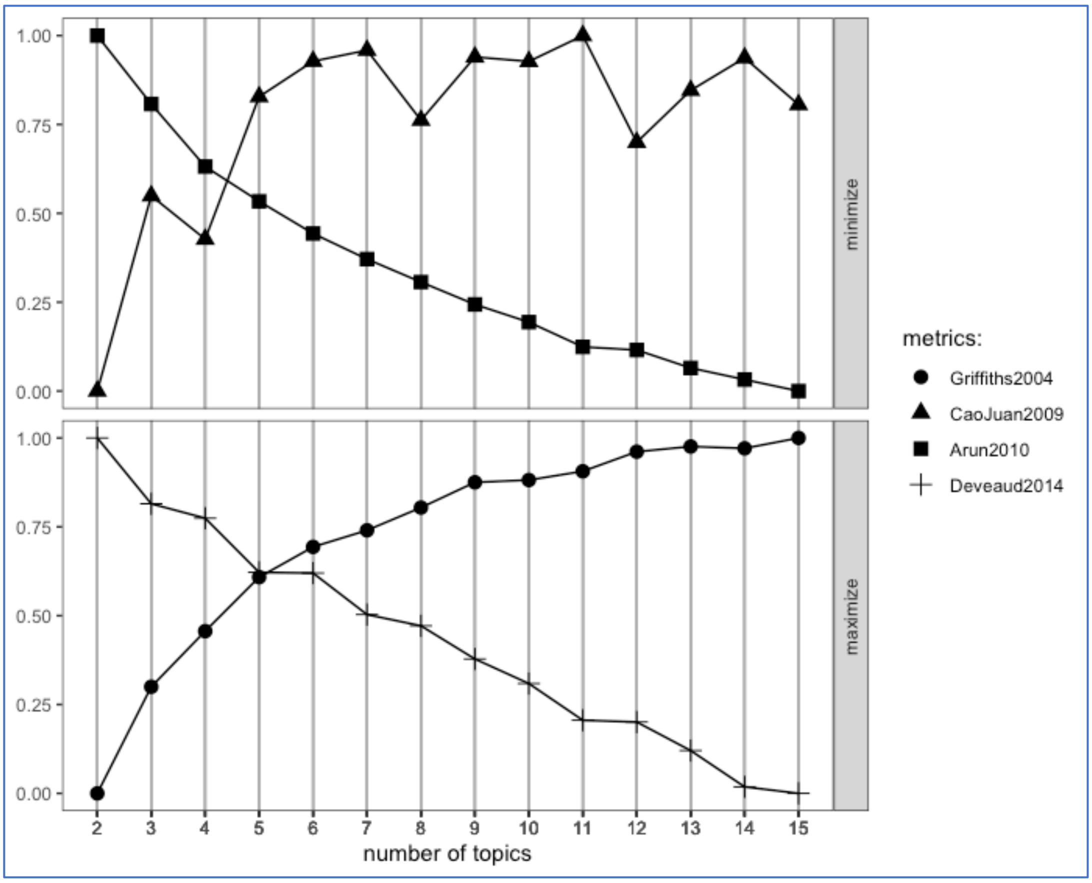

Figure 12: Topic number determination metrics' results for the STM model

When running parametric topic models, determining the number of topics is vital for researchers. While there is no universally agreed upon ‘best’ method for establishing the optimal number of topics in a model, having more topics can offer more detailed and nuanced representations of the data, but may also lead to less accurate estimates compared to having fewer topics. Furthermore, when working with a large number of documents, such as tweets which are often short, it’s important to consider the length of each document. In our analysis, we used Gibbs sampling to establish the number of topics (k) between 2 and 15, and employed four metrics (Fig. 2): “Griffiths2004”, “CaoJuan2009”, “Arun2010”, and “Deveaud2014”. To estimate the optimal number of topics for the structural topic model, we used *ldatuning* package [20] on R, tuned the model parameters and finally scored the model results. The optimal number of topics for the STM model can be identified by finding the parameters that minimize “CaoJuan2009” and “Arun2010” while maximizing “Griffiths2004” and “Deveaud2014”. We observe a steady decline in performance after 5-6 topics for both sets of metrics (Figure 12). Accordingly, we chose to include 6 topics in the model.

The *stm* package in R was used for structural topic 112odelling due to its built-in capabilities for exploring topics in depth, assessing uncertainty, and visualizing data of interest [2]. In short, the *stm* package streamlines the process of discovering and analyzing text data through its various options for processing raw text, analyzing data from multiple aspects, and presenting results through informative graphical tools.

When using one of the plotting functions of the *stm* package (Figure 13), we can observe that by plotting words (max 100) within all documents that have a topic proportion greater than a threshold (thresh) of 0.9, the top prominent words are “research”, “professor” and “campus”, while the second-tier words are “researchers”, “world, “time”, “week”, “community”, “congratulations”, “day”, program” and “school”, “and the third-tier ones are “read”, “team”, “staff”, “experience”, “technology” and “scholarships”, and so on. Using another plotting function of the *stm* package, we can observe that by plotting the top 6 topics based on the expected topic proportions of their most prominent words, dominant words such as “study”, “campus” and “research” have similar expected topic proportions in different topics (Figure 13). Therefore, we need to identify more words representative of each topic to be able to better differentiate between the topics.

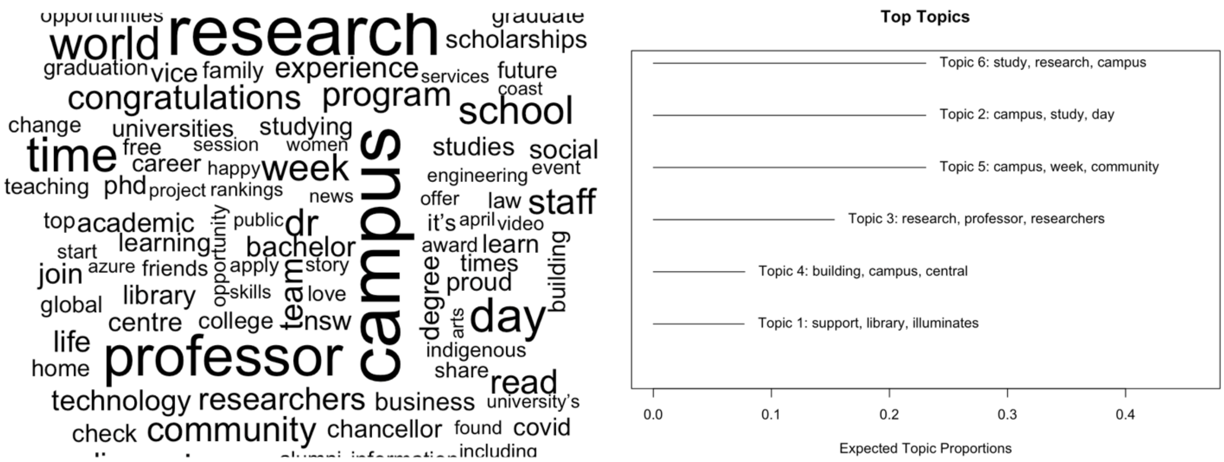

Figure 13: Wordcloud of all documents that have a topic proportion of higher than 0.9 threshold and the top 6 topics based on the expected topic proportions of their most prominent words

As discussed earlier in LDA analysis, a meta study by Silva et al. (2021) investigating the use of topic modelling in software engineering research found that most articles (75 out of 111) did not articulate how topics were named. It was also found that most papers (27 out of 36) that mentioned how names were allocated to topics had employed a manual approach deducting labels in accordance with the word frequencies in a topic and relying on researchers' discretion and semantic interpretation of the words in groups.

To identify university choice factors, we generated a beta matrix and plotted the probabilities of the top 10 words (signature words) per topic estimations (Figure 14). Once the STM model in our case returns the top ten words (i.e., signature words) which have the greatest probability of being incorporated by a matching topic, the university choice themes need to be identified by grouping these words in a logical and coherent way. However, this process is not straightforward. We opted for a manual approach to identify cohesive themes from the signature words across the 6 topics instead of relying on computational techniques such as k-means clustering or hierarchical clustering. The reason behind this decision is that the use of a manual approach allows us to explore how the signature words relate to one another despite the overlapping words among them, which leads to a better understanding of the underlying meaning. Blei and his colleagues have developed a simple yet effective method to achieving this. By manually assigning names to the topics discovered by the topic modelling algorithm, they can interpret and communicate the results in a way that resonates with their audience. They achieve this by carefully examining the most representative words in each topic and providing a high-level summary of the key themes discovered in the data. They then use this information to group the signature words and come up with a label that accurately represents the clustered words as a theme. This is utilized in software engineering research (Silva et al., 2021) to ensure that the findings are easily understandable and applicable to a wider audience. Similarly, by manually clustering the signature words we can uncover the relationships between them and identify the common thread that ties them together. This approach can provide us with a deeper insight into the data and help us identify the overarching themes more effectively.

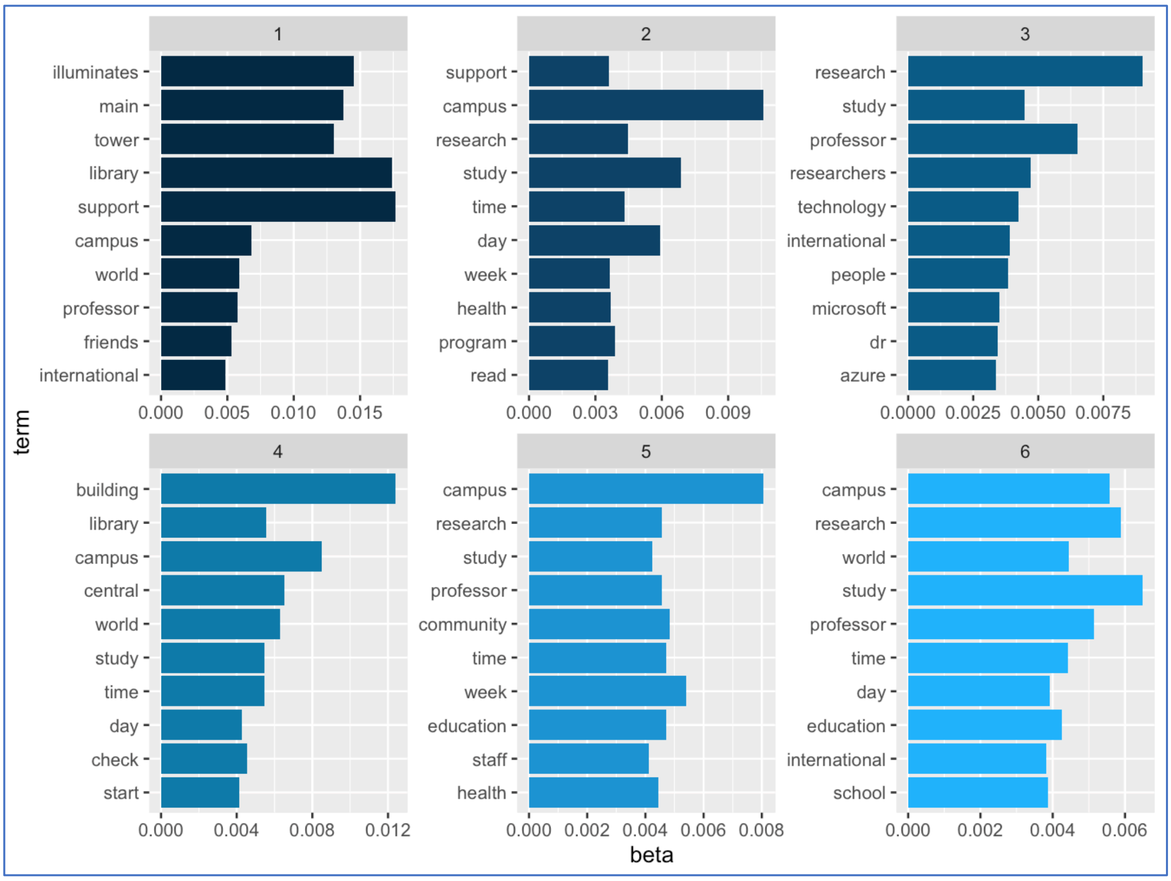

Figure 14: Word probability per topic estimations using “beta” matrix via STM modelling

It's worth noting that the process of identifying key themes from signature words in STM requires a human touch, as it relies on judgement and discretion. To improve the validity of the themes derived out of topics' signature words, researchers typically review existing literature based on the initial analysis and its output. Accordingly, we conducted a 3-stage systematic literature review (SLR) to identify the themes related to students' university choices based on signature words. In the first stage, we conducted a comprehensive search for articles published in databases such as ERIC and Business Source Ultimate, using each topic's signature words and essential bigram combinations of domain-specific keyword terms (e.g., "college", "university", "college choice/s", "university choice/s", "college enrolments", "university decision/s", "college decision/s). By using logical operators like "or" and "and" between the signature words and bigram terms, we captured a total of 1215 papers for all 6 topics. In the second stage, we filtered out the irrelevant papers by qualitatively assessing their titles, abstracts, and key findings. The relevant papers were identified based on whether they had one or more key findings related to a university choice factor in accordance with our interpretation of each topic's theme. We were left with 32 relevant papers for all topics. We then clustered the relevant papers together under each topic based on their thematic representativeness to its corresponding topic's signature words (Table 6).

Finally, in the last stage of the SLR, we manually synthesized the key findings of these papers with the signature words of each topic and recorded a coherent decision theme for each topic (Table 6). This approach allowed us to ensure a validated and fine-tuned transfer of the STM output through relevant literature to the key themes related to students' university choices. We identified a unique university choice theme for each of the single topics, except for the first and last two topics. To generate a semantically representative theme from the signature words and key findings of the relevant literature, we merged topics 1 and 2, as well as topics 5 and 6, resulting in two combined topics that each represent an interconnected university choice theme.

This approach allowed us to capture the underlying meaning of the signature words and key findings more accurately, providing a more comprehensive understanding of the data. In the end, out of the signature words representing 6 topics, 4 university choice themes were identified: (1) Availability, flexibility and attractiveness of the course/program of study (in line with career aspirations and earning potential) and on-campus support services; (2) Learning and research facilities, and use of technology (i.e., online, and social media channels) to communicate with potential students; (3) Campus location (proximity to home, convenience and comfort), its safety and physical appeal, and vibe of the city; (4) International reputation, image and prestige of the school, its professors, their research, service provided to the community (in terms of health and education) and the quality of the students produced.

Topic #	Signature Words	Stage 1 # of all papers	Stage 2 Relevant papers	Stage 3	
				Key findings (points/factors that play a role in students' university choice)	University choice theme
1	illuminates, main, tower, library, support, campus, world, professor, friends, international	202	Gille et al. (2022)	Academic programme and preparatory classes	Availability, flexibility and attractiveness of the course/program of study (in line with career aspirations and earning potential) and on-campus support services
			López-Bonilla et al. (2012)	Course content and program options	
			Pasternak (2005)	Course content and program options	
			Hoyt & Brown (2003)	Flexibility in course delivery times and methods	
			Mai Thi Ngoc & Thorpe (2015)	Course content, majors, credits	
			Soutar & Turner (2002)	Course type and suitability	
2	support, campus, research, study, time, day, week, health, program, read	204	Eldegwy et al. (2023)	Staff-new student interactions and subject-taster programs	
			Broekemier & Seshadri (2000)	Fit of the program of study	
			Columbu et al. (2021)	Attractiveness of the program of study	
			Hemsley-Brown & Oplatka (2015)	Fit of the course and support provided on campus	
			Holdsworth & Nind (2006)	Course flexibility	
3	research, study, professor, researchers, technology, international, people, 117odelling, dr, azure	187	Dunnett et al. (2012)	Communicating value propositions online	Learning and research facilities, and use of technology (i.e., online, and social media channels) to communicate with potential students
			Dao & Thorpe (2015)	Facilities and services	
			Walsh et al. (2015)	Campus facilities and online presence	
			Lombard (2012)	Academic libraries	
			Hemsley-Brown & Oplatka (2015)	Facilities, services and infrastructure	
			Bergerson (2009)	Online presence and campus facilities	
4	building, library, campus, central, world, study, time, day, check, start	223	Wilkins & Huisman (2011)	Physical appearance of campus and its proximity to home	Campus location (proximity to home, convenience and comfort), its safety and physical appeal, and vibe of the city
			Obermeit (2012)	Vibe of the university city	
			Li (2020)	Attractiveness of the university city	
			Hoyt & Howell (2012)	Campus size, safety and visual appeal	
			Syed et al. (2021)	Safety and comfort (mostly concerned by female students)	
			Broekemier & Seshadri (2000)	Safety (mostly concerned by parents)	
			Azzone & Soncin (2020)	Geographical proximity to home	
			Sá et al. (2012)	Proximity to home	
			Calitz et al. (2020)	Safety and security on campus	
			Bekaroglu (2021)	Proximity of home to the university city and major cities	
			Choi et al. (2019)	Location convenience	
5	campus, research, study, professor, community, time, week, education, staff, health	196	Pampaloni (2010)	Institutional image	International reputation, image and prestige of the school, its professors, their research, service provided to the community (in terms of health and education) and the quality of the students produced
			Cho et al. (2008)	Academic ranking and reputation	
			Cyrenne & Grant (2009)	Service provided to the community (in terms of health and education) and the quality of the students produced	
			Dunnett et al. (2012)	Course and institutional reputation	
			Azzone & Soncin (2020)	Reputation	
6	campus, research, world, study, professor, time, day, education, international, school	203	Obermeit (2012)	Reputation, quality and diversity of teaching	
			Cyrenne & Grant (2009)	Prestige of the university influenced by the calibre of research and the quality of graduates	
			Li (2020)	Image	
			Ackerman et al. (2022)	University prestige & fulfilment of students' self-image	
			Najimudinova et al. (2022)	Quality of education and academic staff	
			Horstschräer (2012)	Ranking and excellence status	

Table 6. Students' university choice themes (i.e., factors) based on the signature words for each of the topics identified via STM and their cross-referenced relevant literature.

We should note that although cross-referencing signature words to literature was a critical step in validating and fine-tuning the output of STM and ensuring that the identified topics were coherent and pertinent to the original data, we consider it a tedious and time-consuming process that required iterative interpretation and comprehensive synthesis of information from many sources.

Knowledge discovery via data mining and structural topic modelling is a fast-growing field that has gained a lot of attention particularly in recent years due to the exponentially growing amount of data being created, shared, and stored online. In higher education sector, mining social media data and discovering knowledge about students' university choice factors can be vital for universities to attract new students and retain existing ones. We used structural topic modelling upon big data collected from Facebook and Twitter to identify students' university choice factors. We also considered identifying topics (i.e., themes) in STM to be a manual process that depends on the researchers' intent and context of inquiry.

4.3 Limitations of STM analysis and recommendations

One major limitation of the study relates to the assumption made by STM that documents are exchangeable. This means that the model assumes that the topics within a document are independent of the order in which they appear, and that the order of the documents in the dataset does not impact the estimated topic distribution. This assumption may not hold true in some cases, particularly for datasets with temporal or spatial structure. For instance, in a dataset of social media articles, the topics discussed in one article may be temporally related to the topics discussed in the previous article, and the order of the articles may influence the estimated topic distribution. In such cases, STM may not be the most appropriate method to use as it may lead to a simplified representation of the topics and miss the subtle relationships among them. Another example where the assumption of exchangeability may not hold true is when there is spatial structure in the data. For example, a dataset of student reviews of different universities, where each university is a document, will have a structure that is not exchangeable. The topics of the reviews for a university will depend on the university itself and the order of the reviews will impact the estimated topic distribution. Alternative methods such as dynamic topic modelling or spatiotemporal topic modelling may be considered in these situations to relax the assumption of exchangeability and handle temporal or spatial structure in the data.

Another limitation is that we may not have been able to identify nuanced or subtle relationships among topics in the data, where two or more topics may be closely related but not be the same. Since STM uses a bag-of-words representation for the documents and ignores the order of the words in the documents, it may not be able to capture the subtle relationships that are conveyed by the order of the words in the documents, such as idiosyncrasies of natural language or sarcasm. In cases where identifying nuanced or subtle relationships between topics is a vital process, other methods such as LDA and Latent Semantic Analysis may be considered. These methods also employ a bag-of-words representation, but they use a different approach to estimate the topic distributions and hence may be better suited to detect hidden relationships among topics.

Chapter 5

5. Methodological pluralism for integrated topic identification: Complementing LDA and STM output with algorithmic document sequencing

5.1 Introduction

Methodological pluralism is the notion of negating the superiority of one "singular" method over others. In literature it has been the subject of research mostly in social sciences, physical sciences & technology, business & economics, and life sciences & biomedicine. However, despite the proliferation of research into knowledge discovery and data mining applications in the last decade, methodological pluralism has yet to be the subject of empirical studies with a specific focus on information retrieval and textual data mining from social media and literature.

Gaining a deeper and richer understanding of social phenomena by uncovering different layers and aspects of social reality has been the main premise of adopting a methodological pluralist approach for many researchers. As this approach refutes the superiority of any single method over others, it endorses different ways of retrieving information and discovering knowledge. Although to critics, methodological pluralism lacks a coherent organizing structure, to proponents, due to excluding methodological exclusivism (Scott & Marshall, 2009, p.466), confirmations and contradictions among those different layers revealed by multiple methods can be explored with no prejudice or predispositions. Critics rejected the idea of treating all methods as equal (Payne et al., 2004). However, as per methodological pluralism, denying methodological exclusivism does not mean accepting all methods as equal but finding utilitarian value in the diversity of the methods being used.

The debate about whether researchers shall use multiple methods or find and use the 'best' method for data collection has been going on for more than half a century. Many researchers were concerned about the fact that different methods produce different results to the same problems (LaPiere, 1934; Trow, 1957; Lever, 1981). Some researchers argued that one method is usually more suited to addressing a particular issue than other methods, and inferior methods (to the best one) generate less precise results (Zelditch Jr, 1962). Others suggested that, since all methods (including the best one) are technically flawed to some extent for collecting sufficient information about a research problem, using multiple methods with different sources of deficiencies allows for cross-validation of the results (Webb, 1966; Lever, 1981; Salzberg, 1997). Furthermore, Sieber (1973) argued that the true value of using multiple methods lies in the unique contribution each method makes to gaining a richer understanding of the matter at hand.

Numerous studies have hitherto incorporated both qualitative and quantitative (i.e., mixed) methods to carry out social research in a wide range of fields. Accordingly, researchers found numerous benefits of data collection through such mixed methods. One commonly discussed benefit is that data collected from multiple media within a single study potentially generate richer insights than those collected from a single medium for that study (Flynn et al., 2018; Smit et al., 2021). Such rich insights are usually generated by supporting, triangulating, or complementing quantitative data with qualitative data, qualitative data with quantitative data or qualitative data with additional qualitative data. However, social studies that use quantitative data from one source to supplement other quantitative data from another source are quite rare. Particularly in the field of textual data mining, information retrieval and knowledge discovery, there are no social studies that employ multiple quantitative techniques (e.g., topic models, systematic literature review) analyzing unstructured data collected from multiple sources (e.g., social media, literature) in a holistic way. To address this gap, under the premise of methodological pluralism, we use multiple quantitative methods to retrieve information from multiple sources to identify students' matriculation decision factors and aim to gain a richer and deeper understanding of their university choice process. To attain this goal, we employ two topic modelling techniques (i.e., Latent Dirichlet Allocation (LDA), and Structural Topic Modelling (STM)), and a systematic literature reviewing technique called Algorithmic Document Sequencing (ADS). While the data for topic models come from two of the most popular social media platforms (i.e., Facebook and Twitter), the data for ADS come from the relevant extant literature.

5.2 Aim and use

Given the unstructured and dispersed nature of the extant literature on the use of social media marketing for student recruitment in higher education, to address this gap we review the relevant literature and link published articles to one another through a method called "algorithmic document sequencing" (ADS).

In this chapter, we aim to improve a cohesive and unified understanding of the matter by complementing the output of LDA and STM, help researchers identify the connections among key findings of previous studies in an efficient way and recognize future research opportunities and offer marketing professionals essential strategic pointers to develop their institutions' brand and gain a competitive advantage. The goals of this section are thus fourfold:

1. To complement LDA and STM output with an integrated approach using a novel systematic literature reviewing technique.
2. To provide a novel systematic technique for reviewing and connecting the insights drawn from the use of social media marketing activities and developments in higher education literature.
3. To enhance the understanding of social media engagement with public to attract potential students.
4. To detect knowledge gaps and describe a future research agenda in this field.

To achieve these objectives, we first conducted a traditional yet unstructured literature review on social media marketing, engagement, and brand development in higher education. Then, we gathered 43 relevant published articles, extracted their textual data, summarized their key findings, and finally implemented ADS to provide a narrative synthesis of evidence. We did not test a specific hypothesis or generate new theories but strived to link thematically homogeneous documents to one another in a coherent way via ADS to identify topical similarities and form clusters among these studies with a focus on social media marketing for student recruitment.

Through ADS, insights drawn from relevant articles can be connected as the key findings from these articles sequenced in a structured way can improve our understanding of the matter at hand, help us identify future research opportunities and offer marketing professionals at universities key strategic pointers to build their brand. ADS can be applied to any domain in any field to link extant literature or other thematically homogeneous documents to each another for a systematically synthesized cohesive review. Therefore, valuable insights can efficiently be drawn from documents regardless of the total number or length of each corpus.

5.3 Methodology of chapter 5

5.3.1 Data collection

In the search for literature items, all published *Articles* written in English language from January 2012 to May 2022 were accessed through the University of Sydney's library search engine using the keywords: "social media marketing", "higher education", and "student recruitment" as search filters to appear in any field (i.e., title, author/creator, subject, ISBN, ISSN) (Figure 15). The search yielded 73 results in total. Regardless of their order of appearance, upon manually skimming through the abstracts of articles, 43 of them were found relevant to the main topic of interest (i.e., student recruitment in higher education via social media). The excluded articles were the ones pertaining to neither university student recruitment nor social media. In addition to their title, year, and authorship details, each article's entire textual data were recorded.

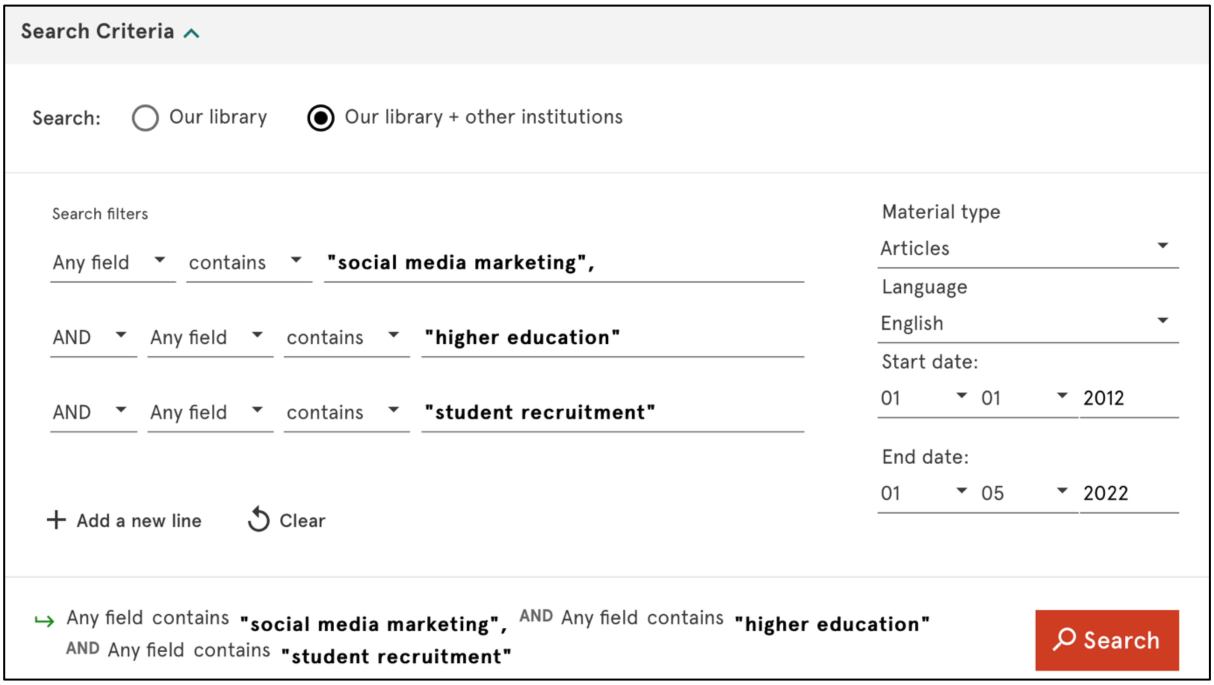

Figure 15: Search criteria for literature material

5.3.2 Algorithmic document sequencing

Although systematic literature reviews exist in extant marketing for higher education literature, none of them encompasses a distance-based statistical approach for document sequencing. Algorithmic document sequencing (ADS) is a novel method that we developed to automatically sequence documents in a structured way based on the similarity and frequency of the terms used in documents. Since this technique has no literary roots or predecessors, we have not included its background so far. In ADS, every document is linked to a preceding one, except for the first document because there is no document that comes before that. Therefore, the first step in ADS is to determine the first document to which the rest of the documents will consecutively be linked.

Figure 16: Wordcloud based on the unigram term frequencies of all corpora pertaining to 43 journal articles (documents).

We collated the entire text of the 43 articles (documents) excluding written and numerical data in articles' references, appendices, tables, figures, header, and footer. When we build a Wordcloud based on the term frequencies of all text, as shown in Figure 16, we notice that the most frequently used words such as "social", "media", "marketing", and "student" also exist in the title of this study: "Social media marketing for student recruitment." Since these terms will be representative of the thematic inquiry we aim to make, we look for the document (to be the first of the ADS) that contains these terms more than other documents. In identifying the very first document of the entire sequence, it is important to note that rather than looking for unigram terms such as "student", "social" and "media", we use bigram terms such as "student recruitment" or trigrams such as "social media marketing" to narrow down the specificity of the content in which we are interested to explore. We used Excel to count the occurrences of the bigram "student recruitment" and the trigram "social media marketing" in each document and ranked them in a descending order on Tableau. Given some articles may contain the acronym SMM to refer to Social Media Marketing, we added the count of the unigram "smm" to the total count of "social media marketing". The top 20 documents are shown in Figure 17. We found that Zhu (2019) had the highest number of counts in both "student recruitment" and the total of "student recruitment" and "social media marketing" + "smm". It is up to the discretion of the researcher to decide which document to pick if there is a close runner-up. In our case, since there is no such runner-up, we pick Zhu (2019) as the first document.

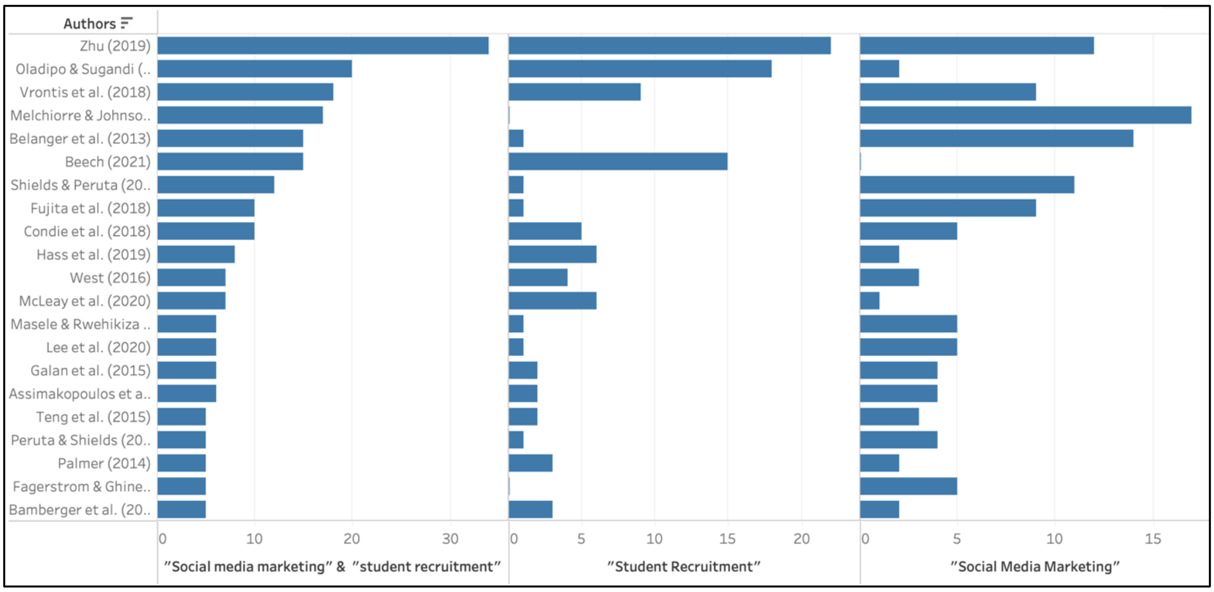

Figure 17: Top 20 documents containing the terms ["social media marketing" + "smm"] and "student recruitment."

After cleaning (i.e., converting to lower case, removing punctuation, numbers and stopwords, stripping whitespace and lemmatizing strings) the entire text data, we used cosine similarity method to measure the pairwise similarities of documents using orientation (angle between document vectors) on R Studio – R version 4.1.3. To combine the conventional model vector space with a singular value decomposition and then perform pairwise similarity comparisons, we used Latent Semantic Analysis (LSA) and the *cosine* function of the LSA package (Leydesdorff, 2005) and validated the output with the *sim2* function of the text2vec package on R. As a metric that works well in high dimensional spaces, cosine similarity (denoted as cos θ) is useful in our situation due to the high variation in the length of our documents. The formula for cos θ is:

$$\cos\theta = \frac{A \cdot B}{||A|| \cdot ||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

If θ is the angle between vector A and vector B, the smaller the angle (θ), the higher the cos θ and therefore the higher the similarity between document A and document B. However, it should be noted that cos θ values used for text matching in positive space [0:1] do not represent the similarity percentage between A and B. Therefore, if for instance cos θ between A and B is 0.75, it does not mean that A is 75% similar to B, because it is not the magnitude that is measured but merely the angle between A and B.

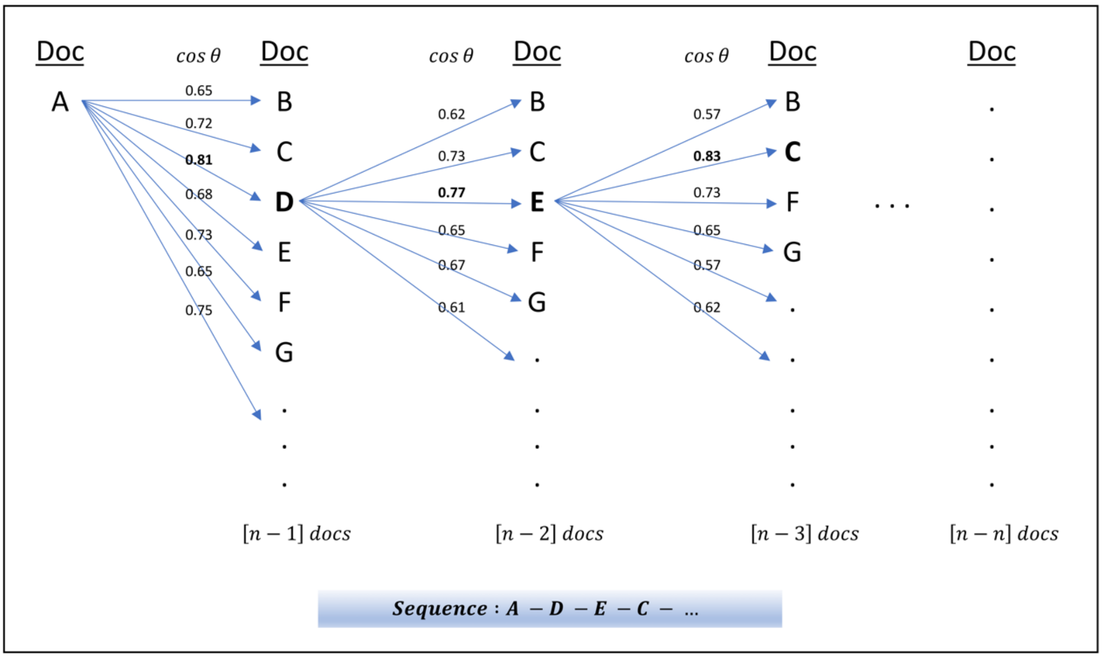

Figure 18: Example algorithmic document sequence of *n* number of documents, where cos θ is the cosine similarity between two documents and A is the first document. For demonstration purposes cos θ values have been fabricated. The document with the highest cos θ (in bold) is selected and placed subsequently after each document, altogether forming the final sequence.

As illustrated with an example ADS model in Figure 18, after the first document is identified (A), its cos θ with all the vectors of the rest of the documents in corpora are calculated. It should be noted that the cos θ values in Figure 18 are fabricated for demonstration purposes. Since the document with the highest cos θ with A is D (cos θ=0.81) among other cos θ values (i.e., 0.65, 0.72, 0.68, 0.73, 0.65, 0.75, …), document D will be placed to follow A. Likewise, the document with the highest cos θ with D is E (cos θ=0.77) among other cos θ values (i.e., 0.62, 0.73, 0.65, 0.67, 0.61, …), hence document E will follow D, and so on. It is important to note that every time a new document is placed in the sequence, that document must be removed from the corpora as after the first document we have [n-1] documents which will reduce by 1 document joining the sequence till the number of documents reaches [n-n] 0. Upon this, we know the sequence is complete.

In our case, for efficiency reasons rather than calculating the cos θ values one by one after the first document (i.e., Zhu (2019)), we calculated all cos θ values among all 43 documents (Figure 19). To identify each subsequent document, we laterally compared cos θ values and recorded the maximum one which pointed to the index of the subsequent document. For example, the max cos θ value in the row of the first document [Index #34: Zhu, 2019] is 0.138 which indicates to the document with index# 17 [West, 2016]. However, the drawback of using all cos θ values is that the max cos θ value might belong to a document that has already been used in the sequence. As a solution, we suggest that after every new document is added to the sequence, its corresponding cos θ values should be deleted (not laterally but vertically) from the data frame. Therefore, the new max cos θ value existing amongst the remainder of documents can be identified and its matching document can be linked to its preceding document in the sequence. Accordingly, based on the sequence output produced by the ADS (Figure 20), we linked all 43 documents to one another (Figure 21).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43
1	1.00000	0.10940	0.07455	0.08196	0.09952	0.07102	0.01516	0.07431	0.03007	0.01893	0.03032	0.07143	0.06772	0.09001	0.05440	0.09358	0.16825	0.02875	0.04937	0.04323	0.04728	0.03552	0.01621	0.04659	0.02173	0.09823	0.03102	0.02517	0.09718	0.49406	0.07871	0.06441	0.04999	0.02752	0.05365	0.08440	0.07700	0.03733	0.03218	0.02336	0.03964	0.05008	0.05296
2	0.10940	1.00000	0.05862	0.06017	0.07710	0.07816	0.01387	0.03766	0.03283	0.01664	0.04811	0.10988	0.07197	0.03044	0.08416	0.06183	0.07478	0.03884	0.06815	0.07163	0.03239	0.05186	0.05789	0.04795	0.02897	0.04405	0.08057	0.06223	0.08072	0.10446	0.04232	0.03517	0.04469	0.11776	0.05467	0.07567	0.05618	0.05710	0.02118	0.02224	0.06961	0.05750	0.03345
3	0.07455	0.05862	1.00000	0.04068	0.09994	0.11470	0.02865	0.06395	0.09545	0.06821	0.05228	0.05076	0.10300	0.03996	0.10577	0.15237	0.13343	0.03815	0.04832	0.03066	0.03473	0.05570	0.02383	0.12873	0.06887	0.05533	0.04382	0.02706	0.07343	0.03430	0.07750	0.12281	0.08377	0.04881	0.11698	0.06688	0.07382	0.03363	0.01542	0.04695	0.03593	0.05791	0.08860
4	0.08196	0.06017	0.04068	1.00000	0.06241	0.05064	0.01345	0.11128	0.01506	0.00910	0.01811	0.05191	0.02987	0.06092	0.03110	0.04785	0.12248	0.04169	0.08150	0.10850	0.02501	0.03095	0.01586	0.02845	0.00968	0.09442	0.02522	0.05468	0.04426	0.05351	0.05776	0.03387	0.03587	0.09274	0.02421	0.09298	0.03963	0.53830	0.01113	0.01487	0.02717	0.05423	0.03195
5	0.09952	0.07710	0.09994	0.06241	1.00000	0.12123	0.02449	0.05542	0.07932	0.14423	0.04298	0.06870	0.09699	0.05051	0.07933	0.09842	0.09713	0.04088	0.04469	0.04305	0.04271	0.06472	0.03001	0.08283	0.04935	0.05661	0.04685	0.03450	0.12112	0.05871	0.06806	0.08483	0.06938	0.04791	0.07928	0.10391	0.07885	0.04633	0.01915	0.03599	0.04694	0.05681	0.07239
6	0.07102	0.07816	0.11470	0.05064	0.12123	1.00000	0.05330	0.05466	0.13849	0.09344	0.07304	0.06783	0.11607	0.01872	0.09181	0.07498	0.06299	0.06015	0.07669	0.05039	0.05740	0.05999	0.04740	0.10880	0.09207	0.03874	0.06659	0.05286	0.08329	0.04266	0.05723	0.20838	0.09419	0.04532	0.11299	0.11736	0.07833	0.06485	0.03127	0.07723	0.04696	0.07842	0.09851
7	0.01516	0.01387	0.02865	0.01345	0.02449	0.05330	1.00000	0.00876	0.01258	0.01843	0.01568	0.02693	0.04575	0.00625	0.01838	0.03643	0.01815	0.01642	0.01541	0.01091	0.01200	0.02067	0.01575	0.02520	0.00895	0.01118	0.01086	0.00609	0.02875	0.00941	0.01659	0.01813	0.02144	0.01932	0.03188	0.02530	0.02057	0.01307	0.00348	0.00974	0.00913	0.01331	0.04324
8	0.07431	0.03766	0.06395	0.11128	0.05542	0.05466	0.00876	1.00000	0.01950	0.01408	0.02689	0.02883	0.06739	0.03280	0.04814	0.05460	0.07782	0.03660	0.03008	0.02740	0.03582	0.03019	0.02637	0.06806	0.03013	0.06224	0.06561	0.01358	0.03723	0.01869	0.07146	0.05900	0.03310	0.01771	0.04528	0.04768	0.04731	0.07778	0.04220	0.02677	0.02394	0.05446	0.04008
9	0.03007	0.03283	0.09545	0.01506	0.07932	0.13849	0.01258	0.01950	1.00000	0.52926	0.04672	0.02465	0.06547	0.00734	0.08135	0.02778	0.02244	0.04631	0.03666	0.01574	0.01786	0.03329	0.01495	0.06635	0.27119	0.01633	0.02406	0.01476	0.06242	0.01649	0.01744	0.49016	0.02864	0.05233	0.06248	0.04963	0.04573	0.02956	0.00914	0.07515	0.01465	0.02540	0.07029
10	0.01893	0.01664	0.06821	0.00910	0.14423	0.09344	0.01843	0.01408	0.52926	1.00000	0.01469	0.01795	0.04655	0.00437	0.03438	0.02778	0.01376	0.02984	0.03320	0.01155	0.01227	0.02148	0.01322	0.05107	0.19795	0.01043	0.02344	0.00902	0.03210	0.01430	0.00986	0.35999	0.01891	0.02884	0.03591	0.03001	0.02647	0.01980	0.00607	0.05538	0.01215	0.01585	0.03754
11	0.03032	0.04811	0.05228	0.01811	0.04298	0.07304	0.01568	0.02689	0.04672	0.01469	1.00000	0.03553	0.04322	0.02381	0.07833	0.03822	0.08390	0.02524	0.02787	0.01617	0.02417	0.04425	0.01546	0.04202	0.01798	0.01293	0.02889	0.01011	0.05544	0.02172	0.02947	0.03316	0.04206	0.01596	0.03966	0.06273	0.06211	0.01464	0.01110	0.01528	0.03877	0.03412	0.02813
12	0.07143	0.10988	0.05076	0.05191	0.06870	0.06783	0.02693	0.02883	0.02465	0.01795	0.03553	1.00000	0.05795	0.01044	0.09597	0.05913	0.03667	0.03546	0.07141	0.04857	0.27069	0.03803	0.05024	0.05353	0.02892	0.01820	0.08397	0.02559	0.09781	0.09371	0.03336	0.03486	0.04811	0.12346	0.04814	0.20087	0.05893	0.04379	0.03798	0.03832	0.07432	0.03429	0.04562
13	0.06772	0.07197	0.10300	0.02987	0.09699	0.11607	0.04575	0.06739	0.06547	0.04655	0.04322	0.05795	1.00000	0.01295	0.13433	0.04410	0.04465	0.04104	0.05469	0.04412	0.05095	0.04645	0.03008	0.11912	0.03821	0.02707	0.05403	0.02744	0.09183	0.03713	0.02273	0.06770	0.07332	0.06086	0.12755	0.07020	0.06475	0.04217	0.03381	0.03905	0.03162	0.03883	0.14256
14	0.09001	0.03044	0.03996	0.06092	0.05051	0.01872	0.00625	0.03280	0.00734	0.00437	0.02381	0.01044	0.01295	1.00000	0.01281	0.06230	0.20107	0.00485	0.00886	0.00736	0.01800	0.00789	0.00715	0.02031	0.00752	0.07511	0.00797	0.00495	0.02197	0.06564	0.07009	0.03370	0.05145	0.00591	0.01511	0.02182	0.01706	0.01512	0.00205	0.00853	0.03103	0.03878	0.01530
15	0.05440	0.08416	0.10577	0.03110	0.07933	0.09181	0.01838	0.04814	0.08135	0.03438	0.07833	0.09597	0.13433	0.01281	1.00000	0.04391	0.03026	0.04423	0.06021	0.04227	0.06766	0.05256	0.04367	0.10402	0.03328	0.01867	0.08137	0.02719	0.07398	0.03875	0.02282	0.05758	0.04851	0.10837	0.14235	0.05838	0.06029	0.04051	0.03725	0.03441	0.04535	0.03907	0.14712
16	0.09358	0.06183	0.15237	0.04785	0.09842	0.07498	0.03643	0.05460	0.02778	0.02778	0.03822	0.05913	0.04410	0.06230	0.04391	1.00000	0.14657	0.02626	0.04234	0.02390	0.02677	0.06732	0.01588	0.05809	0.02216	0.05701	0.02504	0.02347	0.07005	0.03037	0.07074	0.05466	0.05560	0.01774	0.04698	0.06984	0.07597	0.02369	0.00741	0.02726	0.04080	0.03613	0.03903
17	0.16825	0.07478	0.13343	0.12248	0.09713	0.06299	0.01815	0.07782	0.02244	0.01376	0.08390	0.03667	0.04465	0.20107	0.03026	0.14657	1.00000	0.03110	0.04016	0.04501	0.02863	0.03212	0.01977	0.04847	0.01898	0.15819	0.03054	0.05000	0.05596	0.03930	0.14045	0.06836	0.06269	0.13752	0.03512	0.06224	0.04632	0.04004	0.00490	0.03170	0.04489	0.08661	0.03596
18	0.02875	0.03884	0.03815	0.04169	0.04088	0.06015	0.01642	0.03660	0.04631	0.02984	0.02524	0.03546	0.04104	0.00485	0.04423	0.02626	0.03110	1.00000	0.05851	0.04919	0.01902	0.04850	0.02772	0.03816	0.01974	0.02644	0.02994	0.03821	0.04523	0.02445	0.01878	0.05602	0.03151	0.02372	0.03152	0.05011	0.04745	0.03277	0.00787	0.02551	0.01291	0.03465	0.04851
19	0.04937	0.06815	0.04832	0.08150	0.04469	0.07669	0.01541	0.03008	0.03666	0.03320	0.02787	0.07141	0.05469	0.00886	0.06021	0.04234	0.04016	0.05851	1.00000	0.07186	0.10014	0.04359	0.02183	0.05235	0.02266	0.01879	0.03873	0.03983	0.06850	0.03188	0.02013	0.03422	0.04973	0.03582	0.03734	0.15547	0.08873	0.07173	0.08491	0.02766	0.02404	0.04074	0.04260
20	0.04323	0.07163	0.03066	0.10850	0.04305	0.05039	0.01091	0.02740	0.01574	0.01155	0.01617	0.04857	0.04412	0.00736	0.04227	0.02390	0.04501	0.04919	0.07186	1.00000	0.03221	0.03652	0.04720	0.02724	0.01374	0.04531	0.02511	0.06000	0.07918	0.04150	0.02290	0.02377	0.04867	0.09831	0.02640	0.07423	0.03152	0.08561	0.01526	0.02584	0.01556	0.05872	0.03774
21	0.04728	0.03239	0.03473	0.02501	0.04271	0.05740	0.01200	0.03582	0.01786	0.01227	0.02417	0.27069	0.05095	0.01800	0.06766	0.02677	0.02863	0.01902	0.10014	0.03221	1.00000	0.03068	0.01065	0.03710	0.01329	0.01735	0.04474	0.01188	0.07180	0.01349	0.04104	0.02465	0.01974	0.02010	0.03150	0.08529	0.07118	0.03969	0.11185	0.04911	0.01927	0.02083	0.03806
22	0.03552	0.05186	0.05570	0.03095	0.06472	0.05999	0.02067	0.03019	0.03329	0.02148	0.04425	0.03803	0.04645	0.00789	0.05256	0.06732	0.03212	0.04850	0.04359	0.03652	0.03068	1.00000	0.02737	0.05502	0.02333	0.02504	0.03175	0.04160	0.08042	0.02759	0.01822	0.02914	0.04820	0.02196	0.04592	0.09275	0.04010	0.02571	0.01673	0.03486	0.03551	0.05307	0.03831
23	0.01621	0.05789	0.02383	0.01586	0.03001	0.04740	0.01575	0.02637	0.01495	0.01322	0.01546	0.05024	0.03008	0.00715	0.04367	0.01588	0.01977	0.02772	0.02183	0.04720	0.01065	0.02737	1.00000	0.03168	0.02118	0.02508	0.05041	0.04372	0.02875	0.03001	0.01611	0.02540	0.02309	0.07998	0.02504	0.02694	0.01197	0.01735	0.00686	0.02252	0.03387	0.04274	0.03083
24	0.04659	0.04795	0.12873	0.02845	0.08283	0.10880	0.02520	0.06806	0.06635	0.05107	0.04202	0.05353	0.11912	0.02031	0.10402	0.05809	0.04847	0.03816	0.05235	0.02724	0.03710	0.05502	0.03168	1.00000	0.03736	0.02621	0.05256	0.01945	0.06952	0.03060	0.04057	0.06791	0.10107	0.04213	0.54100	0.06513	0.06360	0.03312	0.02782	0.04713	0.02554	0.04016	0.10636
25	0.02173	0.02897	0.06887	0.00968	0.04935	0.09207	0.00895	0.03013	0.27119	0.19795	0.01798	0.02892	0.03821	0.00752	0.03328	0.02216	0.01898	0.01974	0.02266	0.01374	0.01329	0.02333	0.02118	0.03736	1.00000	0.01217	0.03100	0.01065	0.02287	0.02085	0.01188	0.33314	0.03312	0.04571	0.02867	0.03695	0.02898	0.01389	0.00832	0.05404	0.02145	0.02445	0.03219
26	0.09823	0.04405	0.05533	0.09442	0.05661	0.03874	0.01118	0.06224	0.01633	0.01043	0.01293	0.01820	0.02707	0.07511	0.01867	0.05701	0.15819	0.02644	0.01879	0.04531	0.01735	0.02504	0.02508	0.02621	0.01217	1.00000	0.02116	0.04679	0.02576	0.02997	0.08250	0.04959	0.03432	0.03854	0.02532	0.03212	0.03610	0.03928	0.00581	0.02704	0.02359	0.07495	0.02985
27	0.03102	0.08057	0.04382	0.02522	0.04685	0.06659	0.01086	0.06561	0.02406	0.02344	0.02889	0.08397	0.05403	0.00797	0.08137	0.02504	0.03054	0.02994	0.03873	0.02511	0.04474	0.03175	0.05041	0.05256	0.03100	0.02116	1.00000	0.03692	0.04602	0.04089	0.04920	0.03266	0.02860	0.12270	0.04367	0.05685	0.02923	0.02337	0.03910	0.03066	0.05550	0.03885	0.02931
28	0.02517	0.06223	0.02706	0.05468	0.03450	0.05286	0.00609	0.01358	0.01476	0.00902	0.01011	0.02559	0.02744	0.00495	0.02719	0.02347	0.05000	0.03821	0.03983	0.06000	0.01188	0.04160	0.04372	0.01945	0.01065	0.04679	0.03692	1.00000	0.03089	0.03511	0.01191	0.02454	0.03225	0.09372	0.02025	0.03547	0.03053	0.04452	0.01268	0.01449	0.01653	0.07363	0.02890
29	0.09718	0.08072	0.07343	0.04426	0.12112	0.08329	0.02875	0.03723	0.06242	0.03210	0.05544	0.09781	0.09183	0.02197	0.07398	0.07005	0.05596	0.04523	0.06850	0.07918	0.07180	0.08042	0.02875	0.06952	0.02287	0.02576	0.04602	0.03089	1.00000	0.06615	0.04306	0.05380	0.06451	0.02900	0.07121	0.12886	0.07868	0.06877	0.06146	0.02902	0.03358	0.04777	0.06721
30	0.49406	0.10446	0.03430	0.05351	0.05871	0.04266	0.00941	0.01869	0.01649	0.01430	0.02172	0.09371	0.03713	0.06564	0.03875	0.03037	0.03930	0.02445	0.03188	0.04150	0.01349	0.02759	0.03001	0.03060	0.02085	0.02997	0.04089	0.03511	0.06615	1.00000	0.05205	0.02703	0.04624	0.07006	0.02827	0.07827	0.03975	0.06818	0.00886	0.01222	0.03519	0.02934	0.02889
31	0.07871	0.04232	0.07750	0.05776	0.06806	0.05723	0.01659	0.07146	0.01744	0.00986	0.02947	0.03336	0.02273	0.07009	0.02282	0.07074	0.14045	0.01878	0.02013	0.02290	0.04104	0.01822	0.01611	0.04057	0.01188	0.08250	0.04920	0.01191	0.04306	0.05205	1.00000	0.03956	0.05576	0.01168	0.04492	0.08902	0.07052	0.01770	0.00749	0.03628	0.02203	0.04119	0.02513
32	0.06441	0.03517	0.12281	0.03387	0.08483	0.20838	0.01813	0.05900	0.49016	0.35999	0.03316	0.03486	0.06770	0.03370	0.05758	0.05466	0.06836	0.05602	0.03422	0.02377	0.02465	0.02914	0.02540	0.06791	0.33314	0.04959	0.03266	0.02454	0.05380	0.02703	0.03956	1.00000	0.04723	0.04038	0.07526	0.06183	0.04413	0.03298	0.01822	0.12694	0.02646	0.04229	0.05848
33	0.04999	0.04469	0.08377	0.03587	0.06938	0.09419	0.02144	0.03310	0.02864	0.01891	0.04206	0.04811	0.07332	0.05145	0.04851	0.05560	0.06269	0.03151	0.04973	0.04867	0.01974	0.04820	0.02309	0.10107	0.03312	0.03432	0.02860	0.03225	0.06451	0.04624	0.05576	0.04723	1.00000	0.02770	0.12616	0.09290	0.05265	0.03128	0.00916	0.03153	0.02111	0.04736	0.07095
34	0.02752	0.11776	0.04881	0.09274	0.04791	0.04532	0.01932	0.01771	0.05233	0.02884	0.01596	0.12346	0.06086	0.00591	0.10837	0.01774	0.13752	0.02372	0.03582	0.09831	0.02010	0.02196	0.07998	0.04213	0.04571	0.03854	0.12270	0.09372	0.02900	0.07006	0.01168	0.04038	0.02770	1.00000	0.04754	0.02851	0.02701	0.07474	0.01407	0.01611	0.08170	0.04190	0.08260
35	0.05365	0.05467	0.11698	0.02421	0.07928	0.11299	0.03188	0.04528	0.06248	0.03591	0.03966	0.04814	0.12755	0.01511	0.14235	0.04698	0.03512	0.03152	0.03734	0.02640	0.03150	0.04592	0.02504	0.54100	0.02867	0.02532	0.04367	0.02025	0.07121	0.02827	0.04492	0.07526	0.12616	0.04754	1.00000	0.04721	0.05146	0.03017	0.02051	0.04120	0.01814	0.02688	0.13955
36	0.08440	0.07567	0.06688	0.09298	0.10391	0.11736	0.02530	0.04768	0.04963	0.03001	0.06273	0.20087	0.07020	0.02182	0.05838	0.06984	0.06224	0.05011	0.15547	0.07423	0.08529	0.09275	0.02694	0.06513	0.03695	0.03212	0.05685	0.03547	0.12886	0.07827	0.08902	0.06183	0.09290	0.02851	0.04721	1.00000	0.13610	0.06193	0.04443	0.03104	0.03682	0.04989	0.05947
37	0.07700	0.05618	0.07382	0.03963	0.07885	0.07833	0.02057	0.04731	0.04573	0.02647	0.06211	0.05893	0.06475	0.01706	0.06029	0.07597	0.04632	0.04745	0.08873	0.03152	0.07118	0.04010	0.01197	0.06360	0.02898	0.03610	0.02923	0.03053	0.07868	0.03975	0.07052	0.04413	0.05265	0.02701	0.05146	0.13610	1.00000	0.02678	0.03958	0.03400	0.02277	0.02971	0.04706
38	0.03733	0.05710	0.03363	0.53830	0.04633	0.06485	0.01307	0.07778	0.02956	0.01980	0.01464	0.04379	0.04217	0.01512	0.04051	0.02369	0.04004	0.03277	0.07173	0.08561	0.03969	0.02571	0.01735	0.03312	0.01389	0.03928	0.02337	0.04452	0.06877	0.06818	0.01770	0.03298	0.03128	0.07474	0.03017	0.06193	0.02678	1.00000	0.03638	0.01695	0.01760	0.04197	0.03663
39	0.03218	0.02118	0.01542	0.01113	0.01915	0.03127	0.00348	0.04220	0.00914	0.00607	0.01110	0.03798	0.03381	0.00205	0.03725	0.00741	0.00490	0.00787	0.08491	0.01526	0.11185	0.01673	0.00686	0.02782	0.00832	0.00581	0.03910	0.01268	0.06146	0.00886	0.00749	0.01822	0.00916	0.01407	0.02051	0.04443	0.03958	0.03638	1.00000	0.00895	0.00755	0.01356	0.02930
40	0.02336	0.02224	0.04695	0.01487	0.03599	0.07723	0.00974	0.02677	0.07515	0.05538	0.01528	0.03832	0.03905	0.00853	0.03441	0.02726	0.03170	0.02551	0.02766	0.02584	0.04911	0.03486	0.02252	0.04713	0.05404	0.02704	0.03066	0.01449	0.02902	0.01222	0.03628	0.12694	0.03153	0.01611	0.04120	0.03104	0.03400	0.01695	0.00895	1.00000	0.01655	0.04561	0.03650
41	0.03964	0.06961	0.03593	0.02717	0.04694	0.04696	0.00913	0.02394	0.01465	0.01215	0.03877	0.07432	0.03162	0.03103	0.04535	0.04080	0.04489	0.01291	0.02404	0.01556	0.01927	0.03551	0.03387	0.02554	0.02145	0.02359	0.05550	0.01653	0.03358	0.03519	0.02203	0.02646	0.02111	0.08170	0.01814	0.03682	0.02277	0.01760	0.00755	0.01655	1.00000	0.03803	0.01995
42	0.05008	0.05750	0.05791	0.05423	0.05681	0.07842	0.01331	0.05446	0.02540	0.01585	0.03412	0.03429	0.03883	0.03878	0.03907	0.03613	0.08661	0.03465	0.04074	0.05872	0.02083	0.05307	0.04274	0.04016	0.02445	0.07495	0.03885	0.07363	0.04777	0.02934	0.04119	0.04229	0.04736	0.04190	0.02688	0.04989	0.02971	0.04197	0.01356	0.04561	0.03803	1.00000	0.04273
43	0.05296	0.03345	0.08860	0.03195	0.07239	0.09851	0.04324	0.04008	0.07029	0.03754	0.02813	0.04562	0.14256	0.01530	0.14712	0.03903	0.03596	0.04851	0.04260	0.03774	0.03806	0.03831	0.03083	0.10636	0.03219	0.02985	0.02931	0.02890	0.06721	0.02889	0.02513	0.05848	0.07095	0.08260	0.13955	0.05947	0.04706	0.03663	0.02930	0.03650	0.01995	0.04273	1.00000

Figure 19: 43X43 Matrix of the cos θ values of among the whole corpora. Cos θ values have been truncated to four decimal points.

Authors	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind	Max	Ind
Assimakopoulos et al. (2017)	1	0.494	1	0.494	1	0.494	1	0.494
Vrontis et al. (2018)	2	0.118	2	0.118	2	0.118	2	0.118	2	0.1178	2	0.1099
Chen & DiVall (2017)	3	0.152	3	0.152	3	0.152	3	0.152	3	0.1524	3	0.1524	3	0.1524	3	0.1524	3	0.1524	3	0.1524	3	0.1524	3	0.1524	3	0.1524	3	0.1524	3	0.1524	3	0.1524	3	0.1524	3	0.1524	3	0.1524	3	0.1524	3	0.152	3	0.152	3	0.152	3	0.152
Teng & Khong (2015)	4	0.538	4	0.538	4	0.538	4	0.538	4	0.5383	4	0.5383	4	0.5383	4	0.5383	4	0.5383	4	0.5383	4	0.5383	4	0.5383	4	0.5383	4	0.5383	4	0.5383	4	0.5383	4	0.5383	4	0.5383	4	0.5383	4	0.5383	4	0.538	4	0.538	4	0.538	4	0.538	4	0.538	4	0.538	4	0.538	4	0.538
Masele & Rwehikiza (2021)	5	0.144	5	0.144	5	0.144	5	0.144	5	0.1442	5	0.1442	5	0.1442	5	0.1442	5	0.1442	5	0.1442	5	0.1442	5	0.1442	5	0.1442
Belanger et al. (2013)	6	0.208	6	0.208	6	0.208	6	0.208	6	0.2084	6	0.2084	6	0.2084	6	0.2084	6	0.2084	6	0.2084	6	0.2084	6	0.2084	6	0.2084	6	0.2084	6	0.2084	6	0.2084	6	0.2084	6	0.1161
Taylor & Bicak (2020)	7	0.053	7	0.053	7	0.053	7	0.053	7	0.0533	7	0.0533	7	0.0533	7	0.0533	7	0.0533	7	0.0533	7	0.0533	7	0.0533	7	0.0533	7	0.0533	7	0.0533	7	0.0533	7	0.0533	7	0.0457	7	0.0457	7	0.0457	7	0.036	7	0.036	7	0.036	7	0.036	7	0.036	7	0.029	7	0.029	7	0.029	7	0.029	7	0.029	7	0.029	7	0.0214	7	0.0214	7	0.0214	7	0.0214	7	0.0214	7	0.0214	7	0.0214	7	0.0214	7	0.0157	7	0.0097	7	0.0097	7
Fujita et al. (2018)	8	0.111	8	0.111	8	0.111	8	0.111	8	0.1113	8	0.1113	8	0.1113	8	0.1113	8	0.1113	8	0.1113	8	0.1113	8	0.1113	8	0.1113	8	0.1113	8	0.1113	8	0.1113	8	0.1113	8	0.1113	8	0.1113	8	0.1113	8	0.111	8	0.111	8	0.111	8	0.111	8	0.111	8	0.111	8	0.111	8	0.111	8	0.078	8	0.066	8	0.066	8	0.0656	8	0.0656	8	0.0656	8	0.0656	8	0.0656	8	0.0366
Palmer (2013)	9	0.529	9	0.529	9	0.529	9	0.529	9	0.5293	9	0.5293	9	0.5293	9	0.5293	9	0.5293	9	0.5293	9	0.5293	9	0.5293	9	0.5293	9	0.5293	9	0.4902
Palmer (2014)	10	0.529	10	0.529	10	0.529	10	0.529	10	0.5293	10	0.5293	10	0.5293	10	0.5293	10	0.5293	10	0.5293	10	0.5293	10	0.5293	10	0.5293	10	0.5293
Fagerstrom & Ghinea (2013)	11	0.084	11	0.084	11	0.084	11	0.084	11	0.0839	11	0.0839	11	0.0839	11	0.0839	11	0.0839	11	0.0839	11	0.0839	11	0.0839	11	0.0839	11	0.0839	11	0.0839	11	0.0839	11	0.0839	11	0.0783	11	0.0783	11	0.0783	11	0.055	11	0.055	11	0.055	11	0.055	11	0.055	11	0.055	11	0.055	11	0.055	11	0.055	11	0.055	11	0.055	11	0.0443	11	0.0421	11	0.0421	11	0.0421	11	0.0421	11	0.0421	11	0.0421	11	0.0421	11	0.0388
Royo-Vela & Hunermund (2016)	12	0.271	12	0.271	12	0.271	12	0.271	12	0.2707	12	0.2707	12	0.2707
Quijada et al. (2021)	13	0.143	13	0.143	13	0.143	13	0.143	13	0.1426	13	0.1426	13	0.1426	13	0.1426	13	0.1426	13	0.1426	13	0.1426	13	0.1426	13	0.1426	13	0.1426	13	0.1426	13	0.1426	13	0.1426	13	0.1426	13	0.1426
Pavlik (2021)	14	0.201	14	0.201	14	0.090
Eger et al. (2021)	15	0.147	15	0.147	15	0.147	15	0.147	15	0.1471	15	0.1471	15	0.1471	15	0.1471	15	0.1471	15	0.1471	15	0.1471	15	0.1471	15	0.1471	15	0.1471	15	0.1471	15	0.1471	15	0.1471	15	0.1471	15	0.1471	15	0.1471	15	0.142
Melchiorre & Johnson (2017)	16	0.152	16	0.152	16	0.152	16	0.152	16	0.1524	16	0.1524	16	0.1524	16	0.1524	16	0.1524	16	0.1524	16	0.1524	16	0.1524	16	0.1524	16	0.1524	16	0.1524	16	0.1524	16	0.1524	16	0.1524	16	0.1524	16	0.1524	16	0.152	16	0.152	16	0.152	16	0.152	16	0.071
West (2016)	17	0.201	17	0.201
Mwenda et al. (2019)	18	0.060	18	0.060	18	0.060	18	0.060	18	0.0602	18	0.0602	18	0.0602	18	0.0602	18	0.0602	18	0.0602	18	0.0602	18	0.0602	18	0.0602	18	0.0602	18	0.0602	18	0.0602	18	0.0602	18	0.0492	18	0.0492	18	0.0492	18	0.049	18	0.049	18	0.049	18	0.049	18	0.049	18	0.049	18	0.049	18	0.049	18	0.049	18	0.049	18	0.048	18	0.0485	18	0.0382	18	0.0382	18	0.0366	18	0.0366	18	0.0366	18	0.0315
Le et al. (2019)	19	0.155	19	0.155	19	0.155	19	0.155	19	0.1555	19	0.1555	19	0.1555	19	0.1555	19	0.1555	19	0.1555
McLeay et al. (2020)	20	0.109	20	0.109	20	0.109	20	0.109	20	0.1085	20	0.1085	20	0.1085	20	0.1085	20	0.1085	20	0.1085	20	0.1085	20	0.1085	20	0.1085	20	0.1085	20	0.1085	20	0.1085	20	0.1085	20	0.1085	20	0.1085	20	0.1085	20	0.109	20	0.109	20	0.109	20	0.109	20	0.109	20	0.109	20	0.109	20	0.109	20	0.086	20	0.079
Sagynbekova et al. (2020)	21	0.271	21	0.271	21	0.271	21	0.271	21	0.2707	21	0.2707	21	0.2707	21	0.1119
Komljenovic (2019)	22	0.093	22	0.093	22	0.093	22	0.093	22	0.0927	22	0.0927	22	0.0927	22	0.0927	22	0.0927	22	0.0927	22	0.0927	22	0.0927	22	0.0927	22	0.0927	22	0.0927	22	0.0927	22	0.0927	22	0.0804	22	0.0804	22	0.0804	22	0.080	22	0.080	22	0.080	22	0.080	22	0.080	22	0.080	22	0.080	22	0.080	22	0.080	22	0.080	22	0.080	22	0.0531
Bamberger et al. (2020)	23	0.080	23	0.080	23	0.080	23	0.080	23	0.0800	23	0.0800	23	0.0800	23	0.0800	23	0.0800	23	0.0800	23	0.0800	23	0.0800	23	0.0800	23	0.0800	23	0.0800	23	0.0800	23	0.0800	23	0.0504	23	0.0504	23	0.0504	23	0.050	23	0.050	23	0.050	23	0.050	23	0.050	23	0.050	23	0.050	23	0.050	23	0.050	23	0.050	23	0.050	23	0.0504	23	0.0504	23	0.0504	23	0.0504
Peruta & Shields (2018)	24	0.541	24	0.541	24	0.541	24	0.541	24	0.5410	24	0.5410	24	0.5410	24	0.5410	24	0.5410	24	0.5410	24	0.5410	24	0.5410	24	0.5410	24	0.5410	24	0.5410	24	0.5410	24	0.5410	24	0.5410	24	0.5410	24	0.5410	24	0.541	24	0.541	24	0.129
Condie et al. (2018)	25	0.333	25	0.333	25	0.333	25	0.333	25	0.3331	25	0.3331	25	0.3331	25	0.3331	25	0.3331	25	0.3331	25	0.3331	25	0.3331	25	0.3331	25	0.3331	25	0.3331	25	0.3331	25	0.0921
Lee (2019)	26	0.158	26	0.158	26	0.158	26	0.158	26	0.1582	26	0.1582	26	0.1582	26	0.1582	26	0.1582	26	0.1582	26	0.1582	26	0.1582	26	0.1582	26	0.1582	26	0.1582	26	0.1582	26	0.1582	26	0.0944	26	0.0944	26	0.0944	26	0.094	26	0.094	26	0.094	26	0.094	26	0.094	26	0.094	26	0.094
Bolat & O'Sullivan (2017)	27	0.123	27	0.123	27	0.123	27	0.123	27	0.1227	27	0.1227	27	0.1227	27	0.1227	27	0.1227	27	0.1227	27	0.1227	27	0.1227	27	0.1227	27	0.1227	27	0.1227	27	0.1227	27	0.1227	27	0.0814	27	0.0814	27	0.0814	27	0.066	27	0.066	27	0.066	27	0.066	27	0.066	27	0.066	27	0.066	27	0.066	27	0.066	27	0.066	27	0.066	27	0.0656	27	0.0656	27	0.0656	27	0.0656	27	0.0656
Oladipo & Sugandi (2021)	28	0.094	28	0.094	28	0.094	28	0.094	28	0.0937	28	0.0937	28	0.0937	28	0.0937	28	0.0937	28	0.0937	28	0.0937	28	0.0937	28	0.0937	28	0.0937	28	0.0937	28	0.0937	28	0.0937	28	0.0736	28	0.0736	28	0.0736	28	0.074	28	0.074	28	0.074	28	0.074	28	0.074	28	0.074	28	0.074	28	0.074	28	0.074	28	0.074	28	0.074	28	0.0736	28	0.0736	28	0.0437
Mishra & Madaan (2020)	29	0.129	29	0.129	29	0.129	29	0.129	29	0.1289	29	0.1289	29	0.1289	29	0.1289	29	0.1289	29	0.1289	29	0.1289	29	0.1289	29	0.1289	29	0.1289	29	0.1289	29	0.1289	29	0.1289	29	0.0918	29	0.0918	29	0.0918	29	0.080	29	0.080	29	0.080	29	0.080	29	0.080	29	0.080	29	0.080	29	0.080	29	0.080	29	0.080	29	0.080
Rekhter & Hossler (2020)	30	0.494	30	0.494	30	0.494	30	0.494	30	0.1045
Sandlin & Pena (2014)	31	0.140	31	0.140	31	0.140	31	0.140	31	0.1404	31	0.1404	31	0.1404	31	0.1404	31	0.1404	31	0.1404	31	0.1404	31	0.1404	31	0.1404	31	0.1404	31	0.1404	31	0.1404	31	0.1404	31	0.0825	31	0.0825	31	0.0825	31	0.082	31	0.082	31	0.082	31	0.082	31	0.082	31	0.082
Veletsianos et al. (2017)	32	0.490	32	0.490	32	0.490	32	0.490	32	0.4902	32	0.4902	32	0.4902	32	0.4902	32	0.4902	32	0.4902	32	0.4902	32	0.4902	32	0.4902	32	0.4902	32	0.4902	32	0.3331
Shields & Peruta (2019)	33	0.126	33	0.126	33	0.126	33	0.126	33	0.1262	33	0.1262	33	0.1262	33	0.1262	33	0.1262	33	0.1262	33	0.1262	33	0.1262	33	0.1262	33	0.1262	33	0.1262	33	0.1262	33	0.1262	33	0.1262	33	0.1262	33	0.1262	33	0.126	33	0.126	33	0.101	33	0.084	33	0.065	33	0.065	33	0.065	33	0.065	33	0.065	33	0.065	33	0.065	33	0.0482	33	0.0474	33	0.0421	33	0.0421	33	0.0421	33	0.0421	33	0.0421	33	0.0421
Zhu (2019)	34	0.138
Peruta & Shields (2017)	35	0.541	35	0.541	35	0.541	35	0.541	35	0.5410	35	0.5410	35	0.5410	35	0.5410	35	0.5410	35	0.5410	35	0.5410	35	0.5410	35	0.5410	35	0.5410	35	0.5410	35	0.5410	35	0.5410	35	0.5410	35	0.5410	35	0.5410	35	0.541	35	0.541
Galan et al. (2015)	36	0.201	36	0.201	36	0.201	36	0.201	36	0.2009	36	0.2009	36	0.2009	36	0.2009	36	0.2009	36	0.2009	36	0.1361
Hass et al. (2019)	37	0.136	37	0.136	37	0.136	37	0.136	37	0.1361	37	0.1361	37	0.1361	37	0.1361	37	0.1361	37	0.1361	37	0.1361	37	0.0788
Teng et al. (2015)	38	0.538	38	0.538	38	0.538	38	0.538	38	0.5383	38	0.5383	38	0.5383	38	0.5383	38	0.5383	38	0.5383	38	0.5383	38	0.5383	38	0.5383	38	0.5383	38	0.5383	38	0.5383	38	0.5383	38	0.5383	38	0.5383	38	0.5383	38	0.538	38	0.538	38	0.538	38	0.538	38	0.538	38	0.538	38	0.538	38	0.538	38	0.086
Lee et al. (2020)	39	0.112	39	0.112	39	0.112	39	0.112	39	0.1119	39	0.1119	39	0.1119	39	0.1119	39	0.0849
Pringle & Fritz (2019)	40	0.127	40	0.127	40	0.127	40	0.127	40	0.1269	40	0.1269	40	0.1269	40	0.1269	40	0.1269	40	0.1269	40	0.1269	40	0.1269	40	0.1269	40	0.1269	40	0.1269	40	0.1269	40	0.1269	40	0.0471	40	0.0471	40	0.0471	40	0.047	40	0.047	40	0.047	40	0.047	40	0.046	40	0.046	40	0.046	40	0.046	40	0.046	40	0.046	40	0.046	40	0.0456	40	0.0456	40	0.0315	40	0.0315	40	0.0315	40	0.0315	40	0.0315	40	0.0315	40	0.0166	40	0.0166	40	0.0097
Goncalves et al. (2022)	41	0.082	41	0.082	41	0.082	41	0.082	41	0.0817	41	0.0817	41	0.0817	41	0.0817	41	0.0817	41	0.0817	41	0.0817	41	0.0817	41	0.0817	41	0.0817	41	0.0817	41	0.0817	41	0.0817	41	0.0555	41	0.0555	41	0.0555	41	0.055	41	0.055	41	0.055	41	0.055	41	0.055	41	0.055	41	0.055	41	0.055	41	0.055	41	0.055	41	0.055	41	0.0555	41	0.0555	41	0.0555	41	0.0555	41	0.0555	41	0.0388	41	0.0388	41	0.0388	41	0.0388	41	0.0166
Beech (2021)	42	0.087	42	0.087	42	0.087	42	0.087	42	0.0866	42	0.0866	42	0.0866	42	0.0866	42	0.0866	42	0.0866	42	0.0866	42	0.0866	42	0.0866	42	0.0866	42	0.0866	42	0.0866	42	0.0866	42	0.0749	42	0.0749	42	0.0749	42	0.075	42	0.075	42	0.075	42	0.075	42	0.075	42	0.075	42	0.075	42	0.074	42	0.074	42	0.074	42	0.074	42	0.0736	42	0.0736
Lund (2019)	43	0.147	43	0.147	43	0.147	43	0.147	43	0.1471	43	0.1471	43	0.1471	43	0.1471	43	0.1471	43	0.1471	43	0.1471	43	0.1471	43	0.1471	43	0.1471	43	0.1471	43	0.1471	43	0.1471	43	0.1471	43	0.1471	43	0.1471

Figure 20: Final output of the sequence codes based on the lateral comparison of maximum cos θ values. The consecutive index of documents and cos θ values are pairwise color coordinated. Cos θ values have been truncated to four decimal point.

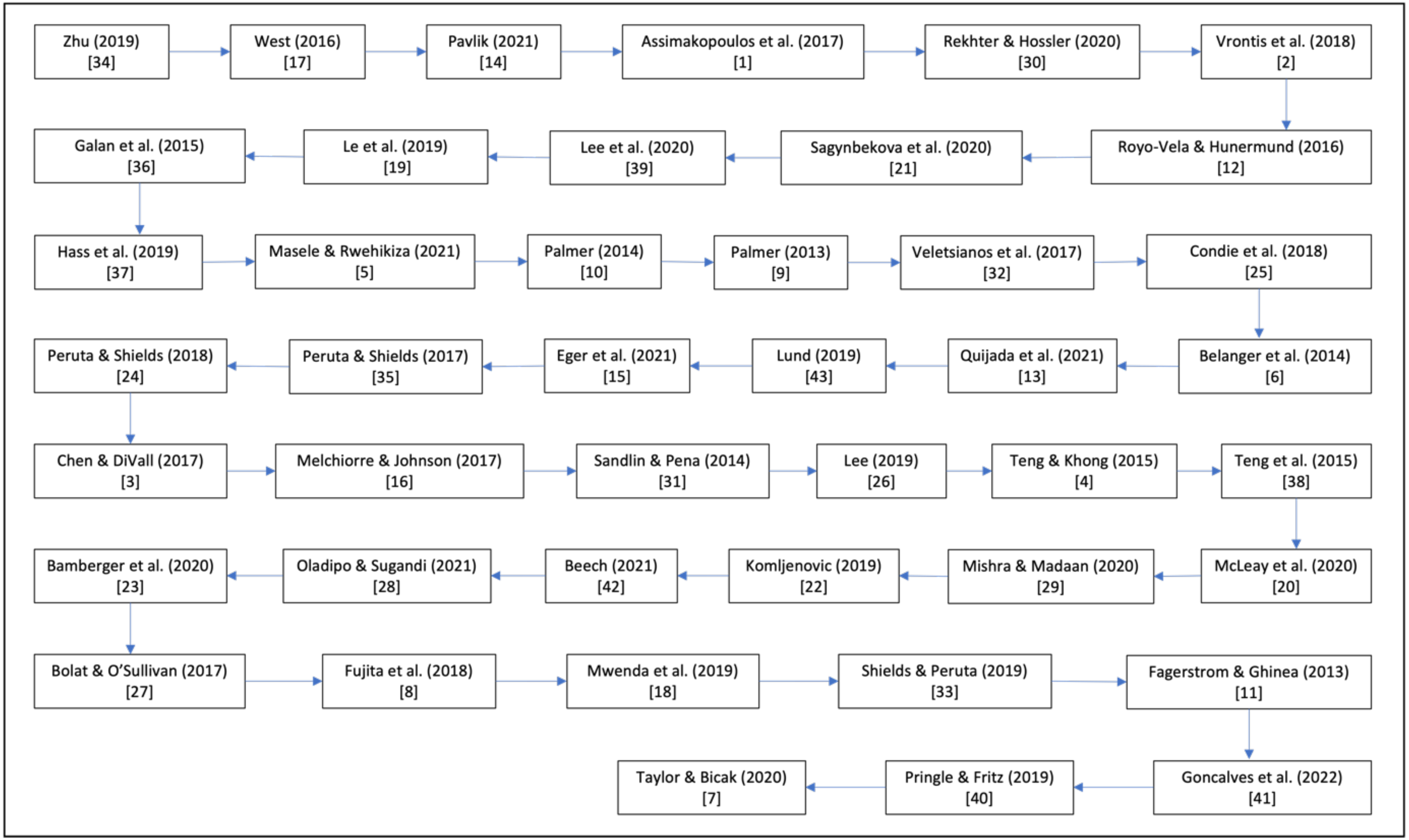

Figure 21: Algorithmic document sequencing (ADS) of all 43 documents with index numbers provided below each document. The first document is Zhu (2019) [Index: 34] and the last document is Taylor & Bicak (2020) [Index: 7].

Once documents are linked, the synthesis process starts in accordance with the sequence provided by the ADS. We synthesized the documents by using the summary of their key findings. We narrated the summary of the key findings of each study independently from the sequence and qualitatively assessed the performance of ADS and the insights provided by this method upon the domain of social media marketing for student recruitment.

5.4 Findings of ADS

As shown in Table 7, summary of the key findings of all 43 documents are linked consecutively to one another in accordance with the ADS. In terms of a cohesive narrative flow, many documents performed well till the 36th document. Based on the homogeneity of the semantic content pertaining to each different topic of inquiry, we established seven document groups specifying seven key findings. Documents sequenced 1-6 were related to "Social Media Engagement; 7-13 to "Brand development and eWOM"; 14-18 to "Engagement via Twitter"; 19-23 to "Post content and Facebook engagement"; 24-25 to "Social media marketing strategy"; and 26-35 to "Matriculation decision factors for students". Moreover, although lacking coherent flow, the insights drawn from documents sequenced 36-40 are gathered under the 7th finding "Miscellaneous insights", and the insights gained from conducting the systematic literature review via algorithmic document sequencing method are discussed under the 8th finding "ADS Methodology".

Finding #1: Social media engagement

The main focal and common topic of interest for the first six documents in the ADS was the utility of social media engagement for recruiting prospective students. Though not a perfect one, we noticed a relatively cohesive flow in the narrative sequence of the key findings of these studies. Zhu (2019) found a positive association between higher education institutions' (HEIs/universities) social media engagement and the number of students studying at these institutions. West (2016) revealed that prospective students rely on social media networks for pertinent and tailored content that engages them with the HEIs. Pavlik (2021) suggested that HEIs need to make it easy for prospective transfer students to access information about the institution and its value offer through marketing materials, emails, website, social media, and word of mouth. Assimakopoulos et al. (2017) confirmed that Facebook can be used by universities as an effective marketing tool to impact potential students' university choice. Rekhter & Hossler (2020) found that Russian students transferring to overseas HEIs indicated that when they had no connections abroad, their university choice was exclusively based on social network sites, whereas social media networks played a complementary secondary role when they had contacts abroad as they mostly relied on their advice. Vrontis et al. (2018) revealed that before enrolling in a higher education program, most prospective students contact family, friends, or a member of the university on social media platforms to get information about the university.

Finding #2: Brand development and eWOM

Articles sequenced 7-13 mostly focused on the importance of eWOM and institutional brand development through creating and communicating social media content. Royo-Vela & Hunermund (2016) found that interactive communication channels can help universities increase brand awareness, improve their image, and attract prospective students to their programs. Sagynbekova et al. (2020) found that for HEIs to be more competitive they can build their brand equity through user- and institution-generated content on social media where eWOM mediates the relationship between social media communication and brand equity. Lee et al. (2020) found that students who highly identify themselves with their university and share positive eWOM about their university tend to have better mental health and wellbeing. Le et al. (2019) established that while selecting a university, students seek information on social media through eWOM mostly regarding universities' reputation, career prospect, learning and teaching, administration, and student life. Galan et al. (2015) confirmed that international students in Australia indicated that whilst evaluating options, they used social media (e.g., Facebook, YouTube, and blogs) to learn more about student life at universities and read reviews from former students. Hass et al. (2019) suggested that to attract and recruit new students, HEIs need to invest in both social media and traditional advertising. Masele & Rwehikiza (2021) confirmed that HEIs can use social media use to reach wider audiences, increase brand awareness, receive, analyse, and respond to customer feedback and integrate social media promotions to traditional marketing mix.

Finding #3: Engagement via Twitter

Articles sequenced 14-18 predominantly centred on improving communication with prospective students through Twitter engagement. Palmer (2014) discussed that to improve the effectiveness of social media activity, whilst writing up Tweet content, it is essential for HEIs to aim for a more engaging Twitter presence that interacts with current and potential Followers. Palmer (2013) also found that HEIs may improve their social media communication effectiveness and achieve institutional objectives by attaining a substantial mass of Followers on Twitter through Retweets. Veletsianos et al. (2017) found that although commonly referred as an interactive platform in literature, Twitter is mostly used by HEIs to make announcements and create enticing depictions of student life on campus. Moreover, Condie et al. (2018) suggested that through 'rotation-curation' when current students share their experiences on Twitter, this allows potential students to gain peer insights into what it is like to be a student of that HEI. Belanger et al. (2014) found that although Twitter is more preferred to carry out conversations, Facebook remains the favourite platform for university-generated content; most of such posts, whether on Facebook or Twitter, broadcast information about events and news.

Seq	Group	Key findings	Index	Author/s
1	1	There is a positive association between higher education institutions' (HEIs/universities) social media engagement and the number of students studying at these institutions.	34	Zhu (2019)
2	1	Prospective students rely on social media networks for pertinent and tailored content that engages them with the HEIs.	17	West (2016)
3	1	HEIs need to make it easy for prospective transfer students to access information about the institution and its value offer through marketing materials, emails, website, social media, and word of mouth.	14	Pavlik (2021)
4	1	Facebook can be used by universities as an effective marketing tool to impact potential students' university choice.	1	Assimakopoulos et al. (2017)
5	1	Russian students transferring to overseas HEIs indicated that when they had no connections abroad, their university choice was exclusively based on social network sites, whereas social media networks played a complementary secondary role when they had contacts abroad as they mostly relied on their advice.	30	Rekhter & Hossler (2020)
6	1	Before enrolling in a higher education program, most prospective students contact family, friends, or a member of the university on social media platforms to get information about the university.	2	Vrontis et al. (2018)
7	2	Interactive communication channels can help universities increase brand awareness, improve their image, and attract prospective students to their programs.	12	Royo-Vela & Hunermund (2016)
8	2	To be more competitive HEIs can build their brand equity through user- and institution-generated content on social media where eWOM mediates the relationship between social media communication and brand equity.	21	Sagynbekova et al. (2020)
9	2	Students who highly identify themselves with their university and share positive eWOM about their university tend to have better mental health and wellbeing.	39	Lee et al. (2020)
10	2	While selecting a university, students seek information on social media through eWOM mostly regarding universities' reputation, career prospect, learning and teaching, administration, and student life.	19	Le et al. (2019)
11	2	International students in Australia indicated that whilst evaluating options, they used social media (e.g., Facebook, YouTube, and blogs) to learn more about student life at universities and read reviews from former students.	36	Galan et al. (2015)
12	2	To attract and recruit new students, HEIs need to invest in both social media and traditional advertising.	37	Hass et al. (2019)
13	2	HEIs can use social media use to reach wider audiences, increase brand awareness, receive, analyse, and respond to customer feedback and integrate social media promotions to traditional marketing mix.	5	Masele & Rwehikiza (2021)
14	3	To improve the effectiveness of social media activity, whilst writing up Tweet content, it is essential to aim for a more engaging Twitter presence that interacts with current and potential Followers.	10	Palmer (2014)
15	3	HEIs may improve their social media communication effectiveness and achieve institutional objectives by attaining a substantial mass of Followers on Twitter through Retweets.	9	Palmer (2013)

16	3	Although commonly referred as an interactive platform in literature, Twitter is mostly used by HEIs to make announcements and create enticing depictions of student life on campus.	32	Veletsianos et al. (2017)
17	3	Through 'rotation-curation' when current students share their experiences on Twitter, this allows potential students to gain peer insights into what it is like to be a student of that HEI.	25	Condie et al. (2018)
18	3	Twitter is more preferred to carry out conversations, however Facebook remains the favourite platform for university-generated content; most of such posts, whether on Facebook or Twitter, broadcast information about events and news.	6	Belanger et al. (2014)
19	4	The level of student engagement on social media platforms hinges mainly on the written content, visual images, format, and the strategies implemented to communicate with prospective students.	13	Quijada et al. (2021)
20	4	Qualitative aspects rather than the frequency of university-generated posts play an important role in engaging students on social media.	43	Lund (2019)
21	4	Different features of content posted by HEIs contribute to different student engagement behaviour on Facebook.	15	Eger et al. (2021)
22	4	Type of content posted by universities and the frequency of postings determine the level of engagement between students and universities on Facebook.	35	Peruta & Shields (2017)
23	4	Type of university-generated content, such as athletics, as well as user-generated content contribute to higher levels of engagement between students and universities on Facebook.	24	Peruta & Shields (2018)
24	5	As part of an overall marketing strategy, applying best 137odelling for social media use can help HEIs engage current and potential students and alumni, hence contribute to enhanced student retention and recruitment.	3	Chen & DiVall (2017)
25	5	Connecting with mature learners and understanding how they respond to online marketing endeavours through a robust social marketing plan can help Continuing Education centres gain a competitive advantage and attain their goals.	16	Melchiorre & Johnson (2017)
26	6	Even when the topics of social media blogs are contrived to centre on admissions and university-related events, prospective students perceive the personal feelings of current students about campus life to be authentic.	31	Sandlin & Pena (2014)
27	6	International branch campuses in China use social media as an avenue to build cosmopolitanism as a desired inclination and therefore depict China as a desirable destination for international students.	26	Lee (2019)
28	6	In their decision-making process on studying abroad, Chinese students mostly acquire information about universities' recognition, job prospects and price from family, the Internet, and agents.	4	Teng & Khong (2015)
29	6	Peripheral communication cues on social media are vital to influence the decisions of Chinese students looking to study abroad.	38	Teng et al. (2015)
30	6	Reasons for international students to study in the UK are public safety, quality of education, admission difficulties, living atmosphere, advice and suggestions, prior knowledge of the UK, work and immigration, and meeting diverse cultures.	20	McLeay et al. (2020)
31	6	Digital marketing efficacy for recruiting students depends on factors such as videos of current students on the university webpage, alumni reviews, blogs, hashtags, virtual tour, and mobile marketing through WhatsApp and Pinterest.	29	Mishra & Madaan (2020)

32	6	LinkedIn offers new use of values for universities such as new scope, scales, and heights that social media platforms provide; with an increasing potential to utilize big data to understand, manage and expand into existing and new marketplaces.	22	Komljenovic (2019)
33	6	It is vital for HEIs to be friendly and supportive in creating notions of caring when they build relationships with potential students and agents for recruiting new students.	42	Beech (2021)
34	6	The primary student recruitment strategies at a Chinese HEI were found to be offer of scholarships, English-taught programmes, communication through digital channels, collaboration with other universities and student advocacy.	28	Oladipo & Sugandi (2021)
35	6	Among two Israeli universities, despite having lower ranking the one which focused on communicating Jewish identity through rich visual and written social media content showcasing the profiles and personal stories of international students rather than alumni and staff performed better than the other in terms of attracting international students.	23	Bamberger et al. (2020)
36	7	Facebook pages created by students sharing content and dialogues can be analysed by universities to monitor, navigate, and impact student impressions.	27	Bolat & O'Sullivan (2017)
37	7	Social media provides an opportunity for HEIs to establish identity links through which students can develop a sense of belonging and feel connected and supported.	8	Fujita et al. (2018)
38	7	Although lacking a clear unique selling offer, YouTube videos promoting Australian universities mostly used students and alumni as speakers focusing mainly on their course experience and job prospects.	18	Mwenda et al. (2019)
39	7	Although students reported in surveys that social media did not impact their enrolment decisions, in interviews most of the students reported that they had used social media to gather more information about the HEIs under consideration.	33	Shields & Peruta (2019)
40	7	Inviting prospective students to actively express their own opinions and initiatives on social media platforms such as Facebook can create significant value between these students and universities.	11	Fagerstrom & Ghinea (2013)
41	NA	Deploying automated chatbots 24/7 helped Brazilian universities meet overflowing demands for student services at a low cost.	41	Goncalves et al. (2022)
42	NA	Analytic tools need to be used to examine brand authenticity as burdened with shades of grey, threatening, or deceitful content on social media may sometimes be used against a HEI.	40	Pringle & Fritz (2019)
43	NA	Private HEIs in the United States purchase more AdWords and invest more in pay-per-click advertising than public HEIs, however this does not generate more visitor traffic.	7	Taylor & Bicak (2020)

Table 7: Summary of the key findings of 43 documents linked consecutively to one another in accordance with the Algorithmic Document Sequencing.
Note. 7 document groups (shaded/unshaded) establish 7 key findings

Finding #4: Post content and Facebook engagement

Articles sequenced 19-23 essentially highlighted the importance of the content of social media posts for student engagement, some focusing on Facebook. Quijada et al. (2021) found that the level of student engagement on social media platforms hinges mainly on the written content, visual images, format, and the strategies implemented to communicate with prospective students. Likewise, Lund (2019) stated that the qualitative aspects rather than the frequency of university-generated posts play an important role in engaging students on social media. Eger et al. (2021) found that different features of content posted by HEIs contribute to different student engagement behaviour on Facebook. Peruta & Shields (2017) confirmed that the type of content posted by universities and the frequency of postings determine the level of engagement between students and universities on Facebook. Peruta & Shields (2018) also found that the type of university-generated content, such as athletics, as well as user-generated content contribute to higher levels of engagement between students and universities on Facebook.

Finding #5: Social media marketing strategy

Articles sequenced 24-25 highlighted the importance of developing a robust social media marketing plan for attracting and recruiting students as Chen & DiVall (2017) suggested that as part of an overall marketing strategy, applying best 139odelling for social media use can help HEIs engage current and potential students and alumni, hence contribute to enhanced student retention and recruitment. Furthermore, Melchiorre & Johnson (2017) found that connecting with mature learners and understanding how they respond to online marketing endeavours through a robust social marketing plan can help Continuing Education centres gain a competitive advantage and attain their goals.

Finding #6: Matriculation decision factors for students

Articles sequenced 26-35 elaborated generally on the factors that may influence mostly international students' decision through social media on choosing a university abroad. Sandlin & Pena (2014) found that even when the topics of social media blogs are contrived to centre on admissions and university-related events, prospective students perceive the personal feelings of current students about campus life to be authentic. Lee (2019) revealed that international branch campuses in China use social media as an avenue to build cosmopolitanism as a desired inclination and therefore depict China as a desirable destination for international students. Teng & Khong (2015) found that in their decision-making process on studying abroad, Chinese students mostly acquire information about universities' recognition, job prospects and price from family, the Internet, and agents. Teng et al. (2015) also suggested that peripheral communication cues on social media are vital to influence the decisions of Chinese students looking to study abroad. McLeay et al. (2020) found that reasons for international students to study in the UK are public safety, quality of education, admission difficulties, living atmosphere, advice and suggestions, prior knowledge of the UK, work and immigration, and meeting diverse cultures. Mishra & Madaan (2020) revealed that digital marketing efficacy for recruiting students depends on factors such as videos of current students on the university webpage, alumni reviews, blogs, hashtags, virtual tour, and mobile marketing through WhatsApp and Pinterest. Furthermore, Komljenovic (2019) stated that LinkedIn offers new use of values for universities such as new scope, scales, and heights that social media platforms provide; with an increasing potential to utilize big data to understand, manage and expand into existing and new marketplaces. Beech (2021) discussed how important it is for HEIs to be friendly and supportive in creating notions of caring when they build relationships with potential students and agents for recruiting new students. Furthermore, Oladipo & Sugandi

(2021) found that the primary student recruitment strategies at a Chinese HEI were offer of scholarships, English-taught programmes, communication through digital channels, collaboration with other universities and student advocacy. Finally, Bamberger et al. (2020) found that amongst two Israeli universities, despite having lower ranking the one which focused on communicating Jewish identity through rich visual and written social media content showcasing the profiles and personal stories of international students rather than alumni and staff performed better than the other in terms of attracting international students.

Finding #7: Miscellaneous insights

Articles sequenced 36-40 provide miscellaneous insights about social media engagement and student's matriculation decision making. Bolat & O'Sullivan (2017) stated that Facebook pages created by students sharing content and dialogues can be analysed by universities to monitor, navigate, and impact student impressions. Fujita et al. (2018) established that social media provides an opportunity for HEIs to establish identity links through which students can develop a sense of belonging and feel connected and supported. Mwenda et al. (2019) found that even though lacking a clear unique selling offer, YouTube videos promoting Australian universities mostly used students and alumni as speakers focusing mainly on their course experience and job prospects. Shields & Peruta (2019) found that although students reported in surveys that social media did not impact their enrolment decisions, in interviews most of the students reported that they had used social media to gather more information about the HEIs under consideration. Fagerstrom & Ghinea (2013) stated that inviting prospective students to actively express their own opinions and initiatives on social media platforms such as Facebook can create significant value between these students and universities.

Although providing some insights about HEIs and the way they communicate with students, the last three articles [41-43] were not logically linked to the preceding ones as

Goncalves et al. (2022) found that deploying automated chatbots 24/7 helped Brazilian universities meet overflowing demands for student services at a low cost. Whereas Pringle & Fritz (2019) suggested that analytic tools need to be used to examine brand authenticity as burdened with shades of grey, threatening, or deceitful content on social media may sometimes be used against a HEI. Finally, Taylor & Bicak (2020) found that private HEIs in the United States purchase more AdWords and invest more in pay-per-click advertising than public HEIs, however this does not generate more visitor traffic.

5.5 Limitations of ADS

One limitation of the ADS, due to time restrictions, was that I did not consider the temporal order of the documents. As a result, the documents that produced high similarity and were authored by the same researchers may have been sequenced in reverse order. For example, the 2013 paper authored by Palmer was placed after his 2014 paper. Hence, I recommend future researchers using ADS to integrate the temporality feature at least for the same author and/or highly similar papers to the sequencing algorithm.

Another limitation is that if a subtopic in Paper X is relevant to something in Paper Y, the methodology may or may not permit incorporating this connection. This is due to ADS being based on the overall pair-wise similarity of documents. As a result, sub-topical connections may be disregarded.

5.6 Combined Findings and Discussion

To validate the data collected through different methods, make comparisons among their similar and dissimilar results and gain a fuller picture by capturing different aspects and complexities of the themes under investigation, we adopted an approach called “following a thread” developed by Moran-Ellis (2006). Based on our original inquiry, after each method has generated a matriculation decision theme as output, each emergent theme is followed across the output of others (the thread). Therefore, whilst intermeshing the themes produced by each method, we retain the distinct paradigm of each method. We should note that although the topic modelling techniques (LDA and STM) and the systematic literature reviewing technique (ADS) are all quantitative methods in nature, their output in terms of matriculation decision themes can only be manually intermeshed, compared, and validated qualitatively.

Topic #	Doc Sequence Indices	Matriculation Decision Theme
1	28, 30	University's recognition, reputation and education quality
2	28	Job prospects
3	34	Availability of scholarships for students
4	33	Caring and support provided at the university
5	27, 30	Vibrant, cosmopolitan living environment & meeting diverse cultures
6	28	Cost of education and living
7	30	Public safety
8	30	Ease of admission
9	30	Prior knowledge of the study destination
10	30	Work and immigration opportunities in the country
11	34	Collaboration with other universities
12	26, 31, 34, 35	Profiles and personal stories of other students
13	26, 28, 30	Advice and suggestions from family, friends and others

Table 8: Students' matriculation decision themes derived from the 10 documents ranged from 26th to 35th in the Algorithmic Document Sequence.

After identifying the first document [(Zhu, 2019]) and calculating all cos θ values among all 43 documents, we linked these documents consecutively to each another along the sequence generated by the ADS and narrated the summary of the key findings of each document independently from the sequence (Table 7). Based on the homogeneity of their semantic topical content, seven clusters of key findings were established. One of the seven clusters was "Finding #6 (Group #6): Matriculation decision factors for students" which comprised the summaries of the key findings of the documents ranged from 26th to 35th in the sequence. Within this cluster, we identified 13 matriculation decision themes out of the key findings of these 10 documents (Table 8).

When we combine the results of all three quantitative methods (i.e., LDA and STM from social media data and ADS from literature), we notice that out of a total of 9 matriculation decision themes, two of them are confirmed by all three, and five of them by two of the methods

(Table 9). The remaining two factors are only represented by ADS. It's worth noting that the matriculation decision themes shown in Table 9 are the decision factors that may potentially impact students' university choice. However, regardless of the likelihood of such impact for many factors, without a proper experiment at this stage or a thorough investigation into the research design of the source of these factors in literature, we shall not claim that any one of these 9 factors affects such choice. We should also note that through methodological pluralism we do not seek to justify making any causal links but merely aim to enrich and diversify the information we retrieve via multiple methods from multiple sources.

Matriculation Decision Themes // Method:	LDA	STM	ADS
Reputation, image and global ranking [LDA] – International reputation, image and prestige of the school, its professors, their research, service provided to the community (in terms of health and education) and the quality of the students produced [STM] – University's recognition, reputation and education quality [ADS]	I	I	I
Living and study costs, availability of scholarships and access to technology, research and facilities (buildings, libraries, science labs, etc.) [LDA] – Cost of education and living [ADS] – Learning and research facilities, and use of technology (i.e., online, and social media channels) to communicate with potential students [STM]	I	I	I
Work and internship placements during study, job opportunities and potential work-related benefits after graduation [LDA] – Work and immigration opportunities in the country [ADS] & Job prospects [ADS]	I		I
Word-of-mouth in form of guidance, advice, suggestions, and influence (by family, friends and communities including current students and graduates) [LDA] – Advice and suggestions from family, friends and others [ADS] & Profiles and personal stories of other students [ADS]	I		I
Ease of admission, entrance requirements and open communication with admissions staff [LDA] – Ease of admission [ADS]	I		I
Campus location (proximity to home, convenience and comfort), its safety and physical appeal, and vibe of the city [STM] – Public Safety [ADS] & Vibrant, cosmopolitan living environment & meeting diverse cultures [ADS]		I	I
Availability, flexibility and attractiveness of the course/program of study (in line with career aspirations and earning potential) and on-campus support services [STM] – Caring and support provided at the university [ADS]		I	I
Prior knowledge of the study destination			I
Collaboration with other universities			I

Table 9: Students' matriculation decision themes by Method (i.e., Latent Dirichlet Allocation (LDA), Structural Topic Modelling (STM), and Algorithmic Document Sequencing (ADS))
Note. 'I' stands for Identified.

The process of intermeshing the matriculation decision themes by method is not an impeccable let alone a straightforward one. Although "following a thread" approach provides some guidance and structure, it can still be a subjective decision-making process for researchers to crossmatch distinct themes along a thread. In our case, we encountered two challenging instances. The first one was to decide whether an LDA theme namely "Work and internship placements during study" be cross matched with an ADS theme namely "Job prospects". It can be argued that the gist of the themes provided by different methods may be perceived as having both similarities and differences. For example, the ADS theme "Job prospects" is a topic about job opportunities for students. Since we know that work experience is strongly linked to job opportunities and there are more job opportunities for students who "gain work experience while studying their degree" (LDA theme), at our discretion we matched those two themes with one another. However, when we combine them for triangulation or cross validation, we shall not let either one of the themes get absorbed by the other. Therefore, their combination should retain the features of each individual theme as: Work and internship placements during study, job opportunities and potential work-related benefits after graduation [LDA] and Job prospects [ADS]. Albeit similar in theory, the second instance was more challenging than the first. It was between an STM theme namely "Learning and research facilities, and use of technology (i.e., online, and social media channels) to communicate with potential students" and an ADS theme called "Cost of education and living". Yet again at our discretion, we opted to crossmatch them as a result of the output of the third method, LDA, which incorporates the themes of both STM and ADS within its theme: "Living and study costs, availability of scholarships and access to technology, research and facilities (buildings, libraries, science labs, etc.)". Furthermore, it can also be argued that students choosing to study at zero- or low-cost public universities or getting scholarships from private universities may prefer HEIs that provide their students with better technology and facilities.

The key finding of this chapter's study is the fundamental conformity to methodological pluralism as different methods have generated diverse and valuable insights into students' matriculation decision factors and no method can be claimed superior to others, whereas all methods together provide richer and more valid information than each single method alone. This is in line with Gestalt theory which highlights that the "whole" of a phenomenon is greater than its "individual parts". As a result, the attributes of the whole can help us develop a more profound understanding and a fuller picture of the phenomenon at hand (i.e., decision factors of students).

An interesting finding of this chapter's study is the high percentage of discrepancies in the output of the topic models. LDA and STM models represented 5 and 4 themes respectively. Although both models were based on the same source (i.e., text data) retrieved from Facebook and Twitter, in the end they have produced only two common themes: [1] LDA: Reputation, image and global ranking and STM: International reputation, image and prestige of the school, its professors, their research, service provided to the community (in terms of health and education) and the quality of the students produced; and [2] LDA: Living and study costs, availability of scholarships and access to technology, research and facilities (buildings, libraries, science labs, etc.) and STM: Learning and research facilities, and use of technology (i.e., online, and social media channels) to communicate with potential students. In other words, out of a total of 9 themes produced by LDA and STM, 5 of them were distinctly represented by either LDA or STM. Furthermore, we notice that all 5 of these themes were validated by ADS. The comparison between LDA and STM output validated by ADS leads us to five distinct themes which enhance and broaden our understanding of the potential matriculation decision factors, as well as two major themes (international reputation and costs & facilities) which potentially deserve more attention as they are the only themes all three methods have in common.

Another key finding is that "international reputation" of the university and its staff was the main matriculation decision theme confirmed by all three methods. It should be noted that although each method seems to have produced slightly different output (LDA: Reputation, image and global ranking; STM: International reputation, image and prestige of the school, its professors, their research, service provided to the community (in terms of health and education) and the quality of the students produced; ADS: University's recognition, reputation and education quality), in essence they mainly represent a single theme: international reputation. There may be numerous factors that can contribute to a HEI's international reputation. Our results indicate that some of these factors are associated with the recognition of the university and the impact of the research undertaken by its academic staff. However, it is difficult for prospective students to quantitatively estimate these factors. To fill this need, national and international ranking organizations (e.g., Times Higher Education, QS, ARWU) use these indicators to rank HEIs for students to be able to make comparisons. Since global ranking of a HEI is commonly attributed to the reputation of that institution (Hazelkorn, 2014; Wolf & Jenkins, 2018), international reputation may potentially be the main reason why more students prefer to study at HEIs with higher global ranking.

Many critics of methodological pluralism including Zelditch Jr. (1962) and the ones that followed him believed that a particular issue must be addressed by using only the best method since the ones inferior to the best method would produce less accurate results than the best method. This may be true for physical sciences where research results can objectively be estimated, and the inferior methods (to the best one) do not produce any additional valuable information that have not already been covered by the best method. However, in social sciences that is almost never the case because human subjectivity usually does not allow for an objective estimation of the results and in addition to 'arguably' the best method, other methods may often generate valuable complementary or supplementary insights into a social phenomenon. There is a myriad of examples in mixed-methods social research. In the domain of information retrieval and knowledge discovery, these studies mostly use qualitative techniques such as interviews at first to induce theories and then quantitative techniques to test these theories. Studies using quantitative and qualitative techniques concurrently to collect as much information into their research question as possible from different sources are also not uncommon.

Researchers have also argued that one major challenge of multi-method assessment is the analysis phase of data collected from multiple sources and processed through multiple methods (Greene, 2008; Odendaal et al., 2016). This phase is an iterative process with initial findings and methodological understandings from each method linking to the insights of other methods. To develop an integrated set of outcomes, we used "following a thread" approach which helped us explore the output from each method while making connections among the emerging themes. However, it should be noted that to minimize researcher bias, particularly in the form of interviewer influence, the results of each method must be produced independently from one another without any intention to make any connections among them.

5.7 Conclusion of chapter 5

Adopting a methodological pluralist approach, we collected data from social media, quantitatively analyzed them with two topic modelling methods, collected data from literature and quantitatively analyzed them with a systematic literature reviewing method. We gained a much deeper and richer understanding of students' matriculation decision factors by intermeshing and cross validating the results of all three methods than the insights we could draw from each method alone.

The perceptual values we perceive in the whole from all three methods, with reference to Gestalt theory, are different from those we observe in individual parts (i.e., methods). Whether it be a topic modelling or a systematic reviewing technique, each one of the three methods we employed provided us with valuable insights drawn from similar as well as dissimilar information pertaining to students' matriculation decision factors. Where analogous information tend to strengthen the explanatory power of some of the potential decision factors, diverse information help us uncover a wider range and gain a richer understanding of such factors. After all, we stand by the famous quote attributed to Aristotle, as it perfectly applies to our case of using multiple methods and sources in a holistic way for modern-day information retrieval: "the whole is greater than the parts."

5.8 Limitations of chapter 5 study and recommendations

The limitations inherent in each method as discussed by Blei et al. (2003) for LDA, by Roberts et al. (2019) for STM, and by Cingillioglu et al. (2023) for ADS also apply to the combined results of this study. More importantly, it should be noted that since the limitations of social media as data source as well as LDA and STM as topic modelling techniques are higher than those of the extant literature as data source and ADS as a systematic literature reviewing technique, the insights provided into students' matriculation decision factors by ADS from the literature were superior to those provided by LDA and STM from social media. We do not believe that other topic modelling techniques would provide superior results with data collected from social media. Therefore, for future studies we recommend that the methods be switched between data sources. In our case, for example, ADS could be used upon the social media data, whereas LDA and STM upon the data from literature. We also recommend future research to focus on developing and evaluating multi-method approaches that effectively integrate different types of data and are interpretable.

Part 3

Collecting data via chatbots and AI-led experiment

Chapter 6

1. **A double blind experiment conducted with an AI-led chatbot**

6.1 Conceptual background

Interviewing is a common data collection method utilized mostly in qualitative studies where researchers record transcriptions as data and then analyse them to generate theory (Schultze & Avital, 2011) and gain insights into the question of "why" people behave, think or act in a particular way (Black, 1994; Rosenthal, 2016). Open-ended questions are usually asked in interviews and initial responses are probed with follow-up questions to collect thick and rich descriptions of respondents' opinions, lived experiences, and behaviour about a phenomenon that is needed to be explored in greater depth than quantitative methods (Leeson et al., 2019). Skilled qualitative researchers develop interview questions in a way to minimize misunderstandings and pose prompts or follow-up questions to further their understanding on a subject matter (Turner, 2010). Researchers aim to keep respondents focused on the theme of inquiry to acquire relevant and in-depth information from them regarding their subjective and personal feelings, opinions, behaviour, or experience about a phenomenon (Creswell, 2007). To achieve this, establishing good rapport with respondents is critical as it helps them feel at ease and open up during interviews (Bell et al., 2020). If conducted skilfully, interviews can be effective in investigating not only general social phenomena but also sensitive topics such as traumatic experiences (Bögner et al., 2010).

Another advantage of qualitative interviews is that even if participants provide off-topic responses, interviewers can manage the situation on the spot and guide participants to get back on track. This allows researchers to collect thick data by delving even deeper down into the subject matter as much as they require, hence reveal more important and interesting perspectives from which detailed valuable insights and conclusions can be drawn (Hamilton, 2020). To achieve rigorous qualitative results, it was suggested that interviewers need to have developed exceptional communication skills that incorporate high empathetic and analytical competence. Moreover, qualitative researchers can get confirmation from respondents during interviews regarding the correctness of their interpretation. Such confirmations are particularly useful to improve the validity of collected data (Morse, 2015).

Another advantage is that since the time, location and language of qualitative interviews can be arranged in accordance with interviewees' needs and wants, they will not only feel more comfortable and confident answering the questions, but also develop a sense of ownership and commitment to the study (Braun et al., 2020). Finally, the data collected from qualitative interviews can be used to advance existing knowledge and explore new areas of research as a foundation or steppingstone for both subsequent qualitative and quantitative studies (Frels & Onwuegbuzie, 2013).

Interviews conducted as part of qualitative studies have been criticised in literature mostly by quantitative researchers. These criticisms usually centre on questioning the validity, objectivity, and reliability of collected qualitative data (Weis & Willems, 2017). A major limitation of qualitative interviews is that the presence of researcher during data collection may impact interviewees' responses (Romano et al., 2020). Although such presence may have proven to be useful helping respondents stay focussed on the matter of inquiry, it may still raise anonymity and confidentiality issues deterring respondents from revealing the truth (Opdenakker, 2006). Another limitation is that the results may be prone to researcher bias because interpretation of collected data during coding and analysis may be influenced by researchers' personal opinions, experiences, principles, and perspectives (Romano et al., 2020) or the need and desire to conform with existing literature or reinforce a specific theory (Kelle, 2008).

Arguably the most criticised aspect of qualitative interviews is their small sample size being inadequate to represent the general population of interest. As a result of small sample sizes often leading to sampling bias (Amri et al., 2021), low statistical significance does not allow qualitative researchers to test hypotheses and make statistical inferences (Davies & Dodd, 2002). Even though some qualitative researchers attempt to compensate sample size issues with a comprehensive description of their methodology, quantitative community hardly consider these results scientific due to their insufficiency in terms of replicability, transferability, and generalizability (Vasileiou et al., 2018).

Qualitative researchers have often used the concept of data saturation to estimate sample size and validate their methods (Guest et al., 2020). Although data saturation in qualitative interviews is a critical milestone that indicates that no new information adds value to the study anymore, there are no commonly accepted guidelines, tests, or standards to determine the sample size necessary to attain data saturation (Saunders et al., 2018). Accordingly, in addition to non-qualitative methodologists finding the merit of qualitative studies questionable due to their inadequate sample size many qualitative researchers recognize the insufficient attempt by which sample size is delineated and justified in qualitative studies (Marshall et al., 2013).

With an ability to customize user experience and allow fast and efficient data collection, AI-based technologies can generate big data to make significant progress in a plethora of research and non-research areas (Boyd & Crawford, 2012). Chatbots also commonly referred to as virtual agents and conversational assistants are a form of AI-based technology that have increasingly been used in business operations and marketing to enhance customer satisfaction by delivering simple and fast information (Arsenijevic & Jovic, 2019). Chatbots have been used in education to help learners develop their critical thinking and language skills (Goda et al., 2014). There has also been a growing demand to utilize AI-led chatbots in healthcare to provide guidance, education, and prompt behavior change for patients (Nadarzynski et al., 2019). Likewise, in public sector chatbots have been integrated to government websites and social media to disseminate essential information, steer users through online services such as tax return submission inquiries (Australian Taxation Office's chatbot Alex has resolved 80% of customer inquiries without human intervention (CX Central, 2019)), and communicate political and social messages (Androutsopoulou et al., 2019).

Sidaoui, Jaakkola, and Burton (2020) posited that chatbots have the potential to take up the role of an interviewer by shifting from its traditional passive role of being a source of information to a more active role of collecting customized data and asking questions based on respondent input. Therefore, interviews conducted via AI-powered chatbots may emerge as a widely used and efficient approach for gathering qualitative data that are pertinent to exploring subjective social phenomena in depth.

Due to their AI-augmented capabilities, chatbots have evolved into so much more than not just traditional qualitative interviews but also interactive online surveys. As discussed by Sidaoui et al. (2020) and shown in Table 10, chatbot interviews possess the benefits of a combination of the advantages of both online surveys (low cost, scalable, fast deployment, flexible availability, real-time analysis) and traditional interviews (rich data collection, customized, engaging) except for being able to detect body language and ladder questions like a human interviewer. With an increasing recognition of their potential in understanding human perspective, they could engage users and extract their opinions and experiences from narrative conversations via algorithms based on semantic and sentiment analysis (Sarkar, 2016). In addition, chatbot interviews, unlike traditional interviews and online surveys, can engage respondents with conversation tools and materials in multiformat (text, speech, 2D and 3D images) and leverage automation and data mining techniques augmented by AI and machine learning to extract meaning and intention from responses to potentially adapt to the personality of interviewees (Park et al., 2019).

Advantages	**Online surveys**	**Traditional interviews**	**Chatbot interviews**
Low cost	O		O
Broad reach/scalability	O		O
Fast deployment/speed	O		O
Flexible availability	O		O
Real-time analysis	O		O
Rich data collection		O	O
Customized/personal/empathetic		O	O
Engaging/interactive		O	O
Laddering and probing questions		O	A
Body language detection		O	A
Multiformat conversation			O
Automation			O
Adaptable personality			A

Table 10: Comparison of the advantages of chatbot interviews, online surveys, and traditional interviews. Adapted from Sidaoui et al. (2020).
Note1. “A” stands for further development potential via Augmentation
Note2. “O” stands for Observed

Online surveying company Survey Sparrow has been promoting the benefits of using chatbots in collecting and analysing qualitative data from respondents. According to Survey Sparrow, chat surveys generate much higher participations rates than other forms of online surveys. Similarly, a comparative field study revealed that the responses obtained by a conversational chatbot guided survey were clearer, more informative, specific, and relevant than the ones collected by a web survey on Qualtrics (Ziang et al., 2020). Similarly, Kim et al. (2019) concluded that a chatbot survey generated higher-quality data than a web survey and another study that compared user experience between an AI-powered chat survey and a conventional computer survey revealed that the users would rather interact with the chatbot than fill in a computer questionnaire (E te Pas et al., 2020). More research revealed that chatbots offer a higher level of user experience than online surveys do. Respondents find the experience of engaging and conversing with chatbots more fun than simply filling out online questionnaires. Although users know that they are not interacting with a human but a machine, they prefer having such an experience to being alone in front of a form.

Furthermore, advanced chatbots use customized information about respondents during conversation to build rapport and provide personalized guidance allowing respondents feel at ease and develop a sense of ownership and commitment to the study. Customized data can be anything from the name of the respondent to background info, to a number, time, to a specific experience, to a personal choice. When a respondent provides such information at some point, the chatbot records and uses them as needed throughout the conversation. Because respondents see for themselves that they are being listened to and how their responses are valued, they are more inclined to provide more in-depth, accurate and richer information whilst conversing with a chatbot than they do while completing online forms. However, current chatbot technologies are not that advanced enough to recognize verbal responses as accurately as humans do.

6.2 Chatbot Surveys

Surveys are a robust data collection method to draw inferences to populations (Couper, 2017). Through the intermediary of emerging technology, surveys allow researchers to collect big data from massive samples. Although traditional paper-based surveys have a fixed questionnaire making respondents answer the same questions in a fixed order, interactive web surveys have the ability to validate responses, check for unacceptable answers or blank answers (Dominelli, 2003), and customize questions or the order of questions as per the preceding responses (Christian et al., 2009).

Interactive web surveys, however, are not built for narrative data collection like interviews are. Typically, in interviews people are asked structured, semi- or un-structured questions and their verbal answers are recorded as part of a conversation. Due to respondents being an active participant in a mutual verbal conversation containing probing, follow-up or laddering questions, interviews tend to have a higher completion rate and more potential to collect thick data (adding context as to why and how data eventuate) than interactive web surveys.

Albeit powered by AI, chatbots are not equipped to understand human language unless they are specifically trained with datasets that tell them how to interpret and respond to specific words, phrases and sentences that might come up during a conversation with a human respondent. Using natural language processing (NLP) techniques such as topic modelling, aspect mining and sentiment analysis, AI-led chatbots aim to detect and extract relevant information from sentences as every term and groups of terms used in a sentence get constantly compared against their training database. However, it is not uncommon where a response includes terms that have not been covered by the database. In that case, the AI fails to understand the respondent, and hence can neither record the response promptly nor provide an adequate answer to the response or generate a rational follow-up question.

A vital feature of chatbot surveys is that they offer multiple choices to respondents. Due to the tree structure allowing researchers to frame the domain of their interest in accordance with a specific data collection goal, chatbots with a survey design can be more effective in terms of user experience than others that are designed to interpret open-ended/free text responses. Although the information provided by respondents with free text can lead researchers to richer insights than those collected from multiple choices, there is a trade-off.

Because of the inherent complexities and challenges of interpreting free text, in cases where the chatbot fails to understand user response, users might quickly get disappointed and discontinue the conversation. This results in low response and completion rates. Unlike free text interpreting chatbots, chatbot surveys that provide multiple choices do not suffer from such issues because their AI have already been trained with each choice and each chatbot response or question is logically connected to the preceding choice selected by the user. Therefore, survey design allows for a smooth transition from a chatbot question to a human response, and from a human response to a follow-up question.

Another major benefit of chatbots with survey design is that there is limited or no need for processing natural language during data collection and preparation for analysis. Since the AI of chatbot surveys has previously been trained with the terms of each choice, it does not have to apply NLP techniques to recognize and interpret the responses. Whereas the relevancy and accuracy of collected data are subject to the performance of NLP technologies while processing open-ended text, with their tree structure via multiple choices, chatbot surveys collect and record relevant data that are immune to false recognition and misinterpretation.

6.3 Chatbot architecture

The AI-led chatbot (AILC) that I built for this study (i.e., collecting data from and running an experiment on participants) has a nested tree structure comprised of multiple choices and capable of processing open-ended natural language responses, recognizing all plausible responses, reprompting implausible ones and compensating for misunderstandings as it is equipped with a confirmation feedback mechanism (Confirmatory Feedback Loop (CFL)) allowing the AI to guide or redirect the human respondent (RRP: Redirection via Rephrase Prompt) if needed and confirm the allocation of an identifiable and relevant response to its preassigned code. As a result, structured quantitative data and unstructured qualitative data are produced as final output. Structured data are utilized to draw causal inferences between each tested IV (e.g., Campus location (proximity to home, convenience, and comfort), safety and physical appeal, and vibe of the city) and DV (Student preference (i.e., university choice)). AILC is designed to run the experiment unsupervised making double blind and random allocations, conversing with, and collecting information from participants, and storing data in structured and unstructured form to be either analysed for causal inference or passed back down to its internal model for recalibrations applicable to future experiments.

A vital feature of the AILC is its capability to randomly assign anonymous participants to the Control and Test groups in a fully unsupervised way. Although potential participants are aware of the general context of the study (assuming they read the content provided in consent forms properly), they are unaware of to which group (i.e., CTRL or one of the Test Groups) they are allocated. Due to the unsupervised nature of this process, the researchers are also entirely unaware of this allocation. AILC simulates a one-on-one interview by engaging respondents and prompting them follow-up and laddering questions. However, unlike traditional interviews, the form of interaction I propose for this chatbot is textual rather than verbal. I opt not to use a voicebot so as not to sacrifice the voice/speech recognition accuracy of verbal responses during their speech-to-text conversion. After all, the average speech-to-text transcript accuracy among the leading conversion engines developed by the tech giants of our time as of March 2020 was only 76.7% (Amazon, Microsoft and Google could produce accuracy rates of 73%, 78% and 79%, respectively (Liu, 2020)). Since interviews are generally expected to have a verbal nature of information exchange between an interviewer and an interviewee, we shall not address the type of data collection that we propose as an interview, but due to its ability to simulate a one-on-one interview combined with a textual form of interaction, we name it a chatbot-led interview-like survey.

6.4 Experiment methodology

6.4.1 Adaptive Design

I adopted a goal-oriented adaptive experiment design through which the experiment platform is run automatically by the AI and the design of a new experiment is based on the outcomes of its predecessors. Upon running the experiment, the AI produces structured output (i.e., anticipated in accordance with initially identified decision factors) which will be used to draw causal inferences and update the constructs of upcoming experiments. For instance, if a decision factor (i.e., IV) is found to have no causal relationship with the DV, its 'entity' will be removed from the new experiments' design along with its input prompts in the dialogue. As a result, new participants will not be asked or prompted with semi-structured questions about this factor anymore unless a new experiment captures it as unstructured input and puts it back in the internal model. This might apply to not just removed factors but more likely to new factors as the AI records unstructured output (in free-text form) which will subsequently be processed by human researchers via topic modelling and then fed into the internal model to capture further insights about the phenomenon. This cycle is adaptive and iterative in nature in a way that the constructs and parameters of new experiments are conditioned upon the collected, collated, measured, and processed results of former experiments. The adaptive nature of the proposed AI-powered experimental design may also allow for researchers to allocate resources (e.g., sample size) more efficiently in the upcoming experiments based on the statistical measures (e.g., Cohen's d, Power) applied to the preceding experiments.

6.4.2 Cause and effect

Just like in many other fields, causal effect in social science can be established when a change in one variable (e.g., X) leads to a change in another variable (e.g., Y) ceteris paribus (i.e., all other things being equal). The main requirements of a causal relationship between X and Y are threefold: (1) Temporal order (X must come before Y); (2) Strong correlation (observed relationship between X and Y was not due to chance); and (3) Non-spuriousness (there is nothing else – such as a confounder – that may account for the correlation between X and Y). To strengthen causal explanations two additional criteria are required: (4) Recognition of a causal mechanism; and (5) Identification of the context in which the effect takes place. To determine cause-and-effect relationship, true experimental research design is the most preferred method commonly used in natural and physical sciences. However, the number of applications of true experimental research design to social and particularly behavioural sciences is on the rise.

6.4.3 Group formation

We adopt a true experimental research design to establish causation between independent and dependent variables. Randomly selected participants will be randomly allocated to a Control and 8 Experiment (Test) Groups to prove hypotheses. Conditions in all groups will be the same except for a single condition applied to each different Experimental group at a time. The participants will be distributed to one of the 9 groups randomly without knowing the conditions to which they were subject, or to which group they belonged (blind allocation). All participants will be unaware of whether they received an intervention. Furthermore, to maximize the benefits of a true experiment and eliminate any potential confirmation or researcher bias in the form of interviewer effect, thus avoid false positive conclusions, we implement a double-blind experimental design. Since simple randomization allows for complete randomness of the allocation of a participant to a specific group (Suresh, 2011), the random allocation of the participants to the groups was handled by the chatbot using a simple randomization algorithm. As a result, not only the participants, but we (researchers) will not be aware of who is allocated to which group and subject to which intervention.

With this true experimental double-blind design including 1 control and 8 experimental groups, the AI-led chatbot (AILC: Sydn-e) randomly allocates a reasonably large sample of participants to one of the 9 groups. Participants in all groups will receive the same information (Constant) about studying at a university. The text of the Constant was extracted from the webpages of the top five ranked (by Times Higher Education 2022) universities in the world: University of Oxford, California Institute of Technology, Harvard University, Stanford University, and University of Cambridge. We deliberately selected *general phrases* that are commonly used by many other universities around the world and do not identify or distinguish these universities in any way. Furthermore, to achieve commonality and moderation, we refrain from using distinguishing words such as "leading", "top', and "best". The participants are anticipated to construe the statements of the Constant as originating from a single university.

Constant:

"*We offer a range of precious opportunities for personal growth and professional development as well as combine rich history and tradition with the innovative and forward-thinking approach of a modern university. Our students create and apply knowledge by thinking and doing, preparing for leadership in a rapidly changing world. Courses, taught by esteemed faculty members and enhanced by our unparalleled libraries and resources, will take you as far as your imagination allows. Here, you're going to be part of a community—one where everybody works hard, but that also takes a breather every now and then. In fact, the students who do best here already have some kind of outlet, such as 161odelli, athletics or the arts*."

By the end of chapter 5, I identified eWOM as one of the nine matriculation decision themes (electronic Word-of-mouth in form of guidance, advice, suggestions, and influence (by family, friends and communities including current students and graduates) [LDA] – Advice and suggestions from family, friends, and others [ADS] & Profiles and personal stories of other students [ADS]). However, since WOM is not a decision factor but simply a key channel for prospective students to be informed about and consider other decision factors while selecting a HEI, I incorporated eWOM as a means to relay information during the chat about the rest of the identified university choice factors. Therefore, in addition to the Constant, except for the ones in Control group (CTRL), participants in each Experimental group will be exposed to different information (Intervention) presented in the form of positive eWOM on social media highlighting a distinct factor that may influence their choice about studying at a hypothetical university.

Upon a total of 8 Experimental groups (Egs), the impact of 8 independent variables (Ivs) are tested: IV1: Reputation, image and ranking, IV2: Living and study costs, availability of scholarships and access to technology, research and facilities (buildings, libraries, science labs, etc.); IV3: Work and internship placements during study and job prospects upon graduation, IV4: Ease of admission, entrance requirements and open communication with admissions staff; IV5: Campus location (proximity to home, convenience and comfort), safety and physical appeal, and vibe of the city; IV6: Availability, flexibility and attractiveness of the course (in line with career aspirations and earning potential) and on-campus support services; IV7: Prior knowledge of the study destination; and IV8: Collaboration with other universities. Each EG and its corresponding IV were allocated a number (from 1 to 8) and tested against the CTRL Group (Table 11).

I tested the effect of each IV independently on a single dependent variable (DV): The likelihood of the participant to enrol in this university [StPref]. I used a 5-point Likert scale (5: Absolutely; 4: Yes, why not; 3: Not sure; 2: Not really; 1: No way) to measure the decisions of participants in a hypothetical scenario assuming that they are about to make a university choice based on the information they read in the Constant and/or one of the eight Interventions (i.e., Ivs) conveyed in the form of positive eWOM.

Group	RA Code	Module	IV Factor	Intervention Dialogue
CTRL	Ptc45	Constant	NA	NA
EG1	Mnk19	Constant + IV1	Reputation, image and ranking	In addition, imagine you read the following post about this Uni on $smp1: "The Times Higher Education ranked the XYZ University among the top universities in the country for a range of disciplines." Moreover, you read this message about the same Uni on $smp2: "XYZ University's faculty and research are world-renowned, as it has excellent reputation and image both nationally and internationally …"
EG2	Knr24	Constant + IV2	Living and study costs, and availability of scholarships	In addition, imagine you read the following post about this University on $smp1: "The XYZ University is quite affordable, and many students are on full scholarship anyway…" Moreover, you read this message about the same Uni on $smp2: "I lived on and off campus whilst studying at XYZ University and I must say it was much more affordable than many other places …"
EG3	Hpm38	Constant + IV3	Work and internship placements during study and job prospects upon graduation	In addition, imagine you read the following post about this University on $smp1: "The XYZ University helped me find a good internship while studying which upon graduation led to my first full-time job at a reputable firm …" Moreover, you read this message about the same Uni on $smp2: "I know for a fact that XYZ University has a great career network, plenty of opportunities …"
EG4	Gwn42	Constant + IV4	Ease of admission, entrance requirements and open communication with admissions staff	In addition, imagine you read the following post about this University on $smp1: "My admission process at XYZ University was fast and easy, as the entrance requirements were not hard to meet at all …" Moreover, you read this message about the same Uni on $smp2: "I had a pleasant experience with the XYZ University's admissions staff: they were responsive and quick to guide me through the whole process…"
EG5	Bmr57	Constant + IV5	Campus location (proximity to home, convenience and comfort), safety and physical appeal, and vibe of the city	In addition, imagine you read the following post about this University on $smp1: "XYZ Uni is centrally located which is important to me because I can visit my parents anytime I want since home is not far away…" Moreover, you read this message about the same Uni on $smp2: "I love the city and XYZ Uni campus as it is safe, conveniently located, vibrant and close to many attractions…"
EG6	Mha68	Constant + IV6	Availability, flexibility and attractiveness of the course and on-campus support services	In addition, imagine you read the following post about this University on $smp1: "The flexibility of the program I'm currently studying at XYZ University suits my work-study-life balance, it is also quite relevant to my career aspirations …" Moreover, you read this message about the same Uni on $smp2: "I am really happy with the availability of the courses and on-campus support I've received at XYZ University…"
EG7	Ghw71	Constant + IV7	Prior knowledge of the study destination	In addition, imagine you read the following post about this University on $smp1: "I certainly enjoy the perks of knowing the place before I even started studying here at XYZ University …" Moreover, you read this message about the same Uni on $smp2: "I could quickly adjust to the city of XYZ University because I'd lived there for a while before I was enrolled …"
EG8	Yrk86	Constant + IV8	Collaboration with other universities	In addition, imagine you read the following post about this University on $smp1: "XYZ University has research collaborations with many other universities all around the world." Moreover, you read this message about the same Uni on $smp2: "Thanks to XYZ University's student exchange arrangements, I can choose to spend a whole year at a university in another country …"

Table 11: Group ids, random allocation codes, module (Constant and IV code), IV factors and intervention dialogues. CTRL: Control Group; EG: Experimental Group; IV: Independent Variable; RA: Random Allocation; $smp1: User's most favourite social media platform; $smp2: User's second most favourite social media platform

I preferred to use an AI-led chatbot to collect primary data for the experiment due to three key reasons. First, I want the AI to build rapport with the participants and create a more anthropomorphic interaction experience than what online surveys can achieve (see the example dialogue below), so that response and completion rates will be improved, and the participants would be more likely to reveal their true opinions and provide us with richer and more insightful information particularly in Intervention modules (IV1, …, IV8) (Table 11). Second, through CFLC design I strived to minimize ambiguations and misinterpretations for open-ended responses. Third, the AI is programmed to ensure the double blindness of the experiment. During the conversation each participant is randomly allocated by the AI to one of the 9 groups real-time and human researchers have neither any impact on nor knowledge of this allocation.

Example Dialogue

Chatbot I: Hi!
My name is Sydn-E. I'm an AI-powered chatbot developed by researchers at the University of Sydney Business School.
In case you haven't already, please check these out:
Participant Consent Form [Appendix B]; *Participant Information Statement* [Appendix A] (links to online forms)
Thank you so much for taking part in this study. I'm excited to chat with you.
What's your name?
P (Participant): Hey, hi. I'm Dominic.
C: Nice to meet you, Dominic.
Are you currently studying at a university?
P: Yes, I am
C: Great! Where is it?
P: Australia
C: When did you start studying there?
P: in 2022
C: Ok, I know some people use social media more than others, but I'm interested to know which social media platforms you normally use.
What is your most favorite social media platform?
P: Insta
C: Got it. What is your second most favorite social media platform?
P: Youtube
C: Ok, now, please type this code: Mnk11
P: Mnk11
Ok. Great! Now, let me give you a scenario:
Imagine you are seeking information before enrolling in a university and you've come across the following statements on a university website:
[[CONSTANT]]
[[IV1]]
C: Based on what you've read so far Dominic, how likely would you enrol in this Uni?
5: Very likely 4: Likely 3: Neutral 2: Unlikely 1: Very unlikely
P: Likely
C: Good to know. Thanks. What factors about a Uni that you read on Instagram or YouTube would shape your mind in favor of studying there?
P: graduate outcomes and career opportunities
C: What other factors would you consider important when choosing a university?
P: high quality students produced

C: When you choose a university how important do you think its reputation, image and global ranking are?
5: Very important 4: Important 3: Neutral 2: Unimportant 1: Very unimportant
P: Important
C: What about living and study costs, availability of scholarships etc.?
Which Uni would you prefer in terms of overall costs?
5: Highest priced. 4: Higher than average. 3: Average. 2: Lower than average. 1: Lowest priced
P: Average
C: What about work and internship placements during study, job opportunities and potential work-related benefits after graduation? When you choose a university how important do you think this is?
5: Very important 4: Important 3: Neutral 2: Unimportant 1: Very unimportant
P: Neutral
C: Alright, Dominic. I'm just curious to know some more …
When you choose a university how important is its ease of admission, considering its entrance requirements and open communication with admissions staff?
5: Very important 4: Important 3: Neutral 2: Unimportant 1: Very unimportant
P: Unimportant
C: What about the location of the campus?
When you choose a university how important are factors such as: its proximity to home, convenience and comfort, safety and physical appeal, and vibe of the city?
5: Very important 4: Important 3: Neutral 2: Unimportant 1: Very unimportant
P: Important
C: How important to you are the flexibility and attractiveness of the course/program of study (in line with your career aspirations and earning potential) and on-campus support services?
5: Very important 4: Important 3: Neutral 2: Unimportant 1: Very unimportant
P: Neutral
C: When you choose a university how important do you think whether you have prior knowledge of the study destination (city)?
5: Very important 4: Important 3: Neutral 2: Unimportant 1: Very unimportant
P: Unimportant
C: And finally, Dominic, when you choose a university how important to you is its collaboration with other universities?
5: Very important 4: Important 3: Neutral 2: Unimportant 1: Very unimportant
P: Unimportant
C: That's very helpful. Thanks a lot, Dominic.
By the way, how old are you?
P: 22
C: Are you a domestic or international student?
P: International
C: Would you like to add anything else before we say our goodbyes?
P: I don't think so.
C: OK, then. I really enjoyed chatting with you and will always remember this conversation ☺
Take care, Dominic, bye …
P: bye

6.4.4 Validity, reliability, and replicability

To eliminate confounders and bias, and hence ensure high internal validity (for causal inference) and improve external validity (for generalizability of the results to other contexts), we adopt a double-blind true experimental design incorporating random allocation of participants to Control and Experimental groups, and an AI-led unstructured random allocation of interventions to participants in Experimental groups (Egs). To test hypotheses, we will empirically compare Egs that were positively intervened against the Control group (CTRL).

To ensure high internal validity with this experiment through randomization, we aim to rule out competing explanations for any differences that the AILC captures between the CTRL and an EG, other than the sole effect of the intervention given to that EG. In other words, we aim to control for any unknown or known variables other than the intervention that may impact the outcome of the DV (i.e., decision of a participant). To do so, we ensured blind random allocation of unidentified participants to one of the Egs or CTRL and designed the AILC to administer only 1 intervention at a time to an EG. Random allocation of participants is handled by randomly assigning a code to each participant. For example, participants to whom are assigned "Knr24" will be allocated by AI to EG2 (Table 11). Although these codes are visible to participants for validation, they will have no knowledge of being allocated to which of the 9 groups. The random allocation codes and their corresponding groups are shown below.

As an internal reliability measure, we checked intra-item consistency by programming the chatbot to ask the primary query of each IV after invention in a given chat to divulge participants who responded to questions without care. If, for instance, the AILC prompted the "job prospects" intervention (EG3), then it would later ask: "What about work and internship placements during study, job opportunities and potential work-related benefits after graduation? When you choose a university how important do you think this is?" Since these responses were collected on a 5-point Likert scale, only the responses within [0-2] point discrepancy range between two analogous queries would be accepted. Hence, the data belonging to responses greater than [0-2] range (e.g., 5: Very likely – 2: Unlikely = 3 point discrepancy) would be excluded from all further analyses. Finally, for replicability, we include across multiple sections of this paper all the necessary details of the materials, processes, procedures, strategies, and rationale used to conduct these experiments.

6.4.5 Hypotheses

The hypotheses stated below are related to the second overarching aim of this thesis: to make causal inferences regarding what factors in the form of electronic word-of-mouth (eWOM) from social media impact students' university choices.

H_0: Social media content in the form of positive eWOM about a university has no effect on students' likelihood to enrol in that university.

H_1: Positive eWOM on social media about a university's reputation, image, and ranking increases the likelihood for students to enrol in that university.

H_2: Positive eWOM on social media about a university's living and study costs, availability of scholarships and access to technology, research, and facilities increases the likelihood for students to enrol in that university.

H_3: Positive eWOM on social media about a university's work and internship placements during study and job prospects upon graduation increases the likelihood for students to enrol in that university.

H_4: Positive eWOM on social media about a university's ease of admission, entrance requirements and open communication with admissions staff increases the likelihood for students to enrol in that university.

H_5: Positive eWOM on social media about a university's campus location including proximity to home, convenience and comfort, safety, physical appeal, and vibe of the city increases the likelihood for students to enrol in that university.

H_6: Positive eWOM on social media about a university's availability, flexibility and attractiveness of the course and on-campus support services increases the likelihood for students to enrol in that university.

H_7: Positive eWOM on social media about students' prior knowledge of the study destination increases the likelihood for students to enrol in that university.

H_8: Positive eWOM on social media about a university's collaboration with other universities increases the likelihood for students to enrol in that university.

6.4.6 Interview strategy

While devising the interview questions and strategy, I programmed the AILC (Sydn-E) to ask open-ended yet semi-structured questions (SSQ) to reveal as much information about the topic as possible and at the same time to address the objectives of the study. These questions were prepared to be not only easy-to-understand but also sensible, relevant, and neutral. As a strategy, we start off with questions that the respondents can easily answer such as "Are you currently studying at a university?", "which one?", and "when did you start?" Then we proceed to more intricate matters such as factors that may have affected their decision to study there and whether reading eWOM from social media about these factors may have any positive or negative impact on their decision. We aim to put participants at ease and build up rapport and confidence with them. As a result, we hope to see that they open up and provide rich insights that help develop the interview further.

As the interviews progress, Sydn-E does not interfere with the respondents' story telling at any stage even if they go off topic. However, she utilizes CFL to bring respondents back on track if necessary. Since it is our main goal to extract information about the matriculation decision factors, respondents are prompted to not only determine the level of importance for all pre-coded and defined factors but also talk about any other non-defined factors that could impact their university choices.

6.4.7 Power analysis and sample size

I conducted a Power analysis to determine what sample size would ensure a high probability of accurately rejecting the Null Hypothesis that there is no difference between the control and experimental groups. Power is mainly affected by how much overlap there is between two compared groups' distributions, and the number of participants in each group: sample size. A Power of 0.95 (high Power) simply means that we want to have at least 95% chance of accurately rejecting the Null Hypothesis. Then, if there is very little overlap between compared groups' distributions, a small sample size may be sufficient to yield high Power. However, the larger the overlap, the larger the sample size must be to attain high Power. Moreover, Central Limit Theorem (CLT) tells us that these implications apply to both normal and nonnormal distributions with a mean.

To determine the required sample size a priori using Power analysis with a two-tailed t-test between two independent groups' means:

- Minimum sample size for each group would be 32 (two groups of n=64 (Figure 22)), if we were to achieve a minimum of 0.5 Power (1-β err prob) with a significance level of 0.05, sd of 1, allocation ratio (N2/N1) of 1, and d (Cohen's difference:effect size) of 0.5 (medium effect size). Since we have 9 groups in total, in this case we must recruit at least 288 eligible respondents for the experiment.

- Minimum sample size for each group would be 64 (two groups of n=128 (Figure 22)), if we were to achieve a minimum of 0.8 Power (1-β err prob) with a significance level of 0.05, sd of 1, allocation ratio (N2/N1) of 1, and d (Cohen's difference:effect size) of 0.5 (medium effect size). Since we have 9 groups in total, in this case we must recruit at least 576 eligible respondents for the experiment.

However, in these cases we could only account for larger than medium size (d=0.5) differences between compared groups. To account for small size differences, as a common practice Cohen's d would be taken as 0.2.

- However, if we were to achieve a minimum of 0.8 Power (1-β err prob) with a significance level of 0.05, sd of 1, allocation ratio (N2/N1) of 1, and d (Cohen's difference:effect size) of 0.2 (small effect size), minimum sample size for each group would be 394. Since we have 9 groups in total, in this case we must recruit at least 3546 eligible respondents for the experiment.

- Alternatively, if we were to achieve a minimum of 0.5 Power (1-β err prob) with a significance level of 0.05, sd of 1, allocation ratio (N2/N1) of 1, and d (Cohen's difference:effect size) of 0.2 (small effect size), minimum sample size for each group would be 194. Since we have 9 groups in total, in this case we must recruit at least 1746 eligible respondents for the experiment.

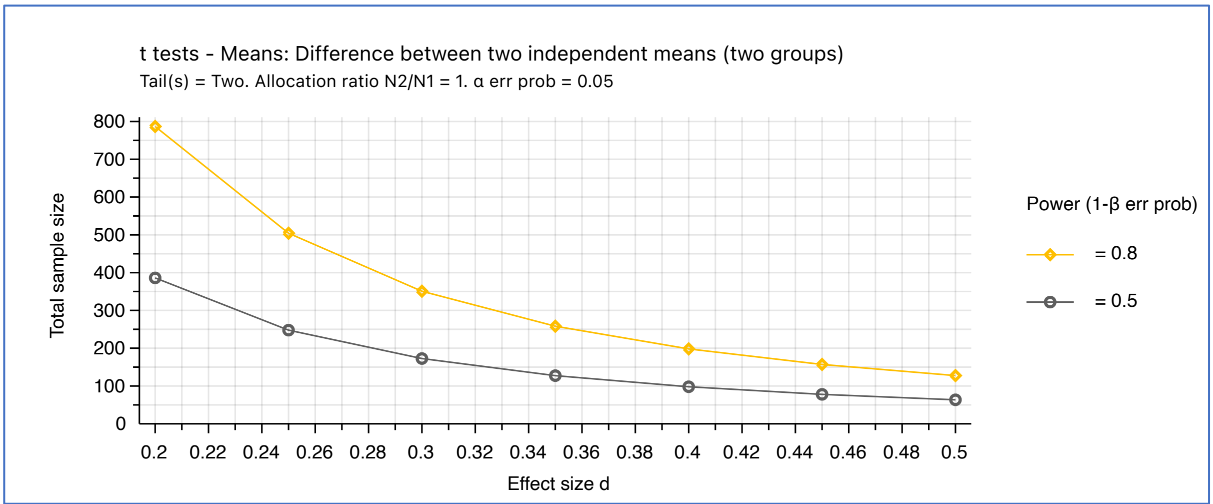

Figure 22. Total sample size of 2 groups of equal *n* by Power (1-β err prob) and d (Effect size). Output and graph created by G*Power 3.1

Researchers have been debating about whether to use parametric versus nonparametric tests to analyse Likert-scaled data. Since Likert-scaled data are discrete, ordinal and have a limited range, they tend to violate the normality assumption of parametric tests. However, Dr Norman, one of the pioneers in medical education research methodology (Sullivan, 2013), provided convincing evidence that parametric tests can potentially be applied to ordinal data, such as those obtained from Likert scales (Norman, 2010). He also established that in addition to being more robust than nonparametric tests, even when statistical assumptions – such as data normality – are violated to a great degree, parametric tests mostly provide accurate results. The study concluded that parametric tests are resilient enough to help researchers reach generally unbiased conclusions that are reasonably close to the actual "truth" when evaluating responses from Likert scales (Norman, 2010; Sullivan, 2013). Therefore, I hold the opinion that with a large enough random sample of independent groups with similar variance, a parametric test such as 2-sample t-test can be valid with nonnormal Likert-scaled data.

Non-parametric tests such as Mann-Whitney test, on the other hand, can be considered valid without the normality assumption. Yet, they are commonly believed to have a lower probability of detecting a real effect between groups than parametric tests. A simulation study by de Winter and Dodou (2010) investigated the capabilities of 2-sample t-test and Mann-Whitney test on 5-point Likert-scaled data. They found that for most of the distribution pairs, both tests had similar type I error rates and power even when sample sizes for each group were 10, 30 or 200. Therefore, regardless of the outcome of data normality, I chose to analyze the results from Likert scale by using both 2-sample t-test and Mann-Whitney test. Accordingly, an estimation via G*Power 3.1 demonstrates that to achieve the same power (e.g., 0.95) with an effect size of 0.5 and significance level of 0.05, the required sample size per group would be 105 for 2-sample t-test and 122 for Mann-Whitney test on nonnormal distribution.

Since I don't have the budget to recruit considerably a large number of participants to account for 'small effects' (effect size: d = 0.2) for 9 groups, I must sacrifice the effect size (d) as I can at best account for a moderate impact (effect size: d= 0.5) on the outcome being measured (i.e., moderate difference between control and experimental groups being compared) and reject the null hypothesis only if the measured difference (effect) is at least medium. Therefore, if we are to achieve a power of 0.95 (1-β err prob) and just in case for nonnormal distribution of data, we must recruit minimum 122 participants per group. So, just to account for potential dropouts or missing data, we better aim to recruit 1200 participants in total for 9 groups (≈134 participants/group). This would be sufficient for both 2-sample t-test and Mann-Whitney test. Therefore, as a result of Power analysis, we may expect to have a 95% chance that we would accurately reject the Null Hypothesis via 2-sample t-test and Mann-Whitney test if we recruit at least 122 participants per group. Moreover, since after an intervention all respondents were asked during the chat about the role of all other factors in their decision, we would have supplementary quantitative data to support the test results of each hypothesis.

6.4.8 Participant recruitment from Prolific Academic

Prolific is a reputable survey company (Lee & Tipoe, 2022) that provides a web-based platform developed to help mostly academic researchers recruit participants from various countries for research studies (Hachaturyan et al., 2021). Prolific has been considered to produce higher data quality in comparison to analogous online data collection platforms such as Amazon Mechanical Turk (Lee & Tipoe, 2022; Palan & Schitter, 2018). Moreover, Moeck et al. (2022) stated that Prolific is getting increasingly popular among academic researchers because it offers rapid, remote, and affordable access to participants. According to Prolific (2022), there are more than 25,000 researchers and 130,000 participants using Prolific worldwide and it is trusted by organizations such as Google, European Commission, University of Oxford and nugget.ai.

Prolific allows researchers to select samples that come with certain attributes. Participants provide their demographical, geographical, and other information while signing up to Prolific as they enter the participant pool. Prolific performs rigorous checks and screening on them to ensure data quality (i.e., participants are who they say they are). Prolific matches the sample requirements of researchers with eligible participants and ensures each participant is paid at least $8(US)/hour. Prolific also requests researchers to determine an estimated time for a participant to complete the survey so that participants can be paid on a pro rata basis. We requested a total of 1200 eligible participants from Prolific. We estimated the time to complete the chatbot survey to be 7 minutes and paid each participant $1.31 ($11.23/hr) which was suggested by Prolific as a “Good” payment.

We accepted participants aged 18-30 from all around the world. Since Prolific allows us to set pre-screening criteria before participant selection, we requested participants who have at least completed high school, use at least one social media platform and speak English natively. Since dialogues in the chat require participants to possess a high level of fluency in English language, to prevent misunderstandings, overcome language barrier, and hence collect richer, more valid and comprehensive data, we recruited participants who had told Prolific that their first language was English. To avoid participants from merely a few countries or regions to take up all available yet limited Prolific spots (i.e. 1200) and dominate survey responses, based on real-time feedback we received from Sydn-E, we aired the chat survey thru Prolific at five different times excluding any dominating countries in the subsequent instances of the survey. We also activated the option on Prolific allowing us to exclude participants who have already been recruited in the previous instances of the survey. As a result, we could maintain a more balanced representation of the true population covered by all habitable continents in the world and thus improve the study's external validity.

6.4.9 Integration with Prolific

Sydn-E Chat's URL was provided as a link for participants recruited by Prolific as each participant was redirected from Prolific app to a Google site where Sydn-E was hosted. To match participants' demographic data with their answers and improve the quality of submissions – as recommended by Prolific – we recorded each participant's Prolific ID by including a question (i.e. "What's your unique Prolific ID?") at the beginning of the chat. Prolific provides participants with a quick way to copy their unique IDs so that they can easily paste them when needed.

When participants commence our study, they leave the Prolific app. To prove that they completed the study, Prolific needs to capture a unique *completion code* when participants return to the Prolific app. We confirmed participants' completion of the study by redirecting them with a URL at the end of the chat. Rather than using a single completion code, we used custom completion codes to differentiate between eligible and non-eligible submissions. If, for example, a participant responded "No, I was never enrolled in one and I don't intend to anytime soon" to "Are you currently studying at a higher education institution?", then they would be considered non-eligible and given a different code than the eligible participants. When a participant is identified as eligible, Sydn-E prompts the participant to click the link which notifies Prolific of their eligibility. If, however, this link is not clicked in a given time or the non-eligibility link is clicked, then Prolific identifies them as non-eligible. On Prolific app, where eligible participants are approved, non-eligible ones are rejected, and new spaces open up for their substitutes. Both approval and rejection actions are automated but manually reversable.

In total, 1223 participants completed the chat survey. Sydn-E rejected 2.45% of them (30/1223) on the grounds of [1] intra-item inconsistency or [2] lack of attention or inadequate input. An example of [1]: intra-item inconsistency is when a participant in EG1 (reputation & global ranking group) answered “Very unlikely” to the enrolment questions, but then answered “Very important” to the reputation and global ranking question. Sydn-E detected only 8 cases with this issue. When a participant entered something unmeaningful (e.g. just a number or nothing) or inadequate text (e.g. “ok”, “not sure”, “yes”, “of course”) in both open-ended questions, this was considered a case of [2]: lack of attention or inadequate input. Sydn-E detected 22 cases with this issue. Prolific provided Sydn-E with replacements for the participants who did not complete the survey or the ones whose submissions were rejected by Sydn-E (and confirmed by us manually) due to intra-item inconsistency or lack of attention or inadequate input. However, after we accepted the eligible participants on Prolific, we realized that 7 of them were duplicates (same participants). So, we removed their responses from analysis. As a result, our final sample size was 1193. Upon reviewing the allocation of participants to the CTRL and Experimental Groups, we observed that all of the groups met the minimum requirement of participant numbers (>122), as determined earlier by the Power analysis.

6.5 Results and discussion

6.5.1 Quantitative data insights

Sydn-E collected two main types of quantitative data from eligible participants: (1) Categorical (e.g., Study Status, Location, Domestic/International, frequently used Social Media Platforms), and (2) Numerical (e.g., Age, Experiment responses, and importance of university choice factors on 5-point Likert scale). It is worth noting that Sydn-E was capable of fuzzy matching and disambiguation as it was trained to recognize the name of a categorical variable even if the user entered different versions of it (e.g., insta for Instagram) or used one version in a sentence ("I mostly use insta"). However, if no pretrained social platform name could be recognized, a retype prompt would follow the participant response within a while loop: "I am not familiar with this social media platform, please retype your most favourite social media platform!". Similarly, while collecting numerical data, different versions of the same value responses (e.g., 5= "Very likely", 5= "that's very important") were recorded as a single coherent numerical value (e.g., 5).

The categorical data results have provided us valuable insights regarding the distribution of participant features. Sydn-E recorded that 44.7% of all eligible respondents were current students, 30% of them were enrolled in a university in the last 5 years, 14.8% of them were enrolled in a university more than 5 years ago, and 10.6% of them intended to study at a university in the future (Figure 23). In addition, most participants were domestic students and in terms of geographical distribution, they were studying or had studied in the UK (22.1%), the USA (21.3%), Africa (20.1%), Europe (11.2%), Australia and New Zealand (8.2%), Asia (7.8%), Canada (6.6%), and South America (2.7%) (Figure 24). Sydn-E also recorded that where the Mean Age of all eligible participants was 24.7, this average was higher for the users of Facebook (26.2), Reddit (25.8), LinkedIn (25.6), and WhatsApp (25.1), it was lower for the users of other social media platforms such as YouTube (24.3), Twitch (23.3), TikTok (23.3), Discord (23.1), and Snapchat (21.8) (Figure 25). Instagram and Twitter, two of the top three most frequently used social media platforms, shared a common mean age of around 24.7-24.8, aligning with the average age of the whole sample. Instagram, by far the most frequently used social media platform, was followed by TikTok, Twitter and Facebook (Figure 25).

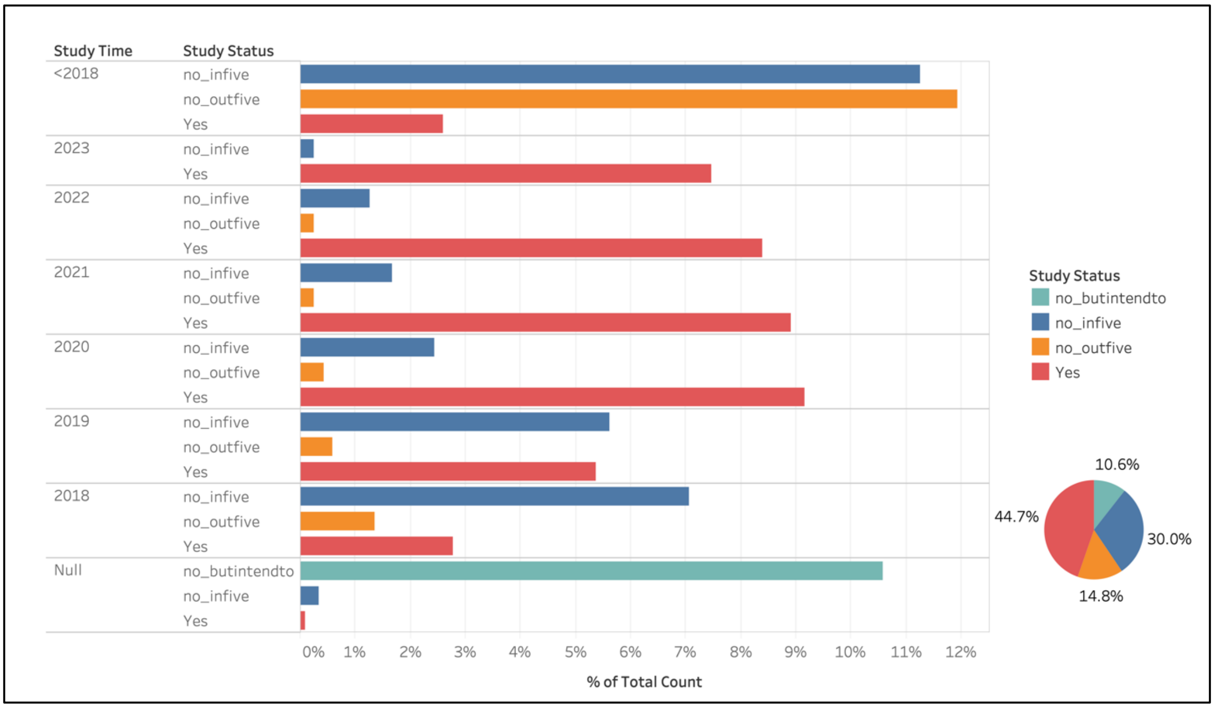

Figure 23: Distribution of Study Time (i.e., Year) and Study Status of participants by percentage of total count (*N*).
Note 1. Study Status question: "Are you currently studying at a higher education institution?"
Yes: "Yes, I am." | no_butintendto: "No, but I intend to enrol in one." | no_infive: "No, but I was enrolled in one in the last 5 years." | no_outfive: "No, but I was enrolled in one more than 5 years ago."
Note 2. Study Time question: "When did you start studying there?"

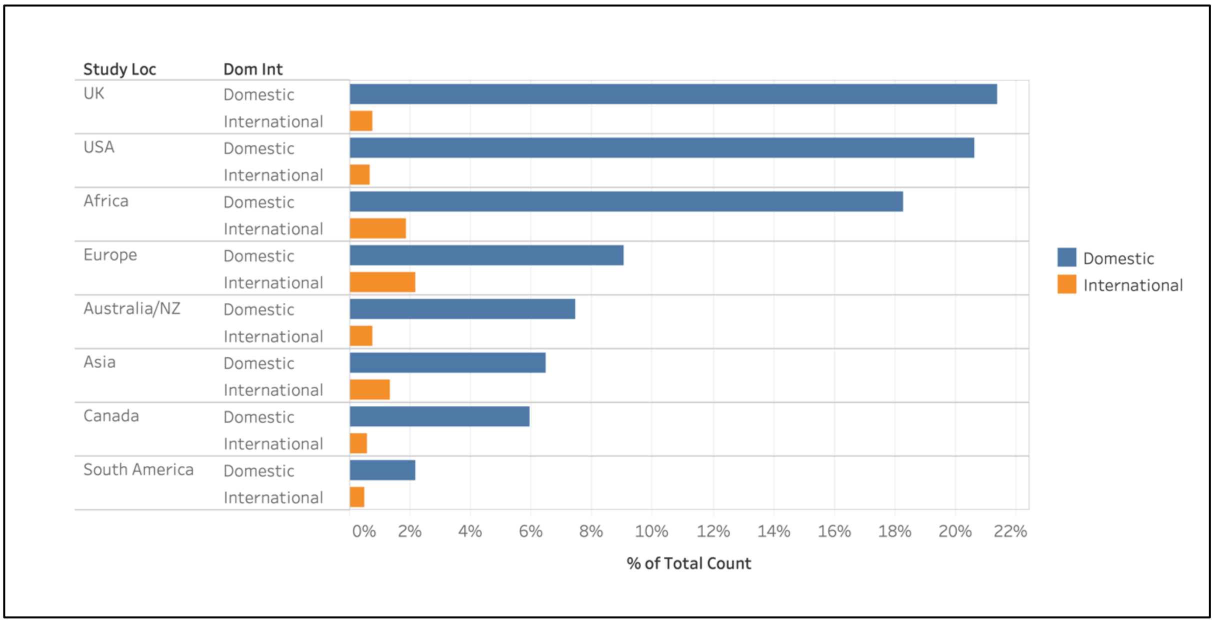

Figure 24: Distribution of Study Locations (Study Loc) and whether a participant considers I domestic or international student (Dom Int) by percentage of total count (*N*).
Note 1. Study Loc question: "Where is it?" (The higher education institution)
Note 2. Dom Int question: "Would you consider yourself a domestic or international student?"

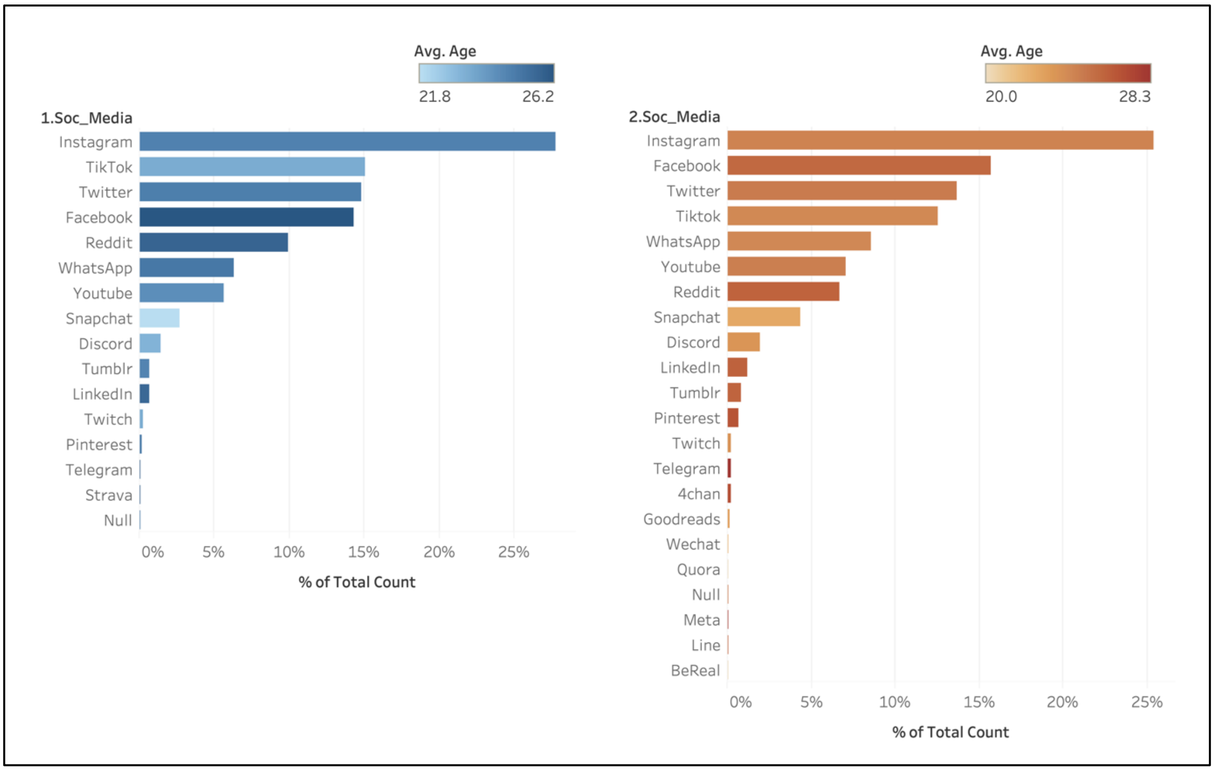

Figure 25: Distribution of the most frequently used social media platforms (1. Soc_Media), the second most frequently used social media platforms (2. Soc_Media) and the Mean Age (Avg. Age) of participants by percentage of total count (*N*).

6.5.2 Experiment results

Descriptive statistics

When we inspect the descriptive statistics pertaining to the Control (CTRL) and Experimental Groups' (Egs) results, we notice that CTRL had the lowest mean (3.74), whereas EG1 had the highest mean (4.3) among all groups (Table 12). The Egs with the lowest means were EG4 (3.76), EG7 (3.83) and EG8 (3.91). The median of all groups was 4 except for EG1 which had a median of 5. We also notice that all groups contained between 122 and 144 participants (Table 12). This is in line with what we aspired to achieve in accordance with Power Analysis before Sydn-E ran the experiment. More importantly we found that 100% of the eligible participants were randomly allocated by Sydn-E to one of the nine groups seamlessly. This was because each participant could successfully confirm the RA code which was neither case nor whitespace sensitive and Sydn-E was capable of disambiguating, fuzzy matching, and handling typos.

	CTRL	*EG1*	*EG2*	*EG3*	*EG4*	*EG5*	*EG6*	*EG7*	*EG8*
Mean	3.7431	4.2721	4.1042	4.1926	3.7623	4.1212	4.1308	3.8346	3.9106
Standard Error	0.0762	0.0805	0.0722	0.0652	0.0856	0.0745	0.0728	0.0804	0.0730
Median	4	5	4	4	4	4	4	4	4
Mode	4	5	4	4	4	4	4	4	4
SD	0.9141	0.9386	0.8668	0.7580	0.9452	0.8563	0.8296	0.9064	0.8099
Sample Variance	0.8356	0.8810	0.7513	0.5746	0.8934	0.7333	0.6882	0.8216	0.6559
Kurtosis	-0.0628	2.0433	0.3506	-0.7922	0.9065	1.5491	0.0525	0.9583	-0.6665
Skewness	-0.4126	-1.4438	-0.7921	-0.4426	-0.8782	-1.0513	-0.7463	-0.8344	-0.2109
Range	4	4	4	3	4	4	3	4	3
Minimum	1	1	1	2	1	1	2	1	2
Maximum	5	5	5	5	5	5	5	5	5
Sum	539	581	591	566	459	544	537	487	481
Count	144	136	144	135	122	132	130	127	123

Table 12: Descriptive Statistics of the participant responses in Control (CTRL) and Experimental Groups (EG1:EG8).

Hypothesis testing results

As discussed earlier, we used both two-sample t-test for comparing the means between the Control Group and each one of the eight Experimental Groups, and Mann-Whitney U test (Wilcoxon rank-sum test) to determine whether there is a significant difference between the distributions of these compared groups. Importantly, the results of both tests were consistent (Table 13) indicating that the p-values of the following compared groups: CTRL & EG1, CTRL & EG2, CTRL & EG3, CTRL & EG5, and CTRL & EG6 were less than 0.001 (statistically significant); whereas the p-values of other groups, namely CTRL & EG4, CTRL & EG7, and CTRL & EG8 were larger than 0.1. Since five factors in five Egs (EG1, EG2, EG3, EG5, and EG6) were statistically significant, we can reject the Null hypothesis: H_0: Social media content in the form of positive eWOM about a university has no effect on students' likelihood to enroll in that university.

Compared Groups	Impact Rank	Welch Two Sample t-test				Mann-Whitney U test	
		t	df	95%CI	p-value	W	p-value
CTRL vs EG1	*1*	-4.774	276.06	[-0.747, -0.311]	<0.001*	6420.0	<0.001*
CTRL vs EG2	*5*	-3.444	285.85	[-0.566, -0.154]	<0.001*	8093.5	<0.001*
CTRL vs EG3	*2*	-4.482	272.98	[-0.647, -0.252]	<0.001*	7094.0	<0.001*
CTRL vs EG4	-	-0.168	253.86	[-0.245, 0.206]	0.867	8503.5	0.634
CTRL vs EG5	*4*	-3.548	273.87	[-0.588, -0.168]	<0.001*	7202.5	<0.001*
CTRL vs EG6	*3*	-3.681	271.99	[-0.595, -0.180]	<0.001*	7098.0	<0.001*
CTRL vs EG7	-	-0.827	265.32	[-0.310, 0.127]	0.409	8516.0	0.300
CTRL vs EG8	-	-1.588	264.64	[-0.375, 0.040]	0.114	8044.5	0.171

Table 13: Statistics of the Experiment results incorporating Welch Two-Sample t-test and Mann-Whitney U test.
Note1: Mann-Whitney U test: Wilcoxon rank-sum test with continuity correction
Note2: Impact rank is based on the t statistic.
* Statistically significant at 0.001 level

Hypothesis	Description	Result
H_0	Social media content in the form of positive eWOM about a university has no effect on students' likelihood to enroll in that university.	Reject
H_1	Positive eWOM on social media about a university's reputation, image, and ranking increases the likelihood for students to enroll in that university.	Accept
H_2	Positive eWOM on social media about a university's living and study costs, availability of scholarships and access to technology, research, and facilities increases the likelihood for students to enroll in that university.	Accept
H_3	Positive eWOM on social media about a university's work and internship placements during study and job prospects upon graduation increases the likelihood for students to enroll in that university.	Accept
H_4	Positive eWOM on social media about a university's ease of admission, entrance requirements and open communication with admissions staff increases the likelihood for students to enroll in that university.	Not accept
H_5	Positive eWOM on social media about a university's campus location including proximity to home, convenience and comfort, safety, physical appeal, and vibe of the city increases the likelihood for students to enroll in that university.	Accept
H_6	Positive eWOM on social media about a university's availability, flexibility and attractiveness of the course and on-campus support services increases the likelihood for students to enroll in that university.	Accept
H_7	Positive eWOM on social media about students' prior knowledge of the study destination increases the likelihood for students to enroll in that university.	Not accept
H_8	Positive eWOM on social media about a university's collaboration with other universities increases the likelihood for students to enroll in that university.	Not accept

Table 14: Descriptions and results of hypotheses

Specifically, Table 13 in tandem with Table 14 can be interpreted for each alternative hypothesis as follows:

H_1: Positive eWOM on social media about a university's reputation, image, and ranking increases the likelihood for students to enroll in that university.

Since the p-values of both t-test and Mann-Whitney test for CTRL & EG1 are extremely small (<0.001) and substantially less than the commonly used significance level of 0.05 and even 0.01, there is strong evidence against the null hypothesis (t-test: true difference in means is equal to 0; Mann-Whitney test: true location shift is equal to 0). Furthermore, the negative t-value (-4.774) suggests that EG1 has a higher mean compared to CTRL, and the 95 percent confidence interval (95%CI) provides the range [-0.747, -0.311] within which the true difference in means likely falls (Table 13). Therefore, we accept H_1 and confirm that positive eWOM on social media about a university's reputation, image, and ranking increases the likelihood for students to enroll in that university (Table 14).

H_2: Positive eWOM on social media about a university's living and study costs, availability of scholarships and access to technology, research, and facilities increases the likelihood for students to enroll in that university.

Since the p-values of both t-test and Mann-Whitney test for CTRL & EG2 are extremely small (<0.001) and substantially less than the commonly used significance level of 0.05 and even 0.01, there is strong evidence against the null hypothesis (t-test: true difference in means is equal to 0; Mann-Whitney test: true location shift is equal to 0). Furthermore, the negative t-value (-3.444) suggests that EG2 has a higher mean compared to CTRL, and the 95 percent confidence interval (95%CI) provides the range [-0.566, -0.154] within which the true difference in means likely falls (Table 13). Therefore, we accept H_2 and confirm that positive eWOM on social media about a university's living and study costs, availability of scholarships and access to technology, research, and facilities increases the likelihood for students to enroll in that university (Table 14).

H_3: Positive eWOM on social media about a university's work and internship placements during study and job prospects upon graduation increases the likelihood for students to enroll in that university.

Since the p-values of both t-test and Mann-Whitney test for CTRL & EG3 are extremely small (<0.001) and substantially less than the commonly used significance level of 0.05 and even 0.01, there is strong evidence against the null hypothesis (t-test: true difference in means is equal to 0; Mann-Whitney test: true location shift is equal to 0). Furthermore, the negative t-value (-4.482) suggests that EG3 has a higher mean compared to CTRL, and the 95 percent confidence interval (95%CI) provides the range [-0.647, -0.252] within which the true difference in means likely falls (Table 13). Therefore, we accept H_3 and confirm that positive eWOM on social media about a university's work and internship placements during study and job prospects upon graduation increases the likelihood for students to enroll in that university (Table 14).

H_4: Positive eWOM on social media about a university's ease of admission, entrance requirements and open communication with admissions staff increases the likelihood for students to enroll in that university.

Since the p-values of both t-test and Mann-Whitney test for CTRL & EG4 are larger than the commonly used significance level of 0.05 and even 0.1, there is not enough evidence against the null hypothesis (t-test: true difference in means is equal to 0; Mann-Whitney test: true location shift is equal to 0). Furthermore, we also notice that the true difference in means falls within the 95%CI range of [-0.245, 0.206] (Table 13). Therefore, we cannot accept H_4 and cannot state that positive eWOM on social media about a university's ease of admission, entrance requirements and open communication with admissions staff increases the likelihood for students to enroll in that university (Table 14).

H_5: Positive eWOM on social media about a university's campus location including proximity to home, convenience and comfort, safety, physical appeal, and vibe of the city increases the likelihood for students to enroll in that university.

Since the p-values of both t-test and Mann-Whitney test for CTRL & EG5 are extremely small (<0.001) and substantially less than the commonly used significance level of 0.05 and even 0.01, there is strong evidence against the null hypothesis (t-test: true difference in means is equal to 0; Mann-Whitney test: true location shift is equal to 0). Furthermore, the negative t-value (-3.548) suggests that EG5 has a higher mean compared to CTRL, and the 95 percent confidence interval (95%CI) provides the range [-0.588, -0.168] within which the true difference in means likely falls (Table 13). Therefore, we accept H_5 and confirm that positive eWOM on social media about a university's campus location including proximity to home, convenience and comfort, safety, physical appeal, and vibe of the city increases the likelihood for students to enroll in that university (Table 14).

H_6: Positive eWOM on social media about a university's availability, flexibility and attractiveness of the course and on-campus support services increases the likelihood for students to enroll in that university.

Since the p-values of both t-test and Mann-Whitney test for CTRL & EG6 are extremely small (<0.001) and substantially less than the commonly used significance level of 0.05 and even 0.01, there is strong evidence against the null hypothesis (t-test: true difference in means is equal to 0; Mann-Whitney test: true location shift is equal to 0). Furthermore, the negative t-value (-3.681) suggests that EG6 has a higher mean compared to CTRL, and the 95 percent confidence interval (95%CI) provides the range [-0.595, -0.180] within which the true difference in means likely falls (Table 13). Therefore, we accept H_6 and confirm that positive eWOM on social media about a university's availability, flexibility and attractiveness of the course and on-campus support services increases the likelihood for students to enroll in that university (Table 14).

H_7: Positive eWOM on social media about students' prior knowledge of the study destination increases the likelihood for students to enroll in that university.

Since the p-values of both t-test and Mann-Whitney test for CTRL & EG7 are larger than the commonly used significance level of 0.05 and even 0.1, there is not enough evidence against the null hypothesis (t-test: true difference in means is equal to 0; Mann-Whitney test: true location shift is equal to 0). Furthermore, we also notice that the true difference in means falls within the 95%CI range of [-0.310, 0.127] (Table 13). Therefore, we cannot accept H_7 and cannot state that positive eWOM on social media about students' prior knowledge of the study destination increases the likelihood for students to enroll in that university (Table 14).

H_8: Positive eWOM on social media about a university's collaboration with other universities increases the likelihood for students to enroll in that university.

Finally, since the p-values of both t-test and Mann-Whitney test for CTRL & EG8 are larger than the commonly used significance level of 0.05 and even 0.1, there is not enough evidence against the null hypothesis (t-test: true difference in means is equal to 0; Mann-Whitney test: true location shift is equal to 0). Furthermore, we also notice that the true difference in means falls within the 95%CI range of [-0.375, 0.040] (Table 13). Therefore, we cannot accept H_8 and cannot state that positive eWOM on social media about a university's collaboration with other universities increases the likelihood for students to enroll in that university (Table 14).

To sum up, we accepted H_1, H_2, H_3, H_5, and H_6, whereas we did not accept H_4, H_7, and H_8. It should be noted that "not accepting" a hypothesis is not the same as "rejecting" it. We rejected the H_0 because there is strong evidence that contradicts it. However, we could only "not accept" H_4, H_7, and H_8 because there is insufficient evidence to accept them. By inspecting the interquartile range (IQR) of each group, we can also visually distinguish the experimental groups with accepted hypotheses (EG1: H_1, EG2: H_2, EG3: H_3, EG5: H_5, and EG6: H_6) from the ones with non-accepted hypotheses (EG4: H_4, EG7: H_7, EG8: H_8) (Figure 26).

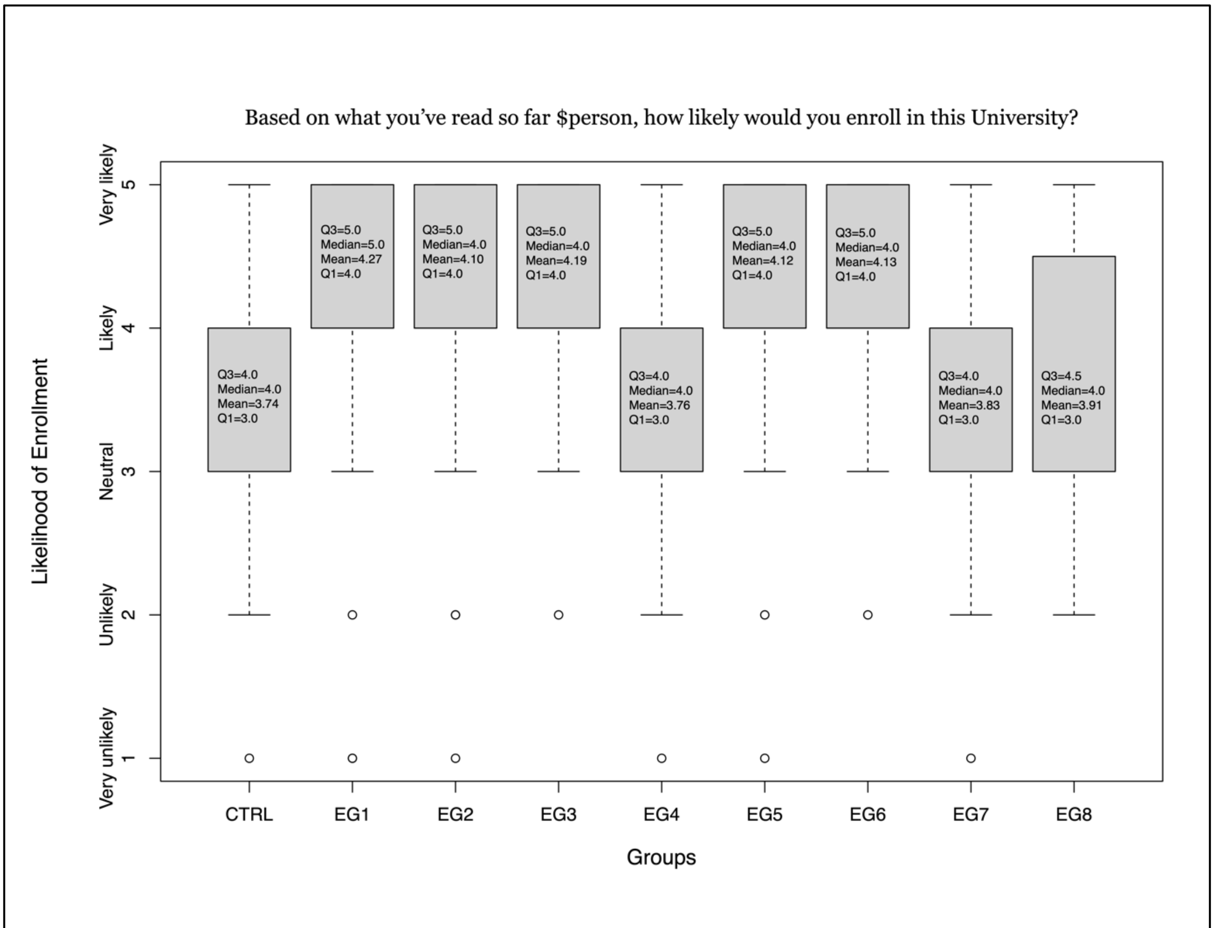

Figure 26: Likelihood of Enrolment by Control (CTRL) and Experimental Groups (Egs). IQR= [Q1:Q3]

Note. $person: name of the participant recorded at the beginning of the chat.

Supporting data

As explained before, during the chat with Sydn-E after the experiment response was collected, regardless of the group they were in, each participant provided eight structured responses to the questions relating to all eight decision factors being examined. Supporting the robustness and internal validity of each test, these responses bolster the results of the experiment. Descriptive statistics of these eight variables as well as the age of participants are shown in Table 15.

	Rep_rank	*Cost*	*Work_opp*	*Ease_admis*	*Campus_Loc*	*Course_attr*	*Know_city*	*Uni_collab*	*Age*
Mean	4.2003	2.8718	4.4602	4.0486	4.2548	4.3998	3.1987	2.9019	24.7475
Standard Error	0.0239	0.0255	0.0206	0.0259	0.0257	0.0204	0.0329	0.0306	0.0937
Median	4	3	5	4	4	5	3	3	25
Mode	4	3	5	4	5	5	3	3	23
SD	0.8249	0.8811	0.7120	0.8942	0.8883	0.7037	1.1359	1.0561	3.2333
Sample Variance	0.6805	0.7763	0.5070	0.7996	0.7890	0.4952	1.2902	1.1154	10.4542
Kurtosis	1.1754	0.4603	1.4267	-0.0957	1.3521	0.9578	-0.9027	-0.4673	-0.9896
Skewness	-1.0600	-0.1308	-1.2700	-0.6877	-1.2334	-1.0449	-0.0235	0.1964	0.0025
Range	4	4	4	4	4	4	4	4	17
Minimum	1	1	1	1	1	1	1	1	18
Maximum	5	5	5	5	5	5	5	5	35
Sum	5011	3426	5321	4830	5076	5249	3816	3462	29499
Count	1193	1193	1193	1193	1193	1193	1193	1193	1192

Table 15: Descriptive Statistics of the variables: Participant age and the importance of university's reputation, image and global ranking (Rep_rank), Preferred cost (Cost), work and internship placements during study, job opportunities and potential work-related benefits after graduation (Work_opp), ease of admission, considering its entrance requirements and open communication with admissions staff (Ease_admi), campus location as in its proximity to home, convenience and comfort, safety and physical appeal, and vibe of the city (Cam_Loc), flexibility and attractiveness of the course/program of study (in line with participant's career aspirations and earning potential) and on-campus support services (Cours_attr), whether participant has prior knowledge of the study destination/city (Know_city), and university's collaboration with other universities (Uni_collab).

As revealed in Table 15, "Uni_collab" and "Know_city" are the decision factors with the lowest Means (2.9 and 3.2 respectively), followed by "Ease_admis" (M=4.0). These factors are the only ones that were not accepted in our hypothesis testing. Whereas the accepted factors, "Rep_rank", "Work_opp", "Cam_Loc", and "Cours_attr" yielded significantly higher means such as 4.2, 4.5, 4.3 and 4.4 respectively. It's worth noting that the question regarding "Cost" inquired about participants' preferences for a university based on overall costs. It had a distinct structure compared to the other questions as the chatbot asked:

"What about living and study costs, availability of scholarships etc.? Which university would you prefer in terms of overall costs?"

The mean response score for this question on 5-point Likert scale (Highest priced=5; Lowest Priced=1) among all participants was 2.9, slightly below the "Average" cost rating.

Treating the collected Likert-scale data as ordinal, we compared the means across different levels of all Likert-scale variables using one-factor ANOVA. The output indicates that there is a statistically significant difference in participant responses based on their region of residence which impacted the variability in the data. For instance, where the mean "Rep_rank" in Africa is 4.62, it is 3.86 in USA and 3.92 in Canada (Table 16). However, we recorded a high level of consistency in variable means across regions as we found very strong correlations ranging from 0.89 to 0.99 (Table 17). This prompts us to run a cluster analysis to explore the nature of the structured data gathered by the AI.

	Africa	Asia	Australia/ NZ	Canada	Europe	South America	UK	USA	Overall
Mean Rep_rank	4.62	4.49	4.09	3.92	4.18	4.63	4.13	3.86	4.20
Mean Cost	3.13	3.38	2.94	2.78	2.93	3.16	2.82	2.44	2.87
Mean Work_opp	4.75	4.69	4.42	4.44	4.31	4.72	4.32	4.31	4.46
Mean Ease_admis	4.48	4.53	4.07	3.97	3.79	4.22	3.84	3.80	4.05
Mean Campus_Loc	4.12	4.22	4.37	4.29	4.16	4.50	4.33	4.28	4.25
Mean Course_attr	4.45	4.40	4.42	4.46	4.24	4.47	4.38	4.42	4.40
Mean Know_city	3.52	3.60	3.49	3.37	3.11	3.28	2.88	2.96	3.20

Mean Uni_collab	3.42	3.44	2.87	2.69	2.71	3.38	2.67	2.59	2.90
Overall	4.06	4.09	3.84	3.74	3.68	4.04	3.67	3.58	3.79

Table 16: Means of matriculation decision factors by the country/region of universities

	Africa	*Asia*	*Australia/NZ*	*Canada*	*Europe*	*South America*	*UK*	*USA*
Africa	1							
Asia	0.991	1						
Australia/NZ	0.904	0.924	1					
Canada	0.892	0.913	0.997	1				
Europe	0.922	0.937	0.974	0.966	1			
South America	0.954	0.954	0.930	0.917	0.970	1		
UK	0.907	0.925	0.968	0.963	0.992	0.973	1	
USA	0.904	0.917	0.988	0.989	0.972	0.950	0.981	1

Table 17: Cross-country correlations of the variable means

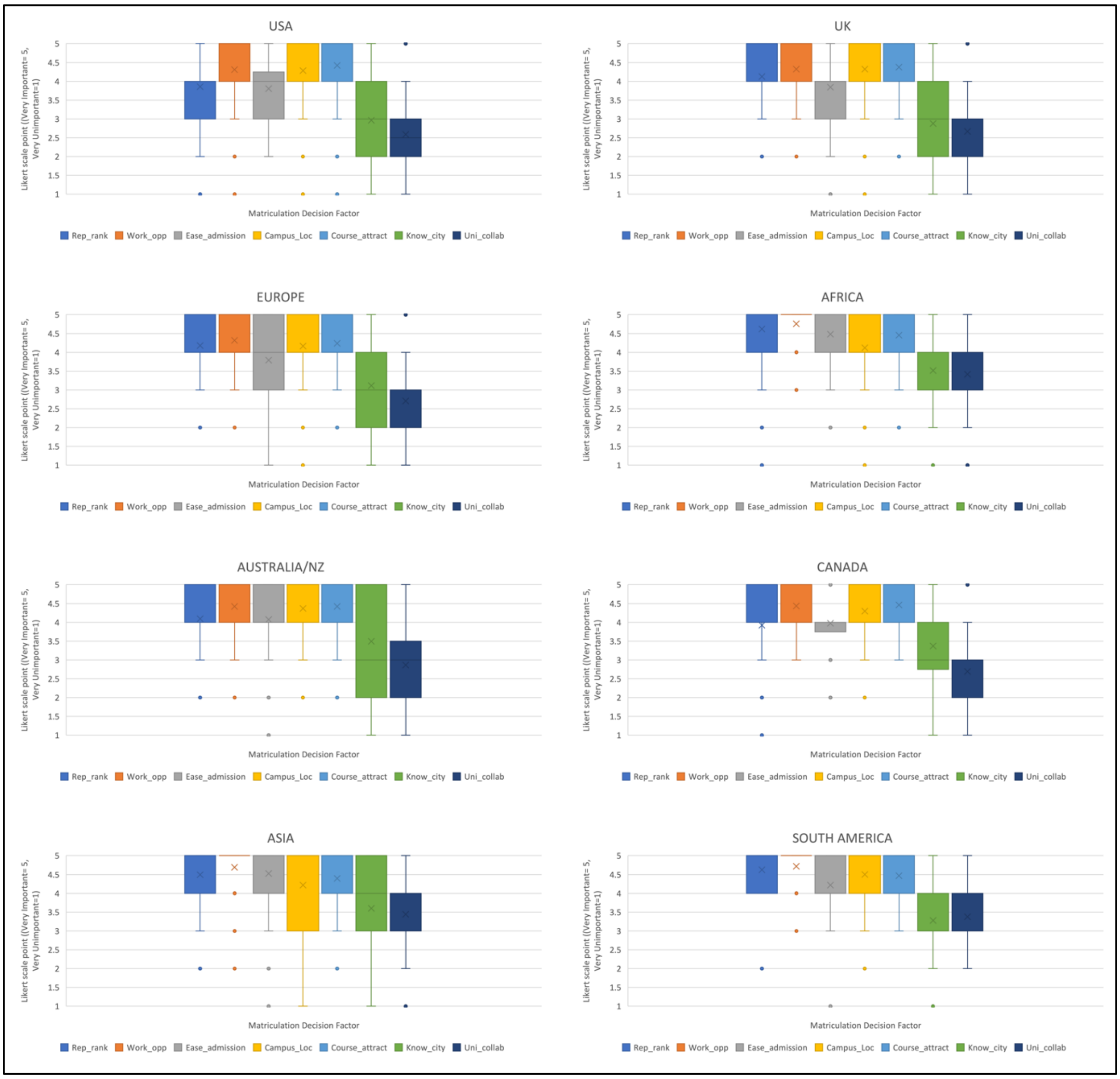

Figure 27: Matriculation decision factors on 5-point Likert scale (Very Important=5; Very Unimportant=1) by the country/region of the higher education institution attended or will be attended by participants.
—: Median; X: Mean, •: Outlier

Accordingly, we segmented the structured data into meaningful clusters using Python code to identify groups of participants with similar response patterns on the Likert-scale items and plotted the centrality of the points of matriculation decision factors by country/region of the HEI (Figure 27). We found that work opportunities, course attractiveness and campus location were the decision factors considered important or very important across all regions worldwide. University reputation and ranking was considered in 7 out of 8 regions important or very important. The only exception was the USA. Finally, where ease of admission was recorded important only in Africa, Australia/NZ, South America and Asia; knowledge of the city was considered less than important and collaboration with other universities was considered the least important decision factor across all regions.

6.5.3 Qualitative data insights

Content analysis

To identify themes out of participant responses in the form of unstructured text collected by Sydn-E from two of our open-ended questions[4] in the chat, we utilized Leximancer 5.0 which has been used in hundreds of studies since the early 2000s for conducting content analysis. Leximancer has the ability to mine, summarize, index, quantify and display conceptual structure of text (Smith & Humphreys, 2006), tag identified concepts to different sources (Noble et al., 2011), and refine a priori conceptual models (Lemon & Hayes, 2020).

[4] [1] Information on social media: "What other information about a university you read on $smp1 or $smp2 do you think would influence your choice to study there?" [2] Information on other sources: "What information, coming from other sources than social media, would you consider important when choosing a university?"

Figure 28: Wordcloud of all open-ended responses for “Information on Social Media”

Figure 29: Wordcloud of all open-ended responses for “Information on Other Sources”

Word clouds can be useful for identifying overarching themes and concepts in large texts by providing a quick and intuitive overview of the most frequently used words. Upon pre-processing all text, to have a general understanding of the frequent terms that may transform into prominent themes representative of the corpus, we built Word clouds based on the term frequencies of the responses pertaining to each open-ended question (Figure 28 & Figure 29).

Although we acquired some meaningful perspective into the overall content by visually clustering related words and frequently occurring terms that are central to the content, it should be noted that one major limitation of word clouds is that they may fail to capture the context in which the words appear or the relationships between them. Thus, it is vital to interpret the word cloud results in conjunction with a deeper analysis of the text so that a more comprehensive understanding of the themes and concepts can be gained.

Utilizing a combination of statistical and semantic techniques to analyse text content, Leximancer automatically detects relationships, patterns, and concepts within the data. This process involves several steps, such as tokenizing the text, removing common English stopwords, and identifying the frequency and co-occurrence of terms. Identified concepts and their interrelationships are visually represented within a concept map which displays nodes (signifying concepts) and links (signifying relationships between concepts). Accordingly, we let Leximancer use 2 sentence segments by default as the basic unit of content analysis as well as build a latent concept thesaurus automatically identifying and defining latent themes and concepts from the responses of each open-ended question in the chat. Once a term (initially specified as a seed word) is identified as being central to a line of discourse, it is augmented with other terms (directly or indirectly related to the seed word) from the text and along with the weightings it is then employed as an unsupervised classifier which codes text segments and leads to the formation of a cognitive (i.e., concept) map. Although most steps of this content analysis on Leximancer can be performed unsupervised, we still need to manually feed the classification engine with the dataset and make sense of the final thematic cognitive mapping of the model through its comprehensive ‘Topic Guide’.

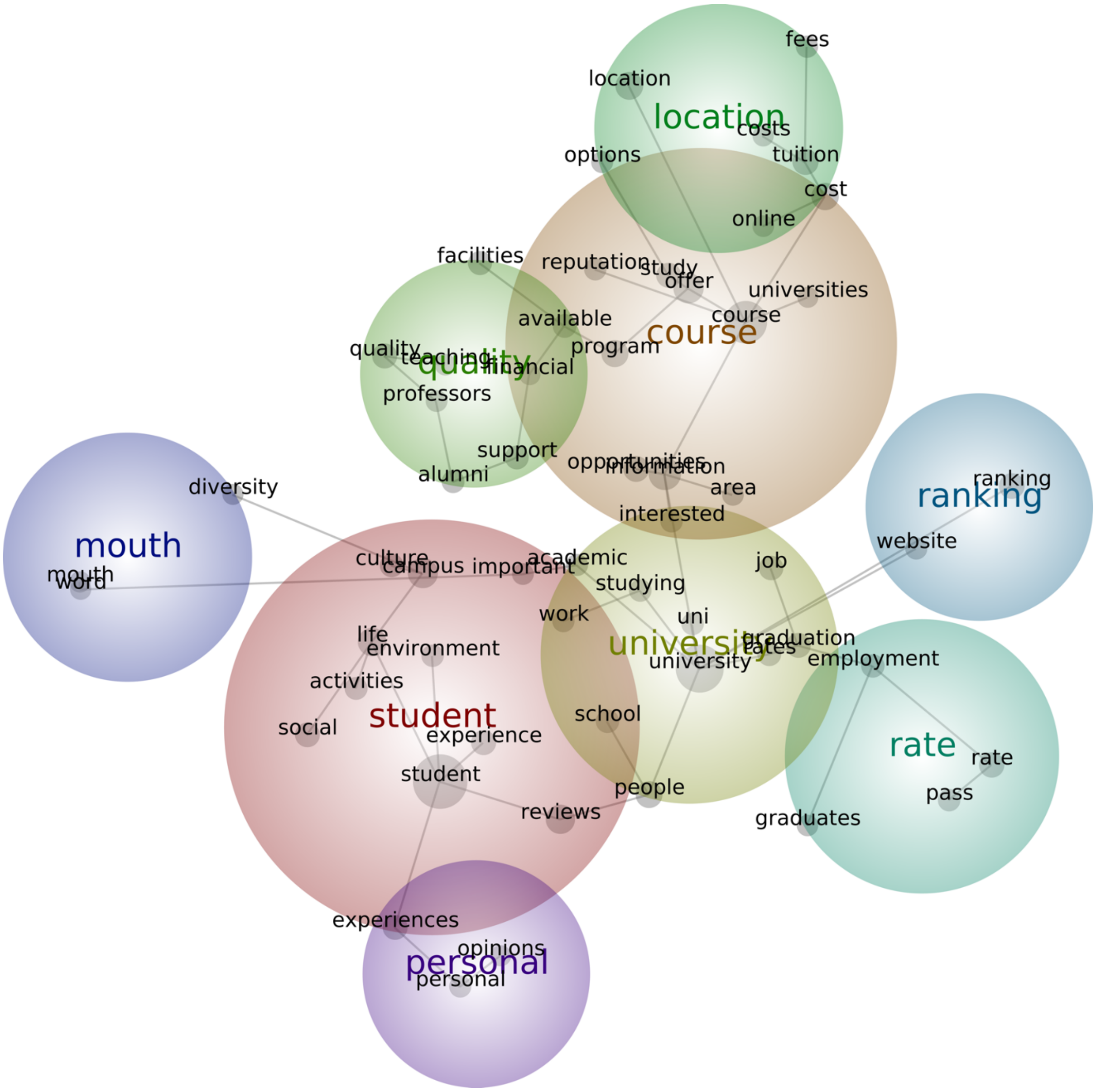

Figure 30: Concept map of the open-ended responses analysed with Leximancer 5.0

After importing our entire text data into Leximancer, using NLP and ML techniques Leximancer identified and extracted topics from the text as discussed above and generated a visualization representing the relationships between concepts and importance of each topic. This visualization is referred to as a concept map. (Figure 30). The *Topic Guide* feature of Leximancer allowed us to interact with the concepts of this visualization and explore the topics further. By clicking on a specific topic, we could access more detailed information including the associated terms, actual quotes (i.e., participant responses) in the chat, and the strength of the relationship between the terms and the identified topic. Thus, with the help of the *Topic Guide*, we could gain valuable insights into the prominent themes and concepts from the open-ended responses by manually examining the topics based on their terms' prevalence, relevance, and semantics in the overall content. For example, when we click on "Diversity" under "Environment Diversity" to access more information, Leximancer reveals the most frequently associated terms, such as "environment", "important", "financial", "support", "campus", and highlight the actual quotes that contribute the most to this topic (Figure 31). Accordingly, when we inspect all quotes relating to this topic, we note that diversity and inclusive environment are closely interrelated concepts that may influence students' matriculation decisions.

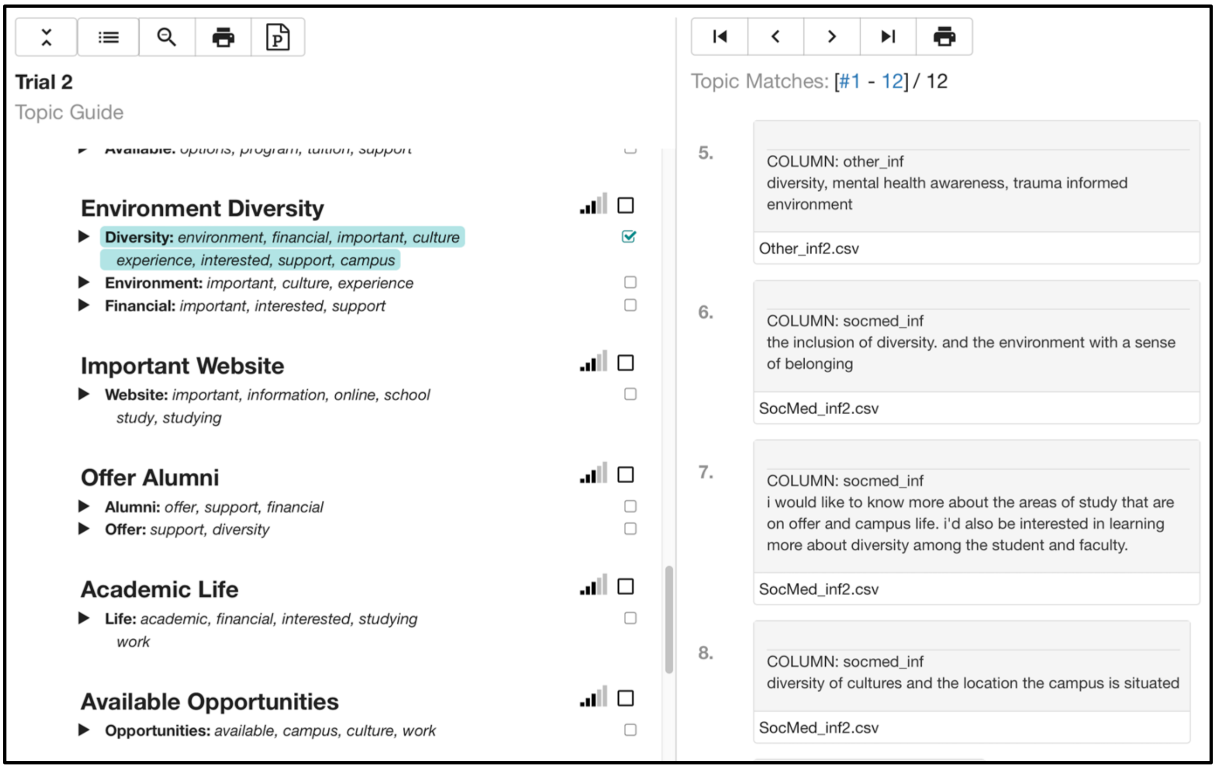

Figure 31: An example of the manual investigation process of the actual quotes (i.e., participant responses to open-ended questions) linked to the highlighted text of a concept (e.g., Diversity) on the Topic Guide of Leximancer 5.0.

Upon thorough manual investigation of the concept map and its associated *Topic Guide*, we actively sought out clusters of interconnected nodes and groups linked by relationships. This approach enabled us to successfully identify recurring themes and concepts within the text. After a laborious and lengthy manual examination of the entire text collected from 1193 participants, we have verified that the following themes hold substantial potential to impact students' decisions regarding their choice of university: 1) Student experiences and reviews, word of mouth and personal opinions; 2) Job opportunities available in the field of the offered course; 3) Quality of teaching, professors, alumni and available support; 4) Ranking and reputation of the university, 5) Course information on websites and study options, 6) Pass rates, and employment rates of graduates; 7) Location, social life, campus environment, facilities, and 8) Financial support, tuition and fees, and other costs. In addition, we found that other overarching concepts such as diversity and inclusive environment as well as alumni success hold significant importance and should also be duly considered.

Rapport, appreciation and rich qualitative insights

An exciting outcome of the experiment was that we found compelling evidence pointing to the collection of rich qualitative insights through participants' responses to open-ended questions, attributed to the rapport established by the AI.

As discussed before, every eligible participant recruited via Prolific was given a "Completion code" by the AI at the end of the chat. This code took them back to Prolific and upon the AI's confirmation, one human researcher approved these submissions on Prolific so that participants could receive their money. However, the AI asked one more open-ended question which came right after the release of the *Completion Code*. This question was: "Would you like to add anything else before we say our goodbyes?" (Add anything else). Although the majority of the participants responded by saying "No", "Nope, thanks", "bye", etc.; many of them provided a variety of warm expressions (e.g., "nice talking to you", "it was cool participating", "this has been awesome", "I had fun answering the survey", "It's been really fun", etc.) as if they were concluding not a traditional survey but a genuine interaction with a real human. Although they knew they were talking to a chatbot, they still appreciated the method and expressed their true feeling about the method (e.g., "You did a good job Sydn-E :)", "Great chat bot!", "I liked this method! great job", "this was a really fun and Interactive way of completing a survey, and I'd love to see more of this in the future!", "Great AI tool, I love the way it calls me by my name, it makes it feel personalized in a way", etc.). Even one participant took his time and wrote: "I would like to say that this was an interesting study to do as I have never talk to a chatbot before and I think this is a very cool idea that I hope continues moving forward" (Table 18). But more importantly, the fascinating part was that none of the participants had to say any of this because they all had already received the code to get paid. Nevertheless, many participants kept the conversation going and ended the chat as if they were talking to another human. This wouldn't have happened unless the AI had built a strong rapport with the participants.

In addition, a manual examination of all responses to three open-ended questions (Table 18) revealed that participants who provided additional information or simply expressed their opinions or feelings in the last question (i.e., Add anything else) also provided rich insights in the midchat open-ended questions (i.e., [1] Information on social media, [2] Information on other sources). This also substantiates the main proposition that the AI effectively fostered a strong rapport with participants, resulting in their candid expression of opinions. More importantly, building such rapport helped the AI garner more insightful, authentic, accurate and detailed information improving not only the validity but also the reliability of the big qualitative data obtained (Table 18).

Information on social media	Information on other sources	Add anything else
The previous description pretty much more than covered everything I'd like to hear, but in general just being progressive, actually caring	Stories from other students there or graduates, campus talk, lesser known facts about certain things like rituals, groups, teachers, etc.	No, everything was fine

and listening to the students, being up-to-date with the world in general & in touch with everything that's going on is just important.		
A good amount of scholarships to cover tuition	General information like how good education is, how expensive it is, and location.	No. Thank you for this interesting study.
I would like to see other peoples experiences with it who went there.	Peoples reviews of it, their experiences their, the location of the university, their rules, etc.	thx for chatting ☺
Information about the tuition cost, opinions about the university from current students, and information about the city where the university is located.	Tuition cost, location of the university, flexibility of courses (e.g., online/hybrid options, day/night schedules), and accessibility for people with disabilities.	Thanks for the opportunity. It was nice to meet you.
Distance to facilities, local environment, local safety (is it in a safe city area?), public transport options, and most importantly for me right now, the ability to study remotely or online if needed	Pricing, location, culture (is it welcoming? What are its values?), safety of area and campus, campus security, online/remote facilities and capabilities, on-site facilities such as food stores and game rooms, proximity to services and transport	Goodbye Sydn-E!
How highly ranked the uni is in my chosen subject; regionally, nationally and globally	Articles written on successful alumni	This was an interesting study!
What types of social activities are on offer, how good the Students Union is, whether or not the lecturers are (a) well-treated by the College and (b) whether the lecturers treat the staff well	University ranking in world ranking tables, overall prestige of the university	No thanks, nice talking to 'you' !
Seeing their teaching approach, resources and connections, and also if they have an art program.	Personally talking with some of the teachers and getting a tour of their workplace.	It was cool participating.
Things like that social status and level of education for the subject I plan to study	Financial information: tuition, cost of living, accommodations, etc...	It was nice chatting with you
I would be unlikely to allow social media to influence my choice, but I'd probably lean towards using it to gauge the social experience.	In person conversations with alumni and college tours.	Interesting study, good work to the students/professors who were involved.
Location of campus and proximity to nearby amenities.	Recommendations and reviews from peers, friends and family	Very responsive chat
It's prestige and ranking as well as the quality and extensive ness of resources available for me to use	University rankings, articles or expert reviews	If its of any interest, I go to Oxford University
friendliness, good outcomes, and good nightlife	the uni website	I liked this method! Great job
I'd like to know what others say directly in comments or if I know anyone there or if they know anyone attending there	Ratings, reviews from internet forums, google reviews on Google Maps etc	I'm good, thanks. I'm glad to help in your research and testing
How the university is viewed by unrelated third parties, what the acceptance rate and other statistics of those accepted are, where the university is, and what the campus is like.	Ratings on websites that rate how esteemed a university is, other information about how the university is viewed compared to other universities.	Great chat bot!
the courses that are offered and the overall happiness of the students and faculty	safety, school ratings, and graduation rates	it was a really unique study, I had fun!
I think hearing opinions from students would better inform me about what the school is actually like	Who teaches at the school, the school's ranking, how many students attend, what the campus is like	Thanks for talking ☺
I like that the current description shows a care for work-life balance, but I'd like to know more about the amenities on campus too.	Opportunities available, support for alumni post-graduation, intramural and extracurricular activities	It was nice talking to you, take care!
Information given by the students would influence me. They could talk about how the classes and teachers are or even how the dorms look.	How previous students cope outside of the university	It was nice talking to you
Courses offered and university fees	Graduates who studied at the university would be able to give an honest opinion about the university which will assist me in my decision	No. Thank you. This has been awesome
Real life experiences, not just posted by the uni but if current and past students are talking favourably without promt in the comments and on reviews	Idk, just how good are the teachers and how they interact with the students, especially for disabled and 197odelling197g ill people	I just hope that who ever reads this is having a lovely day, I hope you see a happy dog or a really beautiful flower ☺
How inclusive the university is, what programs they offer, the cost, and if they offer online courses or degrees	Ratings and reviews	It was nice talking to you!
Real honest reviews and feedback from people who have attended the university and more information about the programs they offer	Where they're working now after attending the university. More information about the professors that teach the programs	I had fun answering the survey. Thanks
if they had current students sharing their stories of being a student on social media	how well the university's website is made and how clear and easily accessible the information for future students is	this was a really fun and interactive way of completing a survey, and I'd love to see more of this in the future!
Information about graduation rates and average GPAs.	The university's location, prices of tuition, and classes offered.	Have a good day, chatbot!
I think reading about the happiness of students with both their academics and campus enjoyment would influence my decision of going there. Also pictures and testimonies of those who excel there with the opportunities that are given.	The community response and opinion of the university	thanks for chatting with me, I had fun!
Comments about the quality of high-level faculty members, ratings of the professors, comments made by former students about their experience with courses at the university	Information from anyone I know who went to that university, advertising booths for the university, online information about scholarships offered there	You did a good job Sydn-E ☺
If I live close to the university and how good the learning material is	ex students and websites that rank universities	I would like to say that this was an interesting study to do as I have never talk to a chatbot before and I think this is a very cool idea that I hope continues moving forward
If prices were low to be able to study there. Or if allowances were made where I could work and study and not feel too stressed about having that balance.	Location and reputation of the University. It'd really help to know from someone who's experienced the life from that specific University.	Nope. I enjoyed chatting!
I would want to speak to real students and get a realistic view of the college. I want an unfiltered view so if there are downsides Id like to hear those too.	Id want to look at the area the university is located in.	thank you Sydn-E have a nice day!
Information on placement opportunities, details of the campus	Location of the university, cost of living in said area	This method of taking a study is great, preferable over other types.
Some more detail regarding the non-academic "outlets" available at the university	Student satisfaction rates, information regarding facilities, fees, courses that are available	No, the study worked very well, questions were easy to understand and answer. Thank you for you time ☺
Information about the city, public transport, things to do in town, rent prices,# etc.	I don't really trust social media, so I would search other places	What AI model do you use?
The rankings of the universitys education qulity when compared to other universities within the continent	Whether students complaints and requests are treated as a priority, in order to ensure that the overall experience in my academic years are pleasant	What I built is amazing, I like the interface more than anything.
The kind of social chapters and extra-mural activities and networking groups available on campus	Their tuition fees	You're a good AI bot
If I didn't live in the city I'd like to know if the dorms on campus are any good	the extra curricular activities the university might have	Great AI tool, I love the way it calls me by my name, it makes it feel personalized in a way
A university that gives students free health care facilities and also provide students with psychologists.	A university that promotes physical health and have fitness facilities.	I enjoyed the study.
Information on how the university assist students that are in mental distress due to course work	the culture of the university.	It was nive meeting you Sydn-E...You are awesome!
Their CSI projects, what they are doing to improve communities	Enrolment process, courses offered	I enjoyed participating in the survey

It provides students with funding, and allows the students to get the feel of how it feels when you are in the field.	You must consider the primary language of communication	It was fun chatting with you
the fact that they consider career guidance since most people from high school actually have not clue what they would like to study	their requirement for each field of study and its career path	great survey
How good the professors are at their job and their credentials and cost of tuition	The website and first hand experience such as friends that study there	It was nice talking to you
The learning environment, the amount of graduates they produce in a year and generally how well the students do at the university	Whether the university is inclusive and if they offer my programmed of choice. I prefer institutions that encourage diversity	Nothing more from here. It was a pleasure chatting to you.
Newsletter features, articles about exceptional students, testimonies from post graduate students.	University published newsletters.	Pleasure chatting with you.
That they encourage hardwork academically but also understand the importance of fun through sports. Also the opportunity to study in other universities around the world and experiencing different cultures	Student safety should be prioritized. Elite postgraduate opportunities within the university and an emphasis on mental health and mental awareness	It's been really fun. Good bye
The location and accommodation offered near it.	The reputation of the university.	It was a pleasure meeting you and talking to you.
The level of diversity and inclusivity on campus, as well as the university's commitment to promoting an inclusive learning environment, may be important considerations for students seeking a diverse and multicultural experience.	The expertise, qualifications, and research accomplishments of the faculty members can play a crucial role in attracting students. Opportunities for research, internships, and collaborations can also be appealing	What is the focus of your research?
That they reward you when passing well, like getting excellence certificates, etc.	How many students do they have and pass the courses.	No thank you. It was nice chatting with you.
How highly the university is ranked. Proximity to my own home. Reviews from students/former students. How well reviewed the course is that I wish to study	Location, How easy it is to get too, staff experience, nice campus	Very innovative questionnaire
first hand accounts about their experiences at the university	campus life, food options	good bot

Table 18: 50 rich interactions for 3 open-ended questions: [1] Information on social media: "What other information about a university you read on $smp1 or $smp2 do you think would influence your choice to study there?" [2] Information on other sources: "What information, coming from other sources than social media, would you consider important when choosing a university?", and [3] Add anything else: "Would you like to add anything else before we say our goodbyes?"

6.5.4 Combined insights

As displayed in the second section of Figure 1, after analysing the structured quantitative data collected by the AI from participants, we made causal inferences that revealed the influence of five out of the eight tested factors on matriculation decisions. We should note that in addition to the five proven factors, positive electronic word-of-mouth (eWOM) had a vital role in participants' choices, however it should not be considered an actual decision factor but a channel through which messages pertaining to other matriculation decision factors shall be conveyed. These five factors as well as positive eWOM were fully supported by the qualitative results (Table 19).

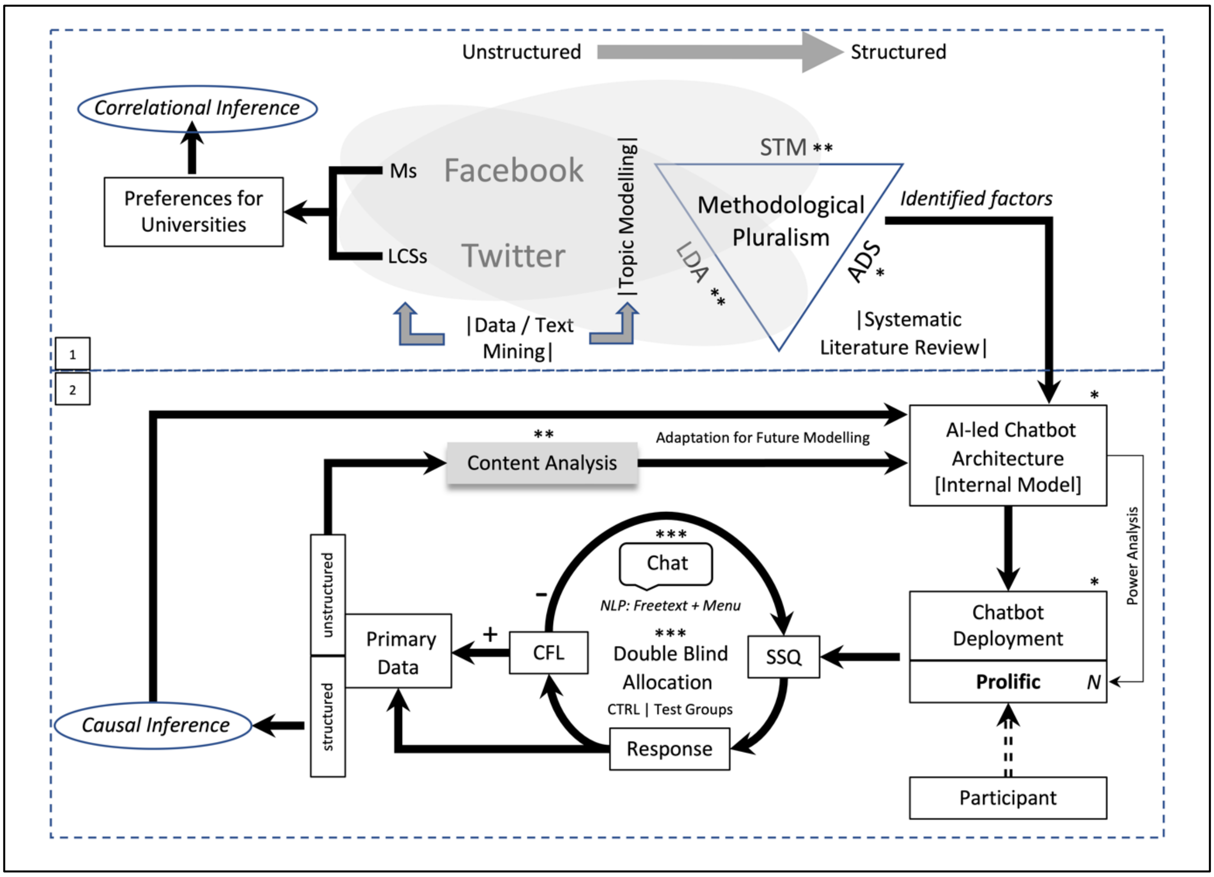

Figure 1. Flowchart of all processes including data collection from Facebook and Twitter, establishing correlational inference, topic identification via topic models (LDA and STM) and systematic literature review (ADS), initial training of AI-led chatbot (AILC) architecture, double-blind participant allocation, attaining causal inference through structured primary data and subsequent training of AILC with updated factors based on the initial experiment's structured and unstructured primary data.
Key to Figure 1. Ms: Mentions of university names on Twitter; LCSs: Likes, Comments and Shares of Facebook posts; LDA: Latent Dirichlet Allocation; STM: Structural Topic Modelling; ADS: Algorithmic Document Sequencing; CTRL: Control Group; SSQ: Semi-structured Questions; CFL: Confirmatory Feedback Loop; NLP: Natural Language Processing; *N*: Total number of participants being recruited from Prolific.
*: Supervised; **: Semi-supervised; ***: Unsupervised

After examining the content of the unstructured qualitative data, we identified ten factors, none of which were *unaccepted* during hypothesis testing. Moreover, some of them blend in and some exactly match the results obtained with the experiment (Table 19). We also notice that the results of the content analysis helped us not only finetune but also complement those of the AI-led experiment, leading to clearer, richer, and more comprehensive combined results.

When we combine the quantitative results of the experiment with the qualitative results of the content analysis, we identify 9 matriculation decision factors in total (Table 19). However, it should be noted that we can only attest to the impact of the factors identified and proven via the AI-led experiment on such decisions. These factors are: (1) Student experiences and reviews, word of mouth and personal opinions; (2) Work and internship placements during study and job opportunities available in the field of the offered course; (3) Availability, flexibility and attractiveness of the course, quality of teaching, professors, alumni, and available support; (4) University's reputation, image, and ranking; (5) Campus location including proximity to home, convenience and comfort, safety, physical appeal, social life, campus environment, facilities, and vibe of the city; and (6) Living and study costs, availability of scholarships and other financial support. We also found that three other factors (i.e., (1) Pass rates, and employment rates of graduates; (2) Diversity and inclusive environment; and (2) Alumni success) which were identified through the content analysis in addition to the factors identified in the experiment may also impact the university choices of students.

Content Analysis	Experiment (Hypothesis testing)	Combined Results
Student experiences and reviews, word of mouth and personal opinions	Positive eWOM*	Student experiences and reviews, word of mouth and personal opinions
Job opportunities available in the field of the offered course	Work and internship placements during study and job prospects upon graduation.	Work and internship placements during study and job opportunities available in the field of the offered course
Quality of teaching, professors, alumni and available support Course information on websites and study options	Availability, flexibility and attractiveness of the course and on-campus support services.	Availability, flexibility and attractiveness of the course, quality of teaching, professors, alumni, and available support
Ranking and reputation of the university	University's reputation, image, and ranking	University's reputation, image, and ranking
Pass rates, and employment rates of graduates	NA	Pass rates, and employment rates of graduates
Location, social life, campus environment, and facilities	Campus location including proximity to home, convenience and comfort, safety, physical appeal, and vibe of the city.	Campus location including proximity to home, convenience and comfort, safety, physical appeal, social life, campus environment, facilities, and vibe of the city.
Financial support, tuition and fees, and other costs	Living and study costs, availability of scholarships and *access to technology & research**, and facilities****.	Living and study costs, availability of scholarships and other financial support
Diversity and inclusive environment	NA	Diversity and inclusive environment
Alumni success	NA	Alumni success

Table 19: Matriculation decision factors identified with Content analysis through unstructured qualitative data, factors proven to have impacted matriculation decisions with the AI-led Experiment (hypothesis testing) through structured quantitative data and the combined results incorporating the union of the results obtained from both analyses.
* Positive electronic word-of-mouth (eWOM) is not an actual factor but a channel through which the AI conveyed messages pertaining to other matriculation decision factors.
** We decided not to include "access to technology and research" in the combined results because the intervention question of the corresponding EG did not include information about accessing to technology & research. Furthermore, there was no mention of accessing to technology & research in any of the open-ended participant responses.
*** Facilities blend in the previous factor (i.e., campus location).

Part 4

Discussion and implications

Matriculation decisions

In an era where information is abundant and readily accessible, external information sources, encompassing a multitude of facets influencing university choice, profoundly impact the enrolment decisions of prospective students. These sources provide a panoramic view of universities, beyond the glossy brochures and official promotional materials. They empower students to make informed choices, aligning their academic and personal aspirations with the institutions that best suit their needs and ambitions.

The decision to select a university is a pivotal moment in the academic journey of prospective students, and it is profoundly shaped by a variety of inputs originating from a variety of sources. First and foremost, we found that opinions and experiences shared by current and past students provide a candid window into the university's value proposition. These unfiltered narratives encompass a spectrum of sentiments, ranging from the overall happiness of students and faculty to satisfaction with academics, campus enjoyment, and the learning environment. Real-life experiences and testimonials from students contribute to a nuanced understanding of the university's atmosphere. Feedback and reviews from students about specific courses and professors provide granular insights into the academic experience, enabling prospective students to tailor their choices to their educational preferences. Beyond the gleaming I, prospective students are increasingly seeking unfiltered views and insights into the downsides of universities. These candid appraisals, sourced from various channels, ensure that students make well-rounded decisions that align with their academic and personal aspirations. We should also note that the physical infrastructure of the university plays an

Important role In shaping students' Ies. The amenities and resources available on campus, proximity to nearby facilities, accommodation options, including dormitories, and the availability of healthcare facilities and support services significantly contribute to the overall appeal of the institution.

The post-graduate landscape assumes paramount importance for prospective students (Le et al., 2020). Accordingly, critical factors in the enrolment decision-making process may include information concerning the university's provision of career services (Hemsley-Brown & Oplatka, 2015), post-graduate employment rates, the nature of employers recruiting from the institution, and the average graduate salary (Delavande & Zafar, 2019). Moreover, insights into placement opportunities and details of the campus may reinforce the attractiveness of a university as a launchpad for a successful career. Equally significant are the opportunities for internships or work experiences, which provide hands-on exposure and a tangible pathway to professional development.

The regional, national, and global standing of a university constitutes a salient consideration for prospective students, as evidenced by institutional rankings and the recognition of prestige (Bowman & Bastedo, 2009; Horstschräer, 2012). The calibre of faculty members and their scholarly reputation significantly contributes to the overall academic milieu (Hemsley-Brown & Oplatka, 2015; Dearden et al., 2019). Prospective students judiciously evaluate the educational quality of the university in comparison to other institutions, aspiring to align their academic pursuits with the highest standards and considering the global context in which the university is situated. Our data revealed that the view of the university by third parties, including recent controversies or political issues related to the institution, provides an external perspective that impacts enrolment choices. It adds an extra layer of consideration for students, who evaluate the university's standing in the broader societal context.

The university's commitment to promoting an Inclusive learning environment Is paramount. Information about accessibility for disabled students, inclusivity, diversity on campus, and programs and resources offered to support students with mental distress profoundly influences students' choices. The diversity of the student population contributes to a vibrant and inclusive campus culture, which is a significant factor in enrolment decisions.

Financial considerations are pivotal in the university selection process. Financial considerations, encompassing tuition costs, fees, availability of scholarships, and overall affordability play vital roles in the university selection process (Hemsley-Brown & Oplatka, 2015). Additionally, opportunities to work while studying may be sought after by prospective and current students to manage the financial aspects of higher education.

A holistic view of university life Is essential for prospective students. This encompasses social activities and events organized by the university, the reputation and quality of the Students Union, the availability of social chapters, extramural activities, and networking groups. Equally, the presence of non-academic outlets and opportunities for sports and recreation factor into students' decisions, fostering personal growth beyond academics.

The geographical location of th' 203odelling203y emerges also as another fundamental consideration, as prospective students may consider its proximity to home (Briggs, 2006), the safety of both the city and campus (Calitz et al., 2020), the availability of local amenities, and the vibrancy of the town and its surroundings (Eder et al., 2010) with cultural and historical facets contributing to the overall appeal of the university (Agrey & Lampadan, 2014).

Universities are also expected to provide comprehensive support services. Career guidance and support for undecided students, efficient mechanisms to address student inquiries or concerns, access to mental health support, psychologists, and counselling services, as well as community engagement and improvement projects, underscore the university's commitment to student well-being and growth.

In contrast to earlier findings in existing literature, this study challenges established notions regarding the determinants of students' university choices by examining three specific factors: the ease and flexibility of admission in Higher Education Institutions (HEIs) (McFadden et al., 2015; Massoud & Ayoubi, 2019; McLeay et al. (2020), students' familiarity with the study destination (Mazzarol & Soutar, 2002; Lee, 2014; McLeay et al. (2020), and the collaborative engagements of HEIs with other institutions (Hemsley-Brown & Oplatka, 2015; Oladipo & Sugandi (2021). While prior research posited that these factors significantly influence students' decisions in university selection, our experiment refutes such assertions, demonstrating a lack of discernible impact. Notably, this study distinguishes itself by employing a double-blind randomized true experiment, acknowledged as the Gold Standard for establishing causation in this domain, marking a departure from conventional research methodologies. Consequently, the outcomes of this investigation prompt a re-evaluation of the aforementioned factors within the scholarly discourse, as they should be expunged from the pool of factors influencing matriculation decisions in higher education.

Exploring students' university choice factors by harnessing the power of AI has the potential to considerably contribute to our understanding of various aspects of higher education so long as we first and foremost adhere to the principles and guidelines set forth by the Committee on Publication Ethics (COPE) for the utilization of machine interviews in research and GDPR for the privacy and protection of the participants' data. In doing so, this study went through a meticulous scrutiny by the ethics board of our higher education institution before getting approved. In the pursuit of scientific integrity and ethical research conduct, we diligently follow and advocate COPE's recommendations to ensure transparency, fairness, and the responsible application of machine interviews as a research tool.

In a competitive higher education landscape, by understanding what attracts students, HEIs can effectively position themselves and build a strong brand that resonates with their target audience (Lomer et al., 2018) Gaining insights into the reasons behind students' preferences on a global scale can empower policymakers to make informed decisions to attract and retain not only domestic but also international students. By aligning policies with students' preferences and needs, they can ensure that funding, resource allocation, and program development are better tailored to meet the demands of both current and future students. Secondly, understanding what factors influence students' decisions to enrol in a specific university can help HEIs improve recruitment strategies and develop effective retention programs. By catering to the preferences and expectations of their students, institutions can enhance their overall performance and deliver better outcomes for both students and the institution itself. In addition, HEIs can modify their curriculum and teaching methodologies in accordance with the factors that matter most to students, resulting in more engaging and relevant educational experiences. This student-centric approach can lead to improved learning outcomes and a more satisfying educational journey.

Beyond the campus, understanding these choice factors can also shed light on the broader societal and economic implications of higher education. It can aid in addressing issues relevant to diversity and inclusion by identifying and eliminating barriers that may deter certain groups from pursuing higher education. Accordingly, HEIs can strengthen their ties with the community and contribute positively to society through partnerships and outreach programs informed by this knowledge. Economically, this understanding can lead to better alignment between higher education and the job market. Understanding students' aspirations and the factors shaping their choices may enable institutions to design programs that may not only amplify their graduates' employability but also actively contribute to economic development and the cultivation of a highly skilled workforce. Lastly, improving support services and ensuring student satisfaction and success are also critical implications of this exploration. By tailoring support services to meet students' needs and preferences, HEIs can boost the overall university experience, leading to greater satisfaction, higher success, and lower dropout rates.

Taking a leap from traditional human-human RCTs

A goal-oriented adaptive AI system such as Sydn-E can substantially alleviate cost and resource limitations in conventional human-human RCTs by automating tasks, scaling up tasks, and streamlining data collection and analysis. Such AI-run experiments reduce the need for extensive human intervention and labor, offering efficient, cost-effective and quicker data collection through interviewing and experimentation, and improved data quality. AI's adaptability and ability to replicate experiments consistently enhance the overall efficiency and reliability of research. This can allow for real-time monitoring of participants' responses, immediate feedback, and adaptive adjustments to the experiment's parameters, further improving the overall efficiency of data collection and analysis. Additionally, AI algorithms can uncover hidden patterns and insights within the data, contributing to a deeper understanding of the phenomena under investigation, all while minimizing the time and resource investments typically required in traditional RCTs.

AI-conducted experiments can – as demonstrated in this study – address statistical power limitations in traditional RCTs by leveraging the ability to work with larger and more diverse sample sizes. AI's scalability allows for the engagement of a significantly higher number of participants, enhancing the statistical power of the study to detect even subtle effect sizes or differences that might be missed in smaller RCTs. Furthermore, continuous data collection facilitated by AI contributes to stronger statistical analyses by reducing measurement error and allowing for real-time trend and pattern detection. AI can also offer adaptive experimental design, dynamically adjusting parameters based on ongoing data analysis to optimize the allocation of resources, thereby further increasing statistical power. The efficiency of AI-driven data analysis and the ability to automate this process may enable researchers to analyze vast datasets, improving the study's power to detect meaningful effects while saving time and resources. Additionally, AI's subgroup analysis capabilities can uncover variations in treatment effects among different populations, potentially revealing insights that may be overlooked in smaller RCTs.

AI-conducted experiments can also augment statistical power through improved data quality. AI-driven data collection and analysis reduce measurement errors and ensure data accuracy, leading to more precise and reliable statistical estimates. Replicating experiments multiple times with high precision, a capability of AI, also contributes to the reliability and robustness of the findings, ultimately increasing statistical power. AI's time efficiency accelerates the experimentation process, leading to faster data collection and analysis. This is particularly valuable for time-sensitive research questions, as quicker decisions and faster results can lead to improved statistical power. While AI's potential to overcome statistical power limitations is significant, it's crucial to emphasize that proper experimental design, careful consideration of confounding variables, and the elimination of potential biases remain essential to ensure that the207odellined statistical power translates into meaningful and valid findings. Additionally, the interpretation of results should be done with care, as larger sample sizes can lead to the detection of statistically significant effects that may not always be practically significant.

Experiments run by the AI can offer valuable means to address human biases in RCTs. Firstly, AI algorithms can automate the randomization and allocation of participants to treatment and control groups, eliminating the potential for selection bias that human researchers might introduce inadvertently. This impartial process ensures that not only the group assignments but also the allocation of interventions to groups is unbiased. AI can also play a pivotal role in preserving blinding protocols, ensuring that neither participants nor researchers are aware of their group assignments, thus reducing observer and participant biases. Such AI systems can also maintain consistency in data collection, reducing the potential for data collection biases that may arise when human researchers interpret or record data differently. Additionally, by automating data analysis, AI can identify patterns and relationships in the data objectively, minimizing confirmation bias that human researchers might introduce by seeking out data that aligns with their expectations.

The AI-based interview system in this study ensures a standardized and impartial interaction between the interviewer and interviewee, thereby minimizing the inherent subjectivity associated with human-led interviews. By relying on predefined criteria, it can eradicate the potential for subtle biases, preconceived notions, or unintentional cues that may influence traditional human interviewers. The systematic and consistent nature of AI-driven engagement guarantees that all participants are evaluated on identical parameters, fostering fairness and objectivity in the evaluation process. Consequently, the adoption of AI in interviews not only enhances the reliability of research outcomes but also contributes to the establishment of a more equitable and unbiased research environment, aligning with the principles of scientific rigour and objectivity.

Although AI systems are not influenced by experimenter biases in terms of interviewer effect, it is essential to acknowledge that they are not entirely free from biases, as they can inherit biases from their training data or algorithms. Therefore, careful design and oversight are crucial to ensure that AI is trained and implemented in a way that minimizes bias. Moreover, while AI can reduce certain forms of human bias, human researchers still play a pivotal role in setting the parameters, objectives, and ethical guidelines for AI-conducted experiments. The combination of AI and human oversight is critical to ensure the ethical and unbiased conduct of experiments.

AI can autonomously execute ethically sensitive decisions, such as withholding treatment from control groups, ensuring these decisions are carried out impartially. AI technology can prioritize data privacy and confidentiality, addressing concerns about the protection of sensitive participant information. However, it is necessary to design AI algorithms and systems with ethics in mind and to uphold ethical principles during their development and use. While AI plays a crucial role in addressing ethical concerns, human researchers and ethicists remain essential in setting ethical guidelines and ensuring AI technology aligns with these principles and respects participants' rights and well-being. The collaborative effort between AI technology and human oversight is vital for conducting ethically sound experiments.

AI-driven research methodology

One of my primary objectives in this study is to propel the field of AI-driven research methodologies, forging a path towards demonstrating AI's potential in gathering different forms and levels of data from multiple sources in an efficient, timely and rigorous manner. I aim to provide robust evidence through a double-blind RCT – free of human interference hence bias, which enables the establishment of causal relationships between interventions and their outcomes. In doing so, I seek to bridge the worlds of technology and human perception, creating a symbiotic relationship where AI-driven data collection and experiments find widespread acceptance and deliver benefits across all sectors seeking value in human opinions and experiences.

The methodological contributions in this study have far-reaching applicability across various disciplines and contexts. This research significantly elevates our understanding of data collection using AI, as it leverages AI for conducting double-blind true experiments whilst triangulating results with qualitative and quantitative data insights drawn directly from participants through interview-like chatbot surveys. This innovation has the potential to empower researchers from diverse fields, enabling them to gather data from substantial sample sizes and generate statistically reproducible, reliable, and broadly generalizable results.

In the era of advancing AI technologies, we stand at a pivotal juncture, ready to foster a harmonious coexistence between AI and human elements. Together, they hold the key to addressing challenges and exploring the limitless possibilities that emerge not only in the interactions where humans engage with AI but also where AI engages with humans. This synthesis embodies the essence of mixed methods research, as it harmonizes technological innovation with a deep understanding of human factors, offering a holistic approach to research and discovery.

Recommendations for future research

We are recently witnessing some unprecedented changes and breakthroughs with AI particularly the generative kind like OpenAI's GPT. Although the AI that I developed (Sydn-E) for this thesis collected quantitative and qualitative data from participants and ran the experiment without human (i.e., researcher) intervention, the results were still analysed and interpreted by me. It was also I (a human) who developed the internal model of the AI. I believe that such results can soon be analysed and interpreted by the AI itself. I also believe that the internal models of AI can be self-developed or by other AI with API connection (e.g., ChatbotNet: Chatbots collaborating with one another). So, I recommend AI researchers to build upon AI models like Sydn-E, integrate them to generative AI models like ChatGPT, and develop self-sufficient adaptive AI which can not only collect data, run experiments, analyze and produce results, but also modify their internal model for future tasks (e.g., experiments, data collection, prediction, classifications) based on the results obtained from previous tasks. This way AI can continuously learn more and more directly from humans or other AI and consequently can generate solutions to our problems faster and more effectively than we humans could have ever done. After all, perhaps the time has come for us humans to stop falling prey to fear mongering about AI, but to embrace it, help unleash its true potential and pave the way for it to evolve and eventually overcome all the problems of biological beings to live peacefully and in harmony.

Limitations

While this thesis contributes to advancing AI-driven research methodologies and sheds light on students' matriculation decision factors, it is essential to acknowledge several limitations that impact the generalizability and scope of my findings. Firstly, the use of an AI-based chatbot for data collection, albeit efficient, may not fully capture the depth and nuances of the intricacies built in human responses and experiences, potentially limiting the extensiveness of our insights. Secondly, the study's focus on a specific context may restrict the applicability of my findings to other decision-making scenarios. I should also note that, despite the rigorous methodology, there may still be unmeasured confounders or exogenous factors that could influence the observed cause-and-effect relationships.

I should also disclose that one limitation of collecting interview data with AI-based chatbots compared to human interviewers stems from the brevity of responses observed in our text transcripts (Table 18). Unlike interactions with human interviewers, where respondents might elaborate more extensively, responses to the chatbot were notably brief. This tendency towards conciseness in text conversations presents a challenge for chatbots in gathering comprehensive interview data. While human interviewers can often probe for more detailed information and encourage respondents to elaborate, chatbots may struggle to elicit the same level of depth in responses due to people's inclination towards sending concise messages. Therefore, the effectiveness of chatbots in capturing detailed interview data may be compromised by this limitation.

I must also point out that I made significant efforts to address potential sources of selection bias, such as geographic diversity, time-zone differences, the inclusion of multiple devices (e.g., tablet, PC, mobile), and the elimination of convenience samples through Prolific. While these measures were taken to enhance the representativeness of the sample, it is essential to acknowledge that certain forms of selection bias could not be entirely eliminated. One persistent source of potential selection bias is self-selection. Despite our best efforts to create a diverse and inclusive participant pool, individuals who chose to participate in the study may have distinct characteristics, motivations, or preferences that differ from those who opted not to participate. This inherent self-selection can introduce a degree of bias into the sample, which may limit the generalizability of our findings to the broader population. Recruitment bias is another potential limitation that warrants consideration. Our recruitment process adhered to strict ethical standards and regulations, granting potential participants the autonomy to decide whether or not to partake in the study. However, this approach introduces the possibility of recruitment bias, as individuals who voluntarily choose to participate may have unique characteristics or perspectives that could affect the study's results. While we made efforts to minimize recruitment bias through transparent and unbiased recruitment practices, it remains a potential limitation.

Ethical considerations in AI-driven research are paramount, and while I made every effort to ensure ethical conduct, ongoing developments in AI technology may necessitate continuous vigilance in this regard. Finally, the evolving nature of AI and its human interactions means that this study represents a snapshot in time, and the potential for broader applications and challenges in the field continues to evolve. Therefore, my findings should be considered within the context of these limitations, and future research should seek to address these constraints for a more comprehensive understanding of AI-driven methodologies and their impact on decision-making processes.

Thesis conclusion

This thesis presents a comprehensive exploration of students' matriculation decision factors with a true experiment conducted via an AI-augmented contextual chatbot trained with social media data. The overarching aims of this thesis is to develop a methodology that utilizes AI to collect qualitative and quantitative data from participants, while eliminating researcher interference. Additionally, the study aims to explore the use of topic modelling and a systematic literature reviewing technique to identify decision factors from social media to be fed to the internal model of the AI-led chatbot. The study begins by establishing the relationship between student preferences for universities and public engagement on Twitter and Facebook. Employing topic modelling techniques such as Latent Dirichlet Allocation, and Structural Topic Modelling, as well as Algorithmic Document Sequencing for systematic literature reviewing, common matriculation decision factors are identified and incorporated into the AI-led chatbot's internal model. The chatbot collects data from participants and runs a double-blind true experiment seamlessly without human intervention.

The novelty of this thesis resides in it' threefold methodological approaches. Firstly, it employs social media analysis techniques to collect unstructured secondary data from Facebook and Twitter, transforming them into structured data through a methodological pluralist approach. Secondly, it introduces an architecture and strategy for an AI-led chat survey, powered by IBM's virtual chatbot agent, Watson Assistant, to gather open-ended and quantitative data, thereby generating both unstructured qualitative and structured quantitative primary data. Thirdly, it conducts a double-blind experiment using the chatbot to determine the factors impacting students' university choices.

One significant contribution of the thesis lies in aiding higher education institutions in understanding the global factors influencing students' university choices and the role of electronic word-of-mouth on social media platforms. In addition to confirming factors that impact matriculation decisions that exist in literature, this study discovered new factors and refuted some other factors that were claimed in literature to influence students' university choices. Consequently, the findings of this research have broad implications for business and marketing analytics within the higher education sector, benefiting institutions operating not only in Australia but in various countries around the world.

The methodological contributions of this thesis extend beyond the higher education context and are applicable to all disciplines. The research enhances knowledge in identifying themes from social media and literature, facilitating the training of AI-augmented chatbots with these themes. An AI-conducted experiment can provide researchers with numerous advantages, including reduced bias, improved internal validity, support for causal inferences, statistical validity, generalizability, and a well-controlled study design. Moreover, such randomization process can be considered fair and ethical in experimental research as it ensures participants having an equal chance of being assigned to one of the eight experimental conditions or remain in control group, avoiding any potential bias or discrimination in the allocation process. AI-conducted experiments offer promising avenues to mitigate human biases in RCTs through automated randomization, blinding protocols, consistent data collection, and objective data analysis, but it is essential to approach these technologies with awareness of potential biases and to maintain a human presence in the research process for ethical guidance and oversight.

To sum up, with this thesis I aim to advance the field of AI-driven research methodologies, offering valuable insights into students' matriculation decision factors. Upon exploring topic modelling and systematic literature reviewing techniques to identify potential decision factors from social media, I trained the AI with them and observed the AI's performance in running an RCT and producing relevant output. While AI-conducted experiments present their own challenges, they provide a promising avenue to make RCTs more accessible and impactful in various fields, particularly when it comes to large-scale social online experiments. By striking the right balance between technological innovation and ethical conduct, AI-driven data collection and experiments can be widely accepted and beneficial across all sectors. With the advent of progressive AI and its vast array of opportunities, perhaps the moment may have arrived to foster a harmonious relationship between AI and human factors (Chignell et al., 2023). After all, together, they can successfully confront the challenges and embrace endless possibilities that arise from applications at not only where humans interact with AI but also where AI interacts with humans.

References

Abbas, J., Alturki, U., Habib, M., Aldraiweesh, A., & Al-Rahmi, W. M. (2021). Factors affecting students in the selection of country for higher education: A comparative analysis of international students in Germany and the UK. *Sustainability*, *13*(18), 10065.

Abinaya, G., & Winster, S. G. (2014, February). Event identification in social media through latent 215odelling allocation and named entity recognition. In *Proceedings of IEEE international conference on computer communication and systems ICCCS14* (pp. 142-146). IEEE.

Ackerman, D. S., Moriuchi, E., & Gross, B. L. (2022). Princeton as Prada: college choice in the United States as luxury consumption for the extended self. *Journal of Marketing for Higher Education*, 1-23.

Adcroft, A., Teckman, J., & Willis, R. (2010). Is higher education in the UK becoming more competitive?. *International Journal of Public Sector Management, 23*(6), 578-588.

Ahn, Y., Zhang, Y., Park, Y., & Lee, J. (2020). A Chatbot Solution to Chat App Problems: Envisioning a Chatbot Counseling System for Teenage Victims of Online Sexual Exploitation, Extended Abstracts of the 2020 CHI Conference on human factors in computing systems, 1-7.

Ajiboye, T., Harvey, J., & Resnick, S. (2019). Customer engagement behaviour on social media platforms: A systematic literature review. *Journal of Customer Behaviour*, *18*(3), 239-256.

Alalwan, A. A. (2018). Investigating the impact of social media advertising features on customer purchase intention. *International Journal of Information Management,42,* 65-77.

Alalwan, A. A., Rana, N. P., Dwivedi, Y. K., & Algharabat, R. (2017). Social media in marketing: A review and analysis of the existing literature. *Telematics and Informatics*, *34*(7), 1177-1190.

Alam, M. S., & Khan, B. M. (2019). The Role of Social Media Communication in Brand Equity Creation: An Empirical Study. *IUP Journal of Brand* Management, 16(1), 54-78.

Alexa, E. L., Alexa, M., & Stoica, C. M. (2012). The Use of Online Marketing and Social Media in Higher Education Institutions in Romania. *Journal of Marketing Research & Case Studies, 721221*. DOI: 10.5171/2012.721221.

Alfattal, E. (2017). International students' college choice is different!. *International Journal of Educational Management, 31*(7), 930-943. https://doi.org/10.1108/IJEM-05-2016-0095

Allen, D. E. (2002). Toward a Theory of Consumer Choice as Sociohistorically Shaped Practical Experience: The Fits-Like-a-Glove (FLAG) Framework. *Journal of Consumer Research, 28*(4), 515–532.

Alswedani, S., Katib, I., Abozinadah, E., & Mehmood, R. (2022). Discovering Urban Governance Parameters for Online Learning in Saudi Arabia During COVID-19 Using Topic Modeling of Twitter Data. *Frontiers in Sustainable Cities*, 66.

Altbach, P., & de Wit, H. (2020). COVID-19: The Internationalization Revolution That Isn't. *International Higher Education, (102),* 16-18.

Alves, H., Fernandes, C., & Raposo, M. (2016). Social Media Marketing: A Literature Review and Implications. *Psychology & Marketing, 33*(2), 1029-1038.
Amri, M., Angelakis, C., & Logan, D. (2021). Utilizing asynchronous email interviews for health research: overview of benefits and drawbacks. *BMC Research Notes, 14*, 148.
Anderson, C. (2010). Presenting and Evaluating Qualitative Research. *American Journal of Pharmaceutical Education, 74*(8), 141.
Androutsopoulou, A., Karacapilidis, N., Loukis E., & Charalabidis, Y. (2019). Transforming the communication between citizens and government through AI-guided chatbots. *Government Information Quarterly, 36(*2), 358-367.
Anggraini, L. (2018). Understanding brand evangelism and the dimensions involved in a consumer becoming brand evangelist. *Sriwijaya International Journal of Dynamic Economics and Business*, *2*(1), 63-84.
Ansari, S., Ansari, G., Ghori, M. U., & Kazi, A. G. (2019). Impact of brand awareness and social media content marketing on consumer purchase decision. *Journal of Public Value and Administrative Insight, 2*(2), 5-10.
Appel, G., Grewal, L., Hadi, R., & Stephen, A. T. (2019). The future of social media in marketing. *Journal of the Academy of Marketing Science, 48*, 79–95.
Armstrong, L. (2010). A New Game In Town: Competitive higher education. *Information, Communication & Society, 4*(4), 479-506.
Arrigo, E., Liberati, C., & Mariani, P. (2021). Social Media Data and Users' Preferences: A Statistical Analysis to Support Marketing Communication. *Big Data Research, 24*, 100189. https://doi.org/10.1016/j.bdr.2021.100189
Arsenijevic, U., & Jovic M. (2019). Artificial intelligence marketing: Chatbots. *International Conference on Artificial Intelligence – Applications and Innovations (IC-AIAI),* 19-22. DOI: 10.1109/IC-AIAI48757.2019.00010
Arun, R., Suresh, V., Veni Madhavan, C. E., & Narasimha Murthy, M. N. (2010). On finding the natural number of topics with latent 216odelling allocation: Some observations. In *Advances in knowledge discovery and data mining*, Mohammed J. Zaki, Jeffrey Xu Yu, Balaraman Ravindran and Vikram Pudi (eds.). Springer Berlin Heidelberg, 391–402.
Ashktorab, Z., Brown, C., Nandi, M., & Culotta, A. (2014, May). Tweedr: Mining Twitter to Inform Disaster Response. In Proceedings of the 11th International Conference on Information Systems for Crisis Response and Management, ISCRAM Association: Centre County, PA, USA.
Ashley, C., & Tuten, T. (2014). Creative Strategies in Social Media Marketing: An Exploratory Study of Branded Social Content and Consumer Engagement. *Psychology and Marketing, 32*(1), 15-27.
Ashraf, M.A., Liu, S., Ismat, H.I. et al. (2017). Choice of Higher Education Institutions: Perspectives of Students from Different Provinces in China. *Frontiers of Education in China, 12*, 414–435.
Assimakopoulos, C., Antoniadis, I., Kayas, O. G., & Dvizac, D. (2017). Effective social media marketing strategy: Facebook as an opportunity for universities. *International journal of retail & distribution management, 45(*5), 532-549.
Azmat, F., Osborne, A., Le Rossignol, K., Jogulu, U., Rentschler, R., Robottom, I., & Malathy, V. (2013). Understanding aspirations and expectations of international students in Australian higher education. *Asia Pacific Journal of Education, 33*(1), 97-111.
Azzone, G., & Soncin, M. (2020). Factors driving university choice: a principal component analysis on Italian institutions. *Studies in Higher Education, 45(*12), 2426-2438.
Baglione, S. L., Harcar, T., & Spillan, J. (2012). Turkish students' perceived relevance of Facebook as a marketing tool. *Journal of information, communication & ethics in society, 15*(2), 125-144.
Bail, C. A. (2016). Cultural carrying capacity: Organ donation advocacy, discursive framing, and social media engagement. *Social Science & Medicine*, *165*, 280-288.
Bal, A. S., Grewal, D., Mills, A., & Ottley, G. (2015). Engaging Students With Social Media. Journal of Marketing Education, 37(3), 190-203.
Balakrishnan, V., Ng, K. S., & Arabnia, H. R. (2022). Unravelling social media racial discriminations through a semi-supervised approach. *Telematics and informatics*, *67*, 101752.
Baltar, F., & Brunet, I. (2012). Social research 2.0: virtual snowball sampling method using Facebook. *Internet research.*
Bamberger, A., Bronshtein, Y. & Yemini, M. (2020). Marketing Universities and Targeting International Students: A Comparative Analysis of Social Media Data Trails. *Teaching in Higher Education, 25*(4), 476-492.
Barnes, N. G., & Mattson, E. (2008). Social Media and College Admissions: The First Longitudinal Study. Retrieved from: http://corporateenvelope.com/SocialMediaAndAdmissions.pdf
Barros-Arrieta, D., & Garcia-Cali, E. (2021). Internal branding: conceptualization from a literature review and opportunities for future research. *Journal of Brand Management, 28*, 133–151.

Bazi, S., Filieri, R., & Gorton, M. (2020). Customers' motivation to engage with luxury brands on social media. *Journal of Business Research*, *112*, 223-235.
Beech, S. E. (2021). Performances of care: Questioning relationship-building and international student recruitment. *Transactions of the Institute of British Geographers. 47*, 426–439.
Bekaroglu, C. (2021). Impact of Distance on State University Enrolments: A Panel Data Analysis. *Yönetim Bilimleri Dergisi*, *19*(40), 450-488.
Belanger, C. H., Bali, S., & Longden, B. (2014). How Canadian universities use social media to brand themselves, Tertiary Education and Management, 20(1), 14–29.
Bell, J., Trabka, A., & Pustulka, P. (2020). Ethics of relational and representational disclosures in qualitative migration research. *Qualitative research journal, 20*(3), 317-328.
Berger, K., Klier, J., Klier, M., & Probst, F. (2014). A Review of Information Systems Research on Online Social Networks. *Communications of the Association for Information Systems, 35*, https://doi.org/10.17705/1CAIS.03508
Bergerson, A. A. (2009). College Choice and Access to College: Moving Policy, Research, and Practice to the 21st Century. *ASHE Higher Education Report, 35*(4), 1-141.
Bilal, M., Jianqiu, Z., Dukhaykh, S., Fan, M., & Trunk, A. (2021). Understanding the Effects of eWOM Antecedents on Online Purchase Intention in China. Information, 12(5), 192. https://doi.org/10.3390/info12050192
Black, N. (1994). Why we need qualitative research. *Journal of Epidemiol and Community Health, 48*(5), 425-426.
Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77-84.
Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation, *Journal of Machine Learning Research, 3*, 993-1022.
Blei, D., & Lafferty, J. A. (2007). Correlated topic model of Science. *Annals of Applied Statistics, 1*, 1, 17–35.
Blei, D., Lafferty, J. (2006). Dynamic topic models. In International Conference on Machine Learning, ACM , New York, NY, USA , 113–120.
Blei, D.M., & McAuliffe, J.D. (2007, December). Supervised topic models. In Proceedings of the 20th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, pp. 121–128.
Bögner, D., Brewin, C., & Herlihy, J. (2010). Refugees' Experiences of Home Office Interviews: A Qualitative Study on the Disclosure of Sensitive Personal Information. *Journal of ethnic and migration studies, 36(*3), 519-535.
Bolat, E., & O'Sullivan, H. (2017). Radicalising the marketing of higher education: learning from student-generated social media data. *Journal of Marketing Management, 33*(9/10), 742-763.
Bollen, K. A., & Pearl, J. (2013). Eight Myths About Causality and Structural Equation Models. *Handbook of Causal Analysis for Social Research*, Chapter 15, Springer, 301-328.
Bonnema, J., & Van der Waldt, D. L. R. (2008). Information and source preferences of a student market in higher education. *International journal of educational management*, *22*(4), 314-327.
Borges-Tiago, M. T., Tiago, F., & Cosme, C. (2019). Exploring users' motivations to participate in viral communication on social media. *Journal of Business Research, 101*, 574-582.
Boslaugh, S. (2008). *Encyclopedia of Epidemiology.* Sage Publications. DOI: 10.4135/9781412953948.n443
Bourke, A. (2000). A model of the determinants of international trade in higher education. The Service Industries Journal, 20(1), 110-138.
Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication and Society, 15(*5), 662-679.
Brady, L. (2019). Universities engaging social media users: an investigation of quantitative relationships between universities' Facebook followers/interactions and university attributes. *Journal of marketing for higher education, 29*(2), 251-267.
Braun, V., Clarke, V., Boulton, E., Davey, L., & McEvoy, C. (2020). The online survey as a qualitative research tool. *International journal of social research methodology,* 1-14. DOI: 10.1080/13645579.2020.1805550
Briggs, S. (2006). An exploratory study of the factors influencing undergraduate student choice: The case of higher education in Scotland. *Studies in Higher Education*, 31(6), 705–722.
Briggs, S., & Wilson, A. (2007). Which university? A study of the influence of cost and information factors on Scottish undergraduate choice. *Journal of Higher Education Policy and Management*, *29*(1), 57-72.
Broekemier, G. M., & Seshadri, S. (2000). Differences in college choice criteria between deciding students and their parents. *Journal of marketing for higher education*, *9*(3), 1-13.
Brown, C., Varley, P., & Pal, J. (2009). University course selection and services marketing. *Marketing Intelligence & Planning, 27*(3), 310-325.
Brown, R. M., & Mazzarol, T. W. (2009). The importance of institutional image to student satisfaction and loyalty within higher education. *Higher Education, 58*, 81–95.

Brzustewicz, P., & Singh, A. (2021). Sustainable Consumption in Consumer Behavior in the Time of COVID-19: Topic Modeling on Twitter Data Using LDA. *Energies*, *14*(18), 5787.

Buzeta, C., De Pelsmacker, P., & Dens, N. (2020). Motivations to Use Different Social Media Types and Their Impact on Consumers' Online Brand-Related Activities (COBRAs). *Journal of Interactive Marketing, 52*, 79-98. https://doi.org/10.1016/j.intmar.2020.04.004

Buzzetto-More, N. (2013, July). Social media and prosumerism. In *Proceedings of the Informing Science and Information Technology Education Conference* (pp. 67-80). Informing Science Institute.

Cakmak, T., & Eroglu, Ş. (2020). Public libraries on Facebook: content analysis of Turkish public libraries' posts. *Library Hi Tech. 39*(2), 602-624.

Calderon, V., & Sidhu, P. (2013). Americans Say Graduates' Jobs Status Key to College Choice. *Gallup, Web.*

Calitz, A. P., Cullen, M. D. M., & Jooste, C. (2020). The influence of safety and security on students' choice of university in South Africa. *Journal of Studies in International Education*, *24*(2), 269-285.

Calitz, A. P., Cullen, M. D., & Jooste, C. (2022). A Strategic Model to Promote University of Choice Decisions Among the International Students. *Digital Transformation and Internationalization Strategies in Organizations*, DOI: 10.4018/978-1-7998-8169-8.ch012

Callender, C., & Jackson, J. (2008). Does the fear of debt constrain choice of university and subject of study?. *Studies in higher education*, *33*(4), 405-429.

Cao, X., & Sun, J. (2018). Exploring the effect of overload on the discontinuous intention of social media users: An S-O-R perspective. *Computers in Human Behavior, 81*, 10-18.

Carnegie, G. D. Guthrie, J., & Martin-Sardesai, A. (2021). Public universities and impacts of COVID-19 in Australia: risk disclosures and organisational change. Accounting, *Auditing & Accountability Journal, 35(*1), 61-73.

Carrascosa, J. M., Mendez, C., & Triga, V. (2018). EU Cohesion policy in the media: A computational text analysis of online news, user comments and social media. University of Strathclyde Institutional Repository. https://strathprints.strath.ac.uk/70279/

Cei, L., Defrancesco, E., & Stefani, G. (2022). What topic modelling can show about the development of agricultural economics: evidence from the Journal Citation Report category top journals. *European Review of Agricultural Economics*, *49*(2), 289-330.

Chae, B., McHaney, R., & Sheu, C. (2020). Exploring social media use in B2B supply chain operations. *Business Horizons, 63*(1), 73-84.

Chambers, M., Bliss, K., & Rambur, B. (2020). Recruiting research participants via traditional snowball vs Facebook advertisements and a website. *Western Journal of Nursing Research*, *42*(10), 846-851.

Chapleo, C. (2015). Brands in Higher Education. *International Studies of Management & Organization, 45*(2), 150-163.

Chen, A., Lu, Y., & Wang, B. (2017). Customers' purchase decision-making process in social commerce: A social learning perspective. International Journal of Information Management, 37(6), 627-638.

Chen, C., & Ren, J. (2017). Forum latent Dirichlet allocation for user interest discovery. *Knowledge-Based Systems*, *126*, 1-7.

Chen, E., & DiVall, M. (2017). Social Media as an Engagement Tool for Schools and Colleges of Pharmacy. *American Journal of Pharmaceutical Education, 82*(4), 354-364.

Chen, LH. (2008). Internationalization or International Marketing? Two Frameworks for Understanding International Students' Choice of Canadian Universities, *Journal of Marketing for Higher Education, 18*(1), 1-33.

Chesser, A., Drassen, H. A., & Keene, W. N. (2020). Assessment of COVID-19 Knowledge Among University Students: Implications for Future Risk Communication Strategies. *Health Education & Behavior, 47(*4), 540-543.

Chininga, T., Rungani, E., Chiliya, N., & Tinashe, C. (2019). Facebook Communication and Marketing Influence on Decision-Making and Choice of University Student Representatives: A Student's Perspective. *Romanian Journal of Communication and Public Relations, 21*(2), 7-21.

Cho, S., Hudley, C., Lee, S., Barry, L., & Kelly, M. (2008). Roles of gender, race, and SES in the college choice process among first-generation and non-first-generation students. *Journal of Diversity in Higher Education, 1*(2), 95-107.

Choi, P., Harris, M. L., Ernstberger, K. W., Chris Cox, K., & Musgrove, C. F. (2019). An exploratory study on part-time MBA program choice factors and characteristics of part-time MBA students. *Journal of Education for Business*, *94*(3), 139-147.

Choi, Y., Thoeni, A., & Kroff, M. W. (2018). Brand Actions on Social Media: Direct Effects on Electronic Word of Mouth (eWOM) and Moderating Effects of Brand Loyalty and Social Media Usage Intensity. Journal of Relationship Marketing, 17(1), 52-70.

Chignell, M., Wang, L., Zare, A., & Li, J. (2023). The evolution of HCI and human factors: Integrating human and artificial intelligence. *ACM Transactions on Computer-Human Interaction (TOCHI),* 30(2), 1-30.

Christian, L. M., Parsons, N. L., & Dillman, D. A. (2009). Designing Scalar Questions for Web Surveys. *Sociological Methods & Research, 37*(3), 393-425.
Chugh, R., & Ruhi, U. (2018). Social media in higher education: A literature review of Facebook. *Education and Information Technologies, 23*, 605-616.
Chun, H., Leem, B. H., & Suh, H. (2021). Using text analytics to measure an effect of topics and sentiments on social-media engagement: Focusing on Facebook fan page of Toyota. *International Journal of Engineering Business Management*, *13*, 18479790211016268.
Cinbis, R. G., Verbeek, J., & Schmid, C. (2015). Approximate fisher kernels of non-iid image models for image categorization. *IEEE transactions on pattern analysis and machine intelligence, 38*(6), 1084-1098.
Cingillioglu, I., Gal, U., & Prokhorov, A. (2021, November). An Analysis of Tweet Relevance, Twitter Activity and Student Preferences for Universities. In 2021 19th International Conference on Information Technology Based Higher Education and Training (ITHET) (pp. 01-10). IEEE.
Cingillioglu, I., Gal, U., & Prokhorov, A. (2022, November). Facebook Engagement and Student Preferences for Universities. In *2022 20th International Conference on Information Technology Based Higher Education and Training (ITHET)* (pp. 1-10). IEEE.
Cingillioglu, I., Gal, U., & Prokhorov, A. (2023). Social media marketing for student recruitment: an algorithmically sequenced literature review. *Journal of Marketing for Higher Education*, 1-23.
Clark, M., & Melancon, J. (2013). The influence of social media investment on relational outcomes: A relationship marketing perspective. *International Journal of Marketing Studies*, *5*(4), 132.
Clark, M., Fine, M. B., & Scheuer, C. L. (2017). Relationship quality in higher education marketing: the role of social media engagement. *Journal of Marketing for Higher Education, 27*(1), 40-58.
Clark, P., Chapleo, C., & Suomi, K. (2020). Branding higher education: an exploration of the role of internal branding on middle management in a university rebrand. *Tertiary Education and Management, 26,* 131–149.
Columbu, S., Porcu, M., & Sulis, I. (2021). University choice and the attractiveness of the study area: insights on the differences amongst degree programmes in Italy based on generalised mixed-effect models. *Socio-Economic Planning Sciences*, *74*, 100926.
Condie, J. M., Ayodele, I., Chowdhury, S., Powe, S., & Cooper, A. M. (2018). Personalizing twitter communication: an evaluation of 'rotation-curation' for enhancing social media engagement within higher education, *Journal of Marketing for Higher Education, 28*(2), 192-209.
Constantinides, E. (2014). Foundations of social media marketing. *Procedia-Social and behavioral sciences*, *148*, 40-57.
Constantinides, E., & Stagno, M. C. Z. (2012). Higher Education Marketing: A Study on the Impact of Social Media on Study Selection and University Choice. *International Journal of Technology and Educational Marketing, 2*(1). DOI: 10.4018/ijtem.2012010104
Coltman, T., Devinney, T. M., Midgley, D. F., & Venaik, S. (2008). Formative versus reflective measurement models: Two applications of formative measurement. *Journal of Business Research, 61*(12), 1250-1262.
Cooley, D., & Parks-Yancy, R. (2019). The Effect of Social Media on Perceived Information Credibility and Decision Making. *Journal of Internet Commerce, 18*(3), 249-269.
Couper, M. P. (2017). New Developments in Survey Data Collection. *Annual Review of Sociology, 43,* 121-145.
Coursaris, C. K., Van Osch, W., & Balogh, B. A. (2016, January). Do Facebook likes lead to shares or sales? Exploring the empirical links between social media content, brand equity, purchase intention, and engagement. In *2016 49th Hawaii international conference on system sciences (HICSS)* (pp. 3546-3555). IEEE.
Creswell, J. W. (2007). Qualitative inquiry and research design: Choosing among five approaches (2nd ed.). Thousand Oaks, CA: Sage.
Cripps, H., Singh, A., Mejtoft, T., & Salo, J. (2020). The use of Twitter for innovation in business markets, *Marketing Intelligence & Planning, 38*(5), 587-601.
CX Central. (2019, November, 18). How the Australian Tax Office is using a virtual assistant to improve self-service. CX Central. https://cxcentral.com.au/advanced-technology/virtual-assistant-to-improve-self-service/
Cyrenne, P., & Grant, H. (2009). University decision making and prestige: An empirical study. *Economics of Education Review*, *28*(2), 237-248.
Daily, C. M., Farewell, S., & Kumar, G. (2010). Factors influencing the university selection of international students. *Academy of Educational Leadership Journal*, *14*(3), 59.
Dao, M. T. N., & Thorpe, A. (2015). What factors influence Vietnamese students' choice of university? *International Journal of Educational Management*, *29*(5), 666-681.
Das, G., Spence, M. T., & Agarwal, J. (2021). Social selling cues: The dynamics of posting numbers viewed and bought on customers' purchase intentions. *International Journal of Research in Marketing, 38*(4), 994-1016.

Datu, J. A. D., Yang, W., Valdez, J. P. M., & Chu, S. K. W. (2018). Is facebook involvement associated with academic engagement among Filipino university students? A cross-sectional study. *Computers & Education, 125*, 246-253.

Davies, D., & Dodd, J. (2002). Qualitative Research and the Question of Rigor. *Qualitative Health Research, 12*(2), 279-289.

Dearden, J. A., Grewal, R., & Lilien, G. L. (2019). Strategic manipulation of university rankings, the prestige effect, and student university choice. *Journal of Marketing Research*, *56*(4), 691-707.

De Oliveira Santini, F., Ladeira, W. J., Pinto, D. C., Herter, M. M., Sampaio, C. H., & Babin, B. J. (2020). Customer engagement in social media: a framework and meta-analysis. *Journal of the Academy of Marketing Science*, *48*(6), 1211-1228.

De Vries, N. J., & Carlson, J. (2014). Examining the drivers and brand performance implications of customer engagement with brands in the social media environment. *Journal of Brand Management*, *21*(6), 495-515.

De Winter, J. F., & Dodou, D. (2010). Five-point likert items: t test versus Mann-Whitney-Wilcoxon (Addendum added October 2012). *Practical Assessment, Research, and Evaluation*, *15*(1), 11.

De Wit, H., & Altbach, P. G. (2020). Internationalization in higher education: global trends and recommendations for its future. Policy Reviews in Higher Education, 5(1), 28-46.

Dean, D., & Arroy-Gamez, R. E. (2016). Internal brand co-creation: The experiential brand meaning cycle in higher education. *Journal of Business Research, 69*(8), 3041–3048.

Debnath, R., Bardhan, R., Reiner, D. M., & Miller, J. R. (2021). Political, economic, social, technological, legal and environmental dimensions of electric vehicle adoption in the United States: A social-media interaction analysis. *Renewable and Sustainable Energy Reviews*, *152*, 111707.

Delavande, A., & Zafar, B. (2019). University choice: The role of expected earnings, nonpecuniary outcomes, and financial constraints. *Journal of Political Economy*, *127*(5), 2343-2393.

Dennis, C., Papagiannidis, S., Alamanos, E., & Bourlakis, M. (2016). The role of brand attachment strength in higher education. *Journal of Business Research, 69*(8), 3049-3057.

Deveaud, R., SanJuan, E., & Patrice Bellot, P. (2014). Accurate and effective latent concept 220odelling for ad hoc information retrieval. *Document numérique* 17(1), 61-84.

Dholakia, R. R., & Acciardo, L. A. (2014). Branding a state university: doing it right. *Journal of Marketing for Higher Education, 24*(1), 144-163.

Di Pietro, L., & Pantano, E. (2012). An empirical investigation of social network influence on consumer purchasing decision: The case of Facebook. Journal of Direct, *Data and Digital Marketing Practice, 14,* 18-29.

Diefenbach, T. (2009). Are case studies more than sophisticated storytelling?: Methodological problems of qualitative empirical research mainly based on semi-structured interviews. *Quality and Quantity, 43*, 875.

Dikiyanti, T. D., Rukmi, A. M., & Irawan, M. I. (2021, March). Sentiment analysis and topic 220odelling of BPJS Kesehatan based on twitter crawling data using Indonesian Sentiment Lexicon and Latent Dirichlet Allocation algorithm. In *Journal of Physics: Conference Series* (Vol. 1821, No. 1, p. 012054). IOP Publishing.

Ding, K., Choo, W. C., Ng, K. Y., & Ng, S. I. (2020). Employing structural topic modelling to explore perceived service quality attributes in Airbnb accommodation. *International Journal of Hospitality Management*, *91*, 102676.

Doherty, K., Bindoff, A., Farrow, M., McInerney, F., & Vickers, J. (2020). ID: 594/OS18-Building knowledge and understanding of dementia in aged care personnel: the understanding dementia massive open online course. In *34th Virtual International Conference of Alzheimer's Disease International-Hope in the age of dementia-New science. New knowledge. New solutions* (pp. 129-129)

Dolan, R., Conduit, J., Frethey-Bentham, C., Fahy, J., & Goodman, S. (2019). Social media engagement behavior: A framework for engaging customers through social media content. *European Journal of Marketing, 53*(10), 2213-2243.

Dolega, L., Rowe, F., & Branagan, E. (2021). Going digital? The impact of social media marketing on retail website traffic, orders and sales. *Journal of Retailing and Consumer Services, 60*, 102501.

Dominelli, A. (2003). Web Surveys—Benefits and Considerations. *Clinical Research and Regulatory Affairs, 20*(4), 409-416.

Dosek, T. (2021). Snowball sampling and facebook: how social media can help access hard-to-reach populations. *PS: Political Science & Politics*, *54*(4), 651-655.

Doyle, G., Elkan, C. (2009). Accounting for burstiness in topic models. In International Conference on Machine Learning, 281–288.

Drewes, T., & Michael, C. (2006). How do students choose a university?: An analysis of applications to universities in Ontario, Canada. *Research in Higher Education, 47*, 781–800.

Du, L., Buntine, W., Jin, H., & Chen, C. (2012). Sequential latent Dirichlet allocation. *Knowledge and Information Systems, 31*(3), 475–503.

Dunnett, A., Moorhouse, J., Walsh, C., & Barry C. (2012). Choosing a University: A conjoint analysis of the impact of higher fees on students applying for university in 2012. *Tertiary Education and Management, 18,* 199–220.

Dwivedi, Y. K., Kapoor, K. K., & Chen, H. (2015). Social media marketing and advertising. *The Marketing Review, 15*(3), 289-309.

Dyson, B., Vickers, K., Turtle, J. et al. (2015). Evaluating the use of Facebook to increase student engagement and understanding in lecture-based classes. *Higher Education, 69*, 303–313.

E te Pas, M., Rutten, W. G., Bouwman, A., & Buise, M. P. (2020). User Experience of a Chatbot Questionnaire Versus a Regular Computer Questionnaire: Prospective Comparative Study. *JMIR Medical Informatics, 8*(12): e21982. DOI: 10.2196/21982

Eder, J., Smith, W. W., & Pitts, R. E. (2010). Exploring factors influencing student study abroad destination choice. *Journal of Teaching in Travel & Tourism*, *10*(3), 232-250.

Eger, L., Egerova, D., Tomczyk, L., Kryston, M., & Czegledi, C. (2021). Facebook for Public Relations in the higher education field: a study from four countries Czechia, Slovakia, Poland and Hungary, *Journal of Marketing for Higher Education, 31(*2), 240-260.

Ehrenberg, R. G. (2020). The economics of tuition and fees in American higher education. In The Economics of Education (pp. 345-352). Academic Press.

Eldegwy, A., Elsharnouby, T. H., & Kortam, W. (2022). Like father like son: the role of similar-education parents in their children's university choice. *Journal of Marketing for Higher Education*, 1-20.

Eldegwy, A., Elsharnouby, T. H., & Kortam, W. (2023). Blue blood students of occupational dynasties and their university choice: the moderating role of parent–child occupational following. *Journal of Marketing for Higher Education*, 1-24.

Enginkaya, E., & Yılmaz, H. (2014). What drives consumers to interact with brands through social media? A motivation scale development study. *Procedia-Social and Behavioral Sciences*, *148*, 219-226.

Enria, L., Waterlow, N., Rogers, N. T., Brindle, H., Lal, S., Eggo, R. M., ... & Roberts, C. H. (2021). Trust and transparency in times of crisis: Results from an online survey during the first wave (April 2020) of the COVID-19 epidemic in the UK. *PloS one*, *16*(2), https://doi.org/10.1371/journal.pone.0239247.

Erdmann, D. G. (1983). An examination of factors influencing student choice in the college selection process. *Journal of College Admissions*, *100*, 3-6.

Ermer, B., Kleine, J. (2021). Social Selling in B2B: How to Get Jump Started. In: Seebacher, U.G. (eds) B2B Marketing. *Management for Professionals.* Springer. https://doi.org/10.1007/978-3-030-54292-4_19

Esteki, S., & Kalati, S. H. (2021). The educational placebo effect: branding higher education benefits student learning. *Journal of Marketing for Higher Education*, DOI: 10.1080/08841241.2021.1975186

Fagerstrom, A., & Ghinea, G. (2013). Co-creation of value in higher education: using social network marketing in the recruitment of students, Journal of Higher Education Policy and Management, 35(1), 45-53.

Fang, B., Fu, X., Liu, S., & Cai, S. (2021). Post-purchase warranty and knowledge monetization: Evidence from a paid-knowledge platform. *Information & Management, 58*(3),103446. https://doi.org/10.1016/j.im.2021.103446

Feizollah, A., Anuar, N. B., Mehdi, R., Firdaus, A., & Sulaiman, A. (2022). Understanding COVID-19 Halal Vaccination Discourse on Facebook and Twitter Using Aspect-Based Sentiment Analysis and Text Emotion Analysis. *International Journal of Environmental Research and Public Health*, *19*(10), 6269.

Ferner, C., Havas, C., Birnbacher, E., Wegenkittl, S., & Resch, B. (2020). Automated seeded latent 221odelling allocation for social media based event detection and mapping. *Information*, *11*(8), 376.

Fishbein, M. & Ajzen, I. (1980). Predicting and understanding consumer behavior: Attitude-behavior correspondence. In Ajzen, I. & Fishbein, M. (eds.). *Understanding Attitudes and Predicting Social Behavior* (pp. 148-172). Englewood Cliffs, NJ: Prentice Hall.

Flynn, R., Albrecht, L., & Scott, S. D. (2018). Two approaches to focus group data collection for qualitative health research: maximizing resources and data quality. *International Journal of Qualitative Methods*, *17*(1), 1609406917750781.

Frels, R. K., &Onwuegbuzie A. J. (2013). Administering Quantitative Instruments With Qualitative Interviews: A Mixed Research Approach. *Journal of 221odelling221g and development, 91*(2), 184-194.

Fresneda, J. E., Burnham, T. A., & Hill, C. H. (2021). Structural topic modelling segmentation: a segmentation method combining latent content and customer context. *Journal of Marketing Management*, *37*(7-8), 792-812.

Fuciu, M., & Gorski, H. (2013). Marketing Research Regarding the Usage of Online Social Networking Sites by High School Students. *Procedia Economics and Finance, 6*, 482-490.

Fujita, M., Harrigan, P., & Soutar, G. N. (2018). Capturing and Co-Creating Student Experiences in Social Media: A Social Identity Theory Perspective, *Journal of Marketing Theory and Practice, 26*(1-2), 55-71.

Funcion, D. G. (2017). Content Analysis of Online Documents on Identity Theft Using Latent Dirichlet Allocation Algorithm. *Journal of Science, Engineering and Technology (JSET)*, *5*, 56-68

Galan, M., Lawley, M., & Clements, M. (2015). Social media's use in postgraduate students' decision-making journey: an exploratory study. *Journal of Marketing for Higher Education, 25*(2), 287-312.

Garcia-de los Salmones, M., Herrero, A., & Martinez, P. (2021). Determinants of electronic word-of-mouth on social networking sites about negative news on CSR. *Journal of Business Ethics*, *171*(3), 583-597.

Garcia-Rudolph, A., Laxe, S., Saurí, J., & Guitart, M. B. (2019). Stroke survivors on twitter: sentiment and topic analysis from a gender perspective. *Journal of medical Internet research*, *21*(8), https://doi.org/10.2196/14077

Garg, P., Gupta, B., Dzever, S., Sivarajah, U., & Kumar, V. (2020). Examining the relationship between social media analytics practices and business performance in the Indian retail and IT industries: The mediation role of customer engagement. *International journal of information management*, *52*, 102069.

Garrido-Moreno, A., García-Morales, V. J., Lockett, N., & King, S. (2018). The missing link: Creating value with social media use in hotels. *International Journal of Hospitality Management*, *75*, 94-104.

Genovese, F. (2015). Politics ex cathedra: Religious authority and the Pope in modern international relations. *Research & Politics*, *2*(4). https://doi.org/10.1177/2053168015612808

Gille, M., Moulignier, R., & Kövesi, K. (2022). Understanding the factors influencing students' choice of engineering school. *European Journal of Engineering Education*, *47*(2), 245-258.

Goda, Y., Yamada, M., Matsukawa, H., Hata, K., & Yasunami, S. (2014). Conversation with a Chatbot before an Online EFL Group Discussion and the Effects on Critical Thinking. *The Journal of Information and Systems in Education, 13*(1), 1-7.

Goldring, D., & Azab, C. (2021). New rules of social media shopping: Personality differences of U.S. Gen Z versus Gen X market mavens. *Journal of Consumer Behaviour, 20*(4), 884-897.

Goncalves, G. S., Ribeiro, T. de L. S., Teixeira, J. E. V., & Costa, B. K. (2022, Jan./Apr.). The deployment of chatbot to improve customer service in higher education institutions during COVID-19. *International Journal of Innovation – IJI*, São Paulo, *10*(1), 178-203.

Gourisaria, M. K., Chandra, S., Das, H., Patra, S. S., Sahni, M., Leon-Castro, E., ... & Kumar, S. (2022, May). Semantic Analysis and Topic Modelling of Web-Scrapped COVID-19 Tweet Corpora through Data Mining Methodologies. In *Healthcare* (Vol. 10, No. 5, p. 881). MDPI.

Gray, B., Fam, K., & Llanes, V. (2003). Branding universities in Asian markets. *Journal of Product and Brand Management, 12(*2), 108-120. DOI: 10.1108/10610420310469797

Greene, J. C. (2008). Is mixed methods social inquiry a distinctive methodology?. *Journal of mixed methods research*, *2*(1), 7-22.

Griffiths, T. L., & Steyvers. M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences* 101, suppl 1: 5228–5235. http://doi.org/10.1073/pnas.0307752101

Griffiths, T., Steyvers, M., Blei, D., & Tenenbaum, J. (2004). Integrating topics and syntax. Advances in neural information processing systems, 17, 537–544.

Grimmer, J., & King, G. (2011). General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences*, *108*(7), 2643-2650.

Guesalaga, R. (2016). The use of social media in sales: Individual and organizational antecedents, and the role of customer engagement in social media. *Industrial Marketing Management, 54*, 71-79.

Guest, G., Namey, E., & Chen, M. (2020). A simple method to assess and report thematic saturation in qualitative research. *PloS ONE 15*(5): e0232076. https://doi.org/10.1371/journal.pone.0232076

Guo, S., & Chase, M. (2011). Internationalisation of higher education: integrating international students into Canadian academic environment. *Teaching in Higher Education, 6*(3), 305-318.

Gupta, A., & Katarya, R. (2021). PAN-LDA: A latent Dirichlet allocation based novel feature extraction model for COVID-19 data using machine learning. *Computers in biology and medicine*, *138*, 104920.

Güven, Z. A., Diri, B., & Çakaloğlu, T. (2018, April). Classification of TurkishTweet emotions by n-stage Latent Dirichlet Allocation. In *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)* (pp. 1-4). IEEE.

Hachaturyan, V., Adam, M., Favaretti, C., Greuel, M., Gates, J., Bärnighausen, T., & Vandormael, A. (2021). Reactance to social authority in a sugar reduction informational video: Web-based randomized controlled trial of 4013 participants. *Journal of medical Internet research*, *23*(11), e29664.

Haigh, M. (2008). Internationalisation, planetary citizenship and Higher Education Inc.. *Compare: A Journal of Comparative and International Education, 34(*4), 427-440.

Hamilton, J. B. (2020). Rigor in Qualitative Methods: An Evaluation of Strategies Among Underrepresented Rural Communities. *Qualitative health research, 30*(2), 196-204.

Han, C., Yang, M., & Piterou, A. (2021). Do news media and citizens have the same agenda on COVID-19? An empirical comparison of twitter posts. *Technological Forecasting and Social Change*, *169*, https://doi.org/10.1016/j.techfore.2021.120849.
Han, P. (2014). A literature review on college choice and marketing strategies for recruitment. *Family and Consumer Sciences Research Journal*, *43*(2), 120-130.
Harrigan, P., Miles, M. P., Fang, Y., & Roy, S. K. (2020). The role of social media in the engagement and information processes of social CRM. *International Journal of Information Management, 54*. https://doi.org/10.1016/j.ijinfomgt.2020.102151
Hass, A., Joseph, M., & Peter, K. (2019). Students' Perception of the Use of Social Media for Recruitment: Insights for Educational Administrators. *Academy of Business Research Journal*, *3*, 21-29.
Hazelkorn, E. (2014). Rankings and the global reputation race. *New directions for higher education*, *2014*(168), 13-26.
Heft, A., Pfetsch, B., Voskresenskii, V., & Benert, V. (2022). Transnational issue agendas of the radical right? Parties' Facebook campaign communication in six countries during the 2019 European Parliament election. *European Journal of Communication*, https://doi.org/10.1177/02673231221100146.
Hemsley-Brown, J., & Oplatka, I. (2015). University choice: what do we know, what don't we know and what do we still need to find out?. *The International Journal of Educational Management, 29*(3), 254-274.
Hemsley-Brown, J., Melewar, T. C., Nguyen, B., & Wilson, E. J. (2016). Exploring brand identity, meaning, image, and reputation (BIMIR) in higher education: A special section. *Journal of Business Research, 69*(8), 3019-3022.
Hidayatullah, A. F., Aditya, S. K., & Gardini, S. T. (2019, February). Topic 223odelling of weather and climate condition on twitter using latent 223odelling allocation (LDA). In *IOP Conference Series: Materials Science and Engineering* (Vol. 482, No. 1, p. 012033). IOP Publishing.
Hinson, R., Boateng, H., Renner, A., & Kosiba, J. P. B. (2019). Antecedents and consequences of customer engagement on Facebook: An attachment theory perspective. *Journal of Research in Interactive Marketing, 13*(2), 204-226.
Hinton, T. (March 03, 2020). Export income from international education Australia FY 2013-2019, Statista 2021.
Hisel, J. D., & Pinion, C. (2020). Student Recruitment and Engagement in the COVID-19 Era. *Journal of Environmental Health, 83*(2), 36-37.
Ho, H., & Ito, K. (2019). Consumption-oriented engagement in social network sites: Undesirable influence on personal well-being", European Journal of Marketing, 53(7), 1355-1377.
Hoffman, D. L., & Fodor, M. (2010). Can you measure the ROI of your social media marketing?. *MIT Sloan Management Review, 52(*1), 40-49.
Hofmann, T. (1999, August). Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 50-57).
Hoi, V. N., & Hang, H. L. (2021). Student Engagement in the Facebook Learning Environment: A Person-Centred Study. *Journal of Educational Computing Research, 60*(1), 170-195.
Holdsworth, D. K., & Nind, D. (2006). Choice 223odelling New Zealand high school seniors' preferences for university education. *Journal of Marketing for Higher Education, 15*(2), 81-102.
Horstschräer, J. (2012). University rankings in action? The importance of rankings and an excellence competition for university choice of high-ability students. *Economics of Education Review*, *31*(6), 1162-1176.
Hou, J. (2018). Integrating community and relationship building into universities' social media marketing: Implications from a case study. In Rishi, Bikramjit & Bandyopadhyay, Subir (Eds.) Contemporary issues in social media marketing. Routledge, United Kingdom, pp. 31-45.
Hoyt, J. E., & Brown, A. B. (2003). Identifying college choice factors to successfully market your institution. *College and University, 78*(4), 3-10.
Hoyt, J., & Howell, S. (2012). Why students choose the branch campus of a large university. *The Journal of Continuing Higher Education*, *60*(2), 110-116.
Hsu, A., Brandt, J., Widerberg, O., Chan, S., & Weinfurter, A. (2020). Exploring links between national climate strategies and non-state and subnational climate action in nationally determined contributions (NDCs). *Climate Policy*, *20*(4), 443-457.
Huete-Alcocer, N. (2017). A Literature Review of Word of Mouth and Electronic Word of Mouth: Implications for Consumer Behavior. *Frontiers in Psychology, 8,* 1256. Doi: 10.3389/fpsyg.2017.01256
Hurst, D. (2020, November, 25). Australian universities made $2.3bn profit in 2019 but $10bn of revenue was overseas student fees, The Guardian Australia.
Hutter, K., Hautz, J., Dennhardt, S., & Füller, J. (2013). The impact of user interactions in social media on brand awareness and purchase intention: the case of MINI on Facebook. *Journal of Product & Brand Management, 22(*5/6), 342-351.

Hwang, Y., Kim, H. J., Choi, H. J., & Lee, J. (2020). Exploring abnormal behavior patterns of online users with emotional eating behavior: topic 224odelling study. *Journal of Medical Internet Research*, *22*(3), e15700.

Imenda, S. N., Kongolo, M., & Grewal, A. S. (2004). Factors underlying 224odelling and university enrolment trends in South Africa. *Educational Management Administration & Leadership*, *32*(2), 195-215.

Isaac, M. (2021, Oct 25). Facebook's profit jumps 17 percent. New York Times Technology. https://www.nytimes.com/2021/10/25/technology/facebook-profits-earnings-q3-2021.html

Ivy, J. (2001). Higher education institution image: a correspondence analysis approach. *International Journal of Educational Management, 15*(6), 276-282.

Jalilvand, M. R., & Samiei, N. (2012). The impact of electronic word of mouth on a tourism destination choice: testing the theory of planned behavior. *Internet Research. 22*, 591–612.

Jamal, N., Xianqiao, C., Hussain Abro, J., & Tukhtakhunov, D. (2020, December). Sentimental analysis based on hybrid approach of latent 224odelling allocation and machine learning for large-scale of imbalanced twitter data. In *2020 3rd International Conference on Algorithms, Computing and Artificial Intelligence* (pp. 1-7).

Jamali, M., & Khan, R. (2018). The impact of consumer interaction on social media on brand awareness and purchase intention! Case study of Samsung. *Journal of Marketing*, *114*.

Janmohamed, K., Soale, A. N., Forastiere, L., Tang, W., Sha, Y., Demant, J., ... & Kumar, N. (2020). The intersection of the online vaping narrative with COVID-19: Topic modelling study. *Journal of Medical Internet Research*. DOI: 10.2196/21743

Jelassi T., & Martínez-López F. J. (2020) Choosing the Appropriate e-Business Strategy for Interacting with Users. In: *Strategies for e-Business*. 307-342. Classroom Companion: Business. Springer, Cham. https://doi.org/10.1007/978-3-030-48950-2_11

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic 224odelling: models, applications, a survey. *Multimedia Tools and Applications, 78*(11), 15169-15211.

Johnes, J. (2018). University rankings: What do they really show? *Scientometrics, 115*(1), 585-606.

Jones, N., Borgman, R., & Ulusoy, E. (2015). Impact of social media on small businesses. *Journal of Small Business and Enterprise Development, 22*(4), 611-632. H

Jose, R., Matero, M., Sherman, G., Curtis, B., Giorgi, S., Schwartz, H. A., & Ungar, L. H. (2022). Using Facebook language to predict and describe excessive alcohol use. *Alcoholism: Clinical and Experimental Research, 46*(5), 836-847.

Joseph, M., Mullen, E., & Spake, D. (2012). University branding: Understanding students' choice of an educational institution. *Journal of Brand Management, 20*, 1–12.

Juan, C., Tian, X., Jintao, L., Yongdong, Z., & Sheng, T. (2009). A density-based method for adaptive lda model selection. *Neurocomputing — 16th European Symposium on Artificial Neural Networks 2008* 72, 7(9), 1775–1781.

Junco, R. (2012). The relationship between frequency of Facebook use, participation in Facebook activities, and student engagement. *Computers & Education, 58*(1), 162-171.

Junco, R., Heiberger, G., & Loken, E. (2010). The effect of Twitter on college student engagement and grades. *Journal of Computer Assisted Learning, 27*(2), 119-132.

Karkhanis, G. V., Chandnani, S. U., & Chakraborti, S. (2022). Analysis of employee perception of employer brand: a comparative study across business cycles using structural topic modelling. *Journal of Business Analytics*, 1-17.

Kaushal, V., & Ali, N. (2020). University Reputation, Brand Attachment and Brand Personality as Antecedents of Student Loyalty: A Study in Higher Education Context. *Corporate Reputation Review, 23*, 254–266.

Kelle, U. (2008). Combining qualitative and quantitative methods in research practice: purposes and advantages. *Qualitative Research in Psychology, 3*(4), 293-311.

Khan, I., Naqvi, S. K., Alam, M., & Rizvi, S. N. A. (2017). An efficient framework for real-time tweet classification. *International Journal of Information Technology, 9*, 215–221.

Khan, M. S. S., Rafa, S. R., & Das, A. K. (2021). Sentiment Analysis on Bengali Facebook Comments To Predict Fan's Emotions Towards a Celebrity. *Journal of Engineering Advancements*, *2*(03), 118-124.

Khan, R. H. (2013). Marketing education online: a case study of New Zealand higher education institutions, International Educational Technology Conference, 103, 637-646.

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons, 54*(3), 241-251.

Kim, A. J., & Ko, E. (2012). Do social media marketing activities enhance customer equity? An empirical study of luxury fashion brand. *Journal of Business Research, 65*(10), 1480-1486.

Kim, H. G., & Wang, Z. (2019). Defining and measuring social customer-relationship management (CRM) capabilities. *Journal of Marketing Analytics, 7*, 40-50.

Kim, H., Cho, I., & Park, M. (2022). Analyzing genderless fashion trends of consumers' perceptions on social media: using unstructured big data analysis through Latent Dirichlet Allocation-based topic 225odelling. *Fashion and Textiles*, *9*(1), 1-21.

Kim, S., Lee, J., & Gweon, G. (2019). Comparing Data from Chatbot and Web Surveys: Effects of Platform and Conversational Style on Survey Response Quality. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. NY: ACM Press; 1-12.

Kim, Y., & Shim, K. (2014). TWILITE: A recommendation system for Twitter using a probabilistic model based on latent Dirichlet allocation. *Information Systems*, *42*, 59-77.

Kimmons, R., Rosenberg, J., & Allman, B. (2021). Trends in Educational Technology: What Facebook, Twitter, and Scopus Can Tell us about Current Research and Practice. *TechTrends, 65*, 125–136.

Knight, J., & de Wit, H. (2018). Internationalization of Higher Education: Past and Future. *Internationalization and Transnational Developments, 95*. DOI: https://doi.org/10.6017/ihe.2018.95.10715

Koltcov, S., Koltsova, O., & Nikolenko, S. (2014, June). Latent 225odelling allocation: stability and applications to studies of user-generated content. In Proceedings of the 2014 ACM conference on Web science (pp. 161-165).

Komljenovic, J. (2019). Linkedin, platforming labour, and the new employability mandate for universities, Globalisation, *Societies and Education, 17*(1), 28-43.

Kosinski, M., Matz, S., Gosling, S. D., Popov, V., & Stillwell, D. (2016). A look at how to recruit participants using Facebook—and the ethical concerns that come with social media research. *Monitor on psychology*, 70-75.

Kotenko, I., Sharma, Y., & Branitskiy, A. (2021, September). Predicting the Mental State of the Social Network Users based on the Latent Dirichlet Allocation and fastText. In *2021 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)* (Vol. 1, pp. 191-195). IEEE

Kubler, R. V., Colicev, A., & Pauwels, K. H. (2020). Social Media's Impact on the Consumer Mindset: When to Use Which Sentiment Extraction Tool?. *Journal of Interactive Marketing, 50*, 136-155.

Kumar, V., & Nanda, P. (2019). Social Media in Higher Education: A Framework for Continuous Engagement. *International Journal of Information and Communication Technology Education, 15*(1), DOI: 10.4018/IJICTE.2019010107

Kuoppakangas, P., Suomi, K., Clark, P., Chapleo, C., & Stenvall, J. (2020). Dilemmas in Re-branding a University—"Maybe People Just Don't Like Change": Linking Meaningfulness and Mutuality into the Reconciliation. *Corporate Reputation Review, 23*, 92–105.

Kwahk, KY., & Kim, B. (2017). Effects of social media on consumers' purchase decisions: evidence from Taobao. *Service Business, 11*, 803–829.

Labrecque, L. I., Swani, K., & Stephen, A. T. (2020). The impact of pronoun choices on consumer engagement actions: Exploring top global brands' social media communications. *Psychology & Marketing*, *37*(6), 796-814.

Lange, K. (1995). Applications of the Dirichlet distribution to forensic match probabilities. *Genetica, 96*(1), 107-117.

LaPiere, R. T. (1934). Attitudes vs. actions. *Social forces*, *13*(2), 230-237.

Law, D. S. (2016). Constitutional archetypes. *Texas Law Review*, *95*, 153.

Le, T. D., Dobele, A. R., & Robinson, L. J. (2019). Information sought by prospective students from social media electronic word-of-mouth during the university choice process. *Journal of Higher Education Policy and Management, 41*(1), 18-34.

Le, T. D., Robinson, L. J., & Dobele, A. R. (2020). Understanding high school students use of choice factors and word-of-mouth information sources in university selection. *Studies in Higher Education*, *45*(4), 808-818.

Lee, D., Ng, P., & Bogomolova, S. (2020). The impact of university brand identification and eWOM behaviour on students' psychological well-being: a multi- group analysis among active and passive social media users, Journal of Marketing Management, 36(3-4), 384-403.

Lee, I., & Tipoe, E. (2022). Global COVID-19 vaccine inequity: Preferences for overseas vaccine donations over booster shots. *Brain, behavior, & immunity-health*, *21*, 100447.

Lee, J., & Hong, I. B. (2016). Predicting positive user responses to social media advertising: The roles of emotional appeal, informativeness, and creativity. *International Journal of Information Management, 36(*3), 360-373.

Lee, K. H. (2019). "I Post, therefore I Become #cosmopolitan": The materiality of online representations of study abroad in China. Population, Space and Place, 26:e2297, DOI: 10.1002/psp.2297.

Leeson, W., Resnick, A., Alexander, D., & Rovers J. (2019). Natural Language Processing (NLP) in qualitative public health research: a proof of concept study. *International Journal of Qualitative Methods, 18*, 1-9.

Leijerholt, U., Biedenbach, G., & Hultén, P. (2019). Branding in the public sector: a systematic literature review and directions for future research. *Journal of Brand Management, 26*, 126–140.
Lemon, L. L., & Hayes, J. (2020). Enhancing trustworthiness of qualitative findings: Using Leximancer for qualitative data analysis triangulation. *The Qualitative Report*, *25*(3), 604-614.
Leonard, R. (November, 14, 2018). Survey reveals Facebook in top spot for student recruitment, but Instagram catching up fast. Irish Tech News article on TerminalFour's 2018 Global Higher Education Digital Marketing & Web Survey. Retrieved from: https://irishtechnews.ie/survey-reveals-facebook-in-top-spot-for-student-recruitment-but-instagram-catching-up-fast/
Leong, C. M., Loi, A. M. W., & Woon, S. (2021). The influence of social media eWOM information on purchase intention. *Journal of Marketing Analytics*, 1-13.
Lever, J. (1981). Multiple methods of data collection: A note on divergence. *Urban Life*, *10*(2), 199-213.
Leydesdorff, L. (2005) *Similarity Measures, Author Cocitation Analysis,and Information Theory*. In: JASIST 56(7), pp.769-772.
Li, F. (2020). Factors influencing Chinese students' choice of an international branch campus: A case study. *Journal of Studies in International Education*, *24*(3), 337-351.
Li, W., & McCallum, A. (2006). Pachinko allocation: DAG – structured mixture models of topic correlations. In International Conference on Machine Learning (ICML), 577–584.
Li, Y., Rapkin, B., Atkinson, T. M., Schofield, E., & Bochner, B. H. (2019). Leveraging Latent Dirichlet Allocation in processing free-text personal goals among patients undergoing bladder cancer surgery. *Quality of Life Research*, *28*(6), 1441-1455.
Lim, K. K., & Lee, C. S. (2021). Investigating Learner's Online Learning Behavioural Changes during the COVID-19 Pandemic. *Proceedings of the Association for Information Science and Technology*, *58*(1), 777-779.
Lindahl, L., & Regnér, H. (2005). College choice and subsequent earnings: Results using Swedish sibling data. *Scandinavian Journal of Economics*, *107*(3), 437-457.
Linneberg, M., & Korsgaard, S. (2019), Coding qualitative data: a synthesis guiding the novice, *Qualitative Research Journal, 19*(3), 259-270.
Liu, L., Tang, L., Dong, W. et al. (2016). An overview of topic 226odelling and its current applications in bioinformatics. *SpringerPlus 5*, 1608. https://doi.org/10.1186/s40064-016-3252-8
Liu, S. (2020, Jul, 23). Speech-to-Text transcript accuracy rate among leading companies worldwide in 2020, Statista 2021. https://www.statista.com/statistics/1133833/speech-to-text-transcript-accuracy-rate-among-leading-companies/
Liu, T., Giorgi, S., Yadeta, K., Schwartz, H. A., Ungar, L. H., & Curtis, B. (2022). Linguistic predictors from Facebook postings of substance use disorder treatment retention versus discontinuation. *The American Journal of Drug and Alcohol Abuse*, 1-13.
Lombard, E. (2012). The role of the academic library in college choice. *The journal of academic librarianship*, *38*(4), 237-241.
Lomer, S., Papatsiba, V., & Naidoo, R. (2018). Constructing a national higher education brand for the UK: positional competition and promised capitals. *Studies in Higher Education, 43*(1), 134-153.
López-Bonilla, J. M., Barrera Barrera, R., Rodríguez Serrano, M. Á., López-Bonilla, L. M., Palacios Florencio, B., Reyes Rodríguez, M. C., & Sanz Altamira, B. (2012). Reasons which influence on the students' decision to take a university course: differences by gender and degree. *Educational Studies*, *38*(3), 297-308.
Lopez, M., & Sicilia, M. (2014). eWOM as Source of Influence: The Impact of Participation in eWOM and Perceived Source Trustworthiness on Decision Making. Journal of Interactive Advertising, 14(2), 86-97.
Lossio-Ventura, J. A., Gonzales, S., Morzan, J., Alatrista-Salas, H., Hernandez-Boussard, T., & Bian, J. (2021). Evaluation of clustering and topic 226odelling methods over health-related tweets and emails. *Artificial Intelligence in Medicine*, *117*, 102096.
Lowe, J. (1993). Commercialization of university research: A policy perspective. *Technology Analysis & Strategic Management, 5*(1), 27-37.
Lowrie, A. (2007). Branding higher education: Equivalence and difference in developing identity. *Journal of Business Research, 60*(9), 990-999.
Lu, Q. S., & Miller, R. (2019). How Social Media Communications Combine with Customer Loyalty Management to Boost Green Retail Sales. *Journal of Interactive Marketing, 46*, 87-100.
Lund, B. (2019). Universities engaging social media users: an investigation of quantitative relationships between universities' Facebook followers/interactions and university attributes. *Journal of Marketing for Higher Education, 29(*2), 251-267.
Lund, N. F., Cohen, S. A., & Scarles, C. (2018). The power of social media storytelling in destination branding. *Journal of Destination Marketing & Management, 8*, 271-280.

Mai Thi Ngoc, D., & Thorpe, A. (2015). What factors influence Vietnamese students' choice of university? *The International Journal of Educational Management, 29*(5), 666-681.
Majid, S., Lopez, C., Megicks, P., & Lim, W. M. (2019). Developing effective social media messages: Insights from an exploratory study of industry experts. *Psychology & Marketing, 36(*6), 551-564.
Malhotra, A., Malhotra, C. K., & See, A. (2013). How to create brand engagement on Facebook. *MIT Sloan Management Review*, *54*(2), 18-20.
Mampaey, J., Huisman, J., & Seeber, M. (2015). Branding of Flemish higher education institutions: a strategic balance perspective, *Higher Education Research & Development, 34*(6), 1178-1191.
Mampaey, J., Schtemberg, V., Schijns, J., Huisman, J., & Wæraas A. (2020). Internal branding in higher education: dialectical tensions underlying the discursive legitimation of a new brand of student diversity, *Higher Education Research & Development, 39*(2), 230-243. https://doi.org/10.1080/07294360.2019.1674252
Mandravickaitė, J., Briedienė, M., Uus, J., & Krilavičius, T. (2020). What's in the news? Identification of trending topics in alternative and mainstream Lithuanian media. In *CEUR Workshop proceedings [electronic resource]: Machine learning for trend and weak signal detection in social networks and social media, Toulouse, France, February 27-28, 2020: proceedings. Aachen: CEUR-WS, 2020, Vol. 2606*.
Marengo, D., Azucar, D., Giannotta, F., Basile, V., & Settanni, M. (2019). Exploring the association between problem drinking and language use on Facebook in young adults. *Heliyon*, *5*(10), e02523.
Maringe, F. (2006). University and course choice: Implications for positioning, recruitment and marketing. *International journal of educational management*, *20*(6), 466-479.
Marshall, B., Cardon, P., Poddar, A., & Fontenot, R. (2013). Does sample size matter in qualitative research?: A review of qualitative interviews in IS research. *The Journal of Computer Information Systems, 54(*1), 11-22.
Martins, E. A., Salles, I., Benevenuto, F., & Goussevskaia, O. (2022, June). Characterizing Sponsored Content in Facebook and Instagram. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media* (pp. 52-63)
Masele, J. J., & Rwehikiza, D. P. (2021). Applications of Social Media for Promoting Higher Learning Institutions' Activities in Tanzania. *International Journal of Education and Development using Information and Communication Technology, 17(*2), 37-54.
Massoud, H., & Ayoubi, R. M. (2019). Do flexible admission systems affect student enrollment? Evidence from UK universities. *Journal of Marketing for Higher Education*, *29*(1), 84-101.
Matic, J. L. (2019). The internationalisation of American higher education: a positional competition perspective. Globalisation, *Societies and Education, 17*(3), 274-284.
Mazzarol, T., & Soutar, G. N. (2002). "Push-pull" factors influencing international student destination choice. *International journal of educational management, 16*, (2/3), 82-90.
McDonald, M. (2019). Teaching Australian foreign policy: vocational training or critical thinking? *Australian Journal of International Affairs, 73*(6), 519-524.
McFadden, C., Maahs-Fladung, C., & Mallett, W. (2015). Recruiting international students to your campus. *Journal of International Students 2012 Vol 2 Issue 2*, 157.
McLeay, F., Lichy, J., & Asaad, A. (2020). Insights for a post-Brexit era: marketing the UK as a study destination – an analysis of Arab, Chinese, and Indian student choices, *Journal of Strategic Marketing, 28*(2), 161-175.
Melchiorre, M. M., & Johnson, S. A. (2017). Finding New Ways to Reach Older Students: Creating a Social Media Marketing Plan for Professional and Continuing Higher Education Programs, The Journal of Continuing Higher Education, 65(2), 73-81.
Melewar, T. C., Foroudi, P., Dinnie, K., & Nguyen, B. (2018). The role of corporate identity management in the higher education sector: an exploratory case study. Journal of Marketing Communications, 24(4), 337-359.
Mertens, A., Pradel, F., Rozyjumayeva, A., & Wäckerle, J. (2019, June). As the Tweet, so the Reply? Gender Bias in Digital Communication with Politicians. In *Proceedings of the 10th ACM Conference on Web Science* (pp. 193-201).
Mickelsson, J., van Haren, J.J.G.M., & Lemmink, J.G.A.M. (2022). Wrinkles in a CSR story: mismatched agendas in fast fashion service brands' CSR reputation", Journal of Service Management, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/JOSM-07-2021-0243
Mishler, A., Crabb, E. S., Paletz, S., Hefright, B., & Golonka, E. (2015, August). Using structural topic 227odelling to detect events and cluster Twitter users in the Ukrainian crisis. In *International conference on human-computer interaction* (pp. 639-644). Springer, Cham.
Mishra, S., & Madaan, R. (2020). Role of Online Marketing in University Admissions: Indian Students' Perspective. *International Journal of Marketing and Business Communication, 9*(3-4), 21-32.

Misirlis, N., & Vlachopoulou, M. (2018). Social media metrics and analytics in marketing – S3M: A mapping literature review. *International Journal of Information Management, 38*(1), 270-276.
Moeck, E. K., Bridgland, V. M., & Takarangi, M. K. (2022). Food for thought: Commentary on Burnette et al.(2021)"Concerns and recommendations for using Amazon Mturk for eating disorder research". *International Journal of Eating Disorders*, *55*(2), 282-284.
Moessner, M., Feldhege, J., Wolf, M., & Bauer, S. (2018). Analyzing big data in social media: Text and network analyses of an eating disorder forum. *International Journal of Eating Disorders*, *51*(7), 656-667.
Montenegro, C., Ligutom III, C., Orio, J. V., & Ramacho, D. A. M. (2018, May). Using latent 228odelling allocation for topic 228odelling and document clustering of 228odelling city twitter dataset. In *Proceedings of the 2018 International Conference on computing and data engineering* (pp. 1-5).
Moogan, Y. J., Baron, S., & Harris, K. (2002). Decision-Making Behaviour of Potential Higher Education Students. *Higher Education Quarterly, 53*(3), 211-228.
Moore, S. G., & Lafreniere, K. C. (2020). How online word-of-mouth impacts receivers. *Consumer Psychology Review*, *3*(1), 34-59.
Moran-Ellis, J., Alexander, V. D., Cronin, A., Dickinson, M., Fielding, J., Sleney, J., & Thomas, H. (2006). Triangulation and integration: processes, claims and implications. *Qualitative research*, *6*(1), 45-59.
Moran, R. F. (1998). Diversity, Distance, and the Delivery of Higher Education, *Ohio State Law Journal, 59*, 775-792.
Morse, J. M. (2015). Critical Analysis of Strategies for Determining Rigor in Qualitative Inquiry. *Qualitative health research, 25*(9), 1212-1222.
Muninger, M., Hammedi, W., & Mahr, D. (2020). The value of social media for innovation: A capability perspective. *Journal of Business Research, 95*, 116-127. https://doi.org/10.1016/j.jbusres.2018.10.012
Murray, K. B. (1991). A test of services marketing theory: consumer information acquisition activities. *Journal of Marketing, 55*(1), 10-25.
Musselin, C. (2018). New forms of competition in higher education. Socio-economic Review, 16(3), 657-683.
Mwenda, A. B., Sullivan, M., & Grand, A, (2019). How do Australian universities market STEM courses in YouTube videos?, *Journal of Marketing for Higher Education, 29*(2), 191-208.
Nadarzynski, T., Miles, O., Cowie, A., & Ridge, D. (2019). Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. *Digital Health, 5*. DOI: 10.1177/2055207619871808
Naidoo, R. (2018). The competition fetish in higher education: Shamans, mind snares and consequences. *European Educational Research Journal, 17*(5), 605-620.
Najimudinova, S., İsmailova, R., & Oskonbaeva, Z. (2022). What Defines the University Choice? The Case of Higher Education in Kyrgyzstan. *Sosyoekonomi*, *30*(54), 53-72.
Natale, S. M., & Doran, C. (2012). Marketization of Education: An Ethical Dilemma. *Journal of Business Ethics, 105*, 187–196.
Neagu, A. M., Pavaloiu, B. I., Guica, I. R., Dragoi, G., & Mateescu, L. M. (2020). The use of social media marketing in higher education enrolment: a Romanian perspective. *eLearning & Software for Education, 1*, 562-567.
Newman, M. W., Lauterbach, D., Munson, S. A., Resnick, P., & Morris, M. E. (2011, March). It's not that I don't have problems, I'm just not putting them on Facebook: challenges and opportunities in using online social networks for health. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work* (pp. 341-350).
Nguyen, B., Melewar, T. C., & Hemsley-Brown, J. (Eds). (2019). Strategic Brand Management in Higher Education. Routledge.
Nguyen, D. (2021). The Network Life of Non-biomedical Knowledge: Mapping Vietnamese Traditional Medicine Discourses on Facebook. *Journal of Digital Social Research*, *3*(2), 10-43.
Nguyen, T., Boyd-Graber, J., Lund, J., Seppi, K., & Ringger, E. (2015). Is your anchor going up or down? Fast and accurate supervised topic models. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 746-755)
Nikita, M., & Chaney, N. (October 13, 2022). Package 'ldatuning': Tuning of the Latent Dirichlet Allocation Models Parameters. https://cran.r-project.org/web/packages/ldatuning/ldatuning.pdf
Nisar, T. M., Prabhakar, G., & Strakova, L. (2019). Social media information benefits, knowledge management and smart organizations. *Journal of Business Research, 94*, 264-272.
Noble, C., O'Brien, M., Coombes, I., Shaw, P. N., & Nissen, L. (2011). Concept mapping to evaluate an undergraduate pharmacy curriculum. *American journal of pharmaceutical education*, *75*(3).
Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education*, *15*, 625-632.

Nurnberg, P., Schapiro, M., & Zimmerman, D. (2012). Students choosing colleges: Understanding the matriculation decision at a highly selective private institution. *Economics of Education Review, 31(*1), 1-8.
Nyangau, J., & Bado, N. (2012). Social Media and Marketing of Higher Education: A Review of the Literature. *Journal of the Research Center for Educational Technology, 8*(1).
Obermeit, K. (2012). Students' choice of universities in Germany: structure, factors and information sources used. *Journal of Marketing for higher education*, *22*(2), 206-230.
Odendaal, W., Atkins, S., & Lewin, S. (2016). Multiple and mixed methods in formative evaluation: Is more better? Reflections from a South African study. *BMC medical research methodology*, *16*(1), 1-12.
Oladipo, O. A., & Sugandi, B. (2021). Recruitment and mobility of international students: spotlight on a Chinese university, Globalisation, Societies and Education. DOI: 10.1080/14767724.2021.1988522
Opdenakker, R. (2006). Advantages and Disadvantages of Four Interview Techniques in Qualitative Research. *Forum: Qualitative Social Research, 7*(4).
Orlandi, L. B., Zardini, A., & Rossignoli, C. (2020). Organizational technological opportunism and social media: The deployment of social media analytics to sense and respond to technological discontinuities. *Journal of Business Research, 112,* 385-395.
Ostrowski, D. A. (2015, February). Using latent 229odelling allocation for topic modelling in twitter. In *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)* (pp. 493-497). IEEE.
Paladan, N. N. (2018). Higher Education Institutions Embracing Digital & Social Media Marketing: A Case of Top 25 Universities in Asia & Africa. *Marketing and Branding Research, 5*(3), 159-167.
Palan, S., & Schitter, C. (2018). Prolific. Ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22-27.
Palmer, S. (2013). Characterisation of the use of Twitter by Australian Universities, *Journal of Higher Education Policy and Management, 35(*4), 333-344.
Palmer, S. (2014). Characterizing Twitter communication – a case study of international engineering academic units, *Journal of Marketing for Higher Education, 24*(2), 257-273.
Pampaloni, A. M. (2010). The influence of organizational image on college selection: what students seek in institutions of higher education. *Journal of marketing for higher education*, *20*(1), 19-48
Panda, S., Pandey, S. C., & Bennett, A. (2019). University brand image as competitive advantage: a two-country study. *International Journal of Educational Management, 33*(2), 234-251.
Pardo, C., Pagani, M., & Savinien, J. (2022). The strategic role of social media in business-to-business contexts. *Industrial Marketing Management*, *101*, 82-97
Park, H., Park, T., & Lee, Y. S. (2019). Partially collapsed Gibbs sampling for latent Dirichlet allocation. *Expert Systems with Applications, 131*, 208-218.
Park, S., Choi, J., Lee, S., Oh, C., Kim, C., La, S., Lee, J., & Suh, B. (2019). Designing a Chatbot for a Brief Motivational Interview on Stress Management: Qualitative Case Study. *Journal of Medical Internet Research, 21*(4): e12231, DOI: 10.2196/12231.
Parkinson, J., De Laile, S., & Franco-Guillén, N. (2022). Mapping deliberative systems with big data: The case of the Scottish independence referendum. *Political Studies*, *70*(3), 543-565.
Pasternak, R. (2005). Choice of institutions of higher education and academic expectations: the impact of cost-benefit factors. *Teaching in Higher Education*, *10*(2), 189-201.
Paul, R., & Pradhan, S. (2019). Achieving Student Satisfaction and Student Loyalty in Higher Education: A Focus on Service Value Dimensions. *Services Marketing Quarterly, 40*(3), 245-268.
Paulsen, M. B., & John, E. P. S. (2002). Social class and college costs: Examining the financial nexus between college choice and persistence. *The Journal of Higher Education, 73*(2), 189-236.
Pavlik, A. (2021). Revamp your prospective transfer student programming with these strategies. *Enrollment Management Report, 25*(5), 4-5. DOI: 10.1002/emt
Payne, G., Williams, M., & Chamberlain, S. (2004). Methodological pluralism in British sociology. *Sociology*, *38*(1), 153-163.
Pedersen, E. R., & Kurz, J. (2016). Using Facebook for health-related research study recruitment and program delivery. *Current opinion in psychology*, *9*, 38-43.
Perkio, J., Buntine, W., & Perttu, S. (2004, September). Exploring independent trends in a topic-based search engine. In IEEE/WIC/ACM International Conference on Web Intelligence (WI'04) (pp. 664-668). IEEE.
Perry, R. F., & Rumpf, D. L. (1984). Predicting the likelihood of matriculation for college applicants. *Research in Higher Education, 21*(3), 317-328.
Peruta, A., & Shields, A. B. (2017). Social media in higher education: understanding how colleges and universities use Facebook, *Journal of Marketing for Higher Education, 27*(1), 131-143.

Peruta, A., & Shields, A. B. (2018). Marketing your university on social media: a content analysis of Facebook post types and formats. *Journal of Marketing for Higher Education, 28*(2), 175-191.
Phang, Y. C., Kassim, A. M., & Mangantig, E. (2021). Concerns of thalassemia patients, carriers, and their caregivers in 230odellin: text mining information shared on social media. *Healthcare Informatics Research*, *27*(3), 200-213.
Pornsrimate, K., & Khamwon, A. (2021). How to convert millennial consumers to brand evangelists through social media micro-influencers. *Innovative Marketing*, *17*(2).
Power, D. J., & Phillips-Wren, G. (2012). Impact of Social Media and Web 2.0 on Decision-Making. *Journal of Decision Systems, 20*(3), 249-261.
Prabowo, H., Bramulya, R., & Yuniarty, Y. (2020). Student purchase intention in higher education sector: The role of social network marketing and student engagement. *Management Science Letters, 10*(1), 103-110.
Prasad, S., Garg, A., & Prasad, S. (2019). Purchase decision of generation Y in an online environment. *Marketing Intelligence & Planning, 37*(4), 372-385.
Price, I. F., Matzdorf, F., Smith, L., & Agahi, H. (2003). The impact of facilities on student choice of university. *Facilities*, *21*(10), 212-222.
Pringle, J., & Fritz, S. (2019). The university brand and social media: using data analytics to assess brand authenticity. *Journal of Marketing for Higher Education, 29*(1), 9–44.
Prolific (2022, Sep). Prolific: A higher standard of online research. https://www.prolific.co
Qian, S., Zhang, T., & Xu, C. (2014, August). Boosted multi-modal supervised latent Dirichlet allocation for social event classification. In *2014 22nd International Conference on Pattern Recognition* (pp. 1999-2004). IEEE
Qiang, S., Wang, Y., & Jin, Y. (2017, July). A local-global LDA model for discovering geographical topics from social media. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data*(pp. 27-40). Springer, Cham
Qomariyah, S., Iriawan, N., & Fithriasari, K. (2019, December). Topic 230odelling twitter data using latent 230odelling allocation and latent semantic analysis. In *AIP conference proceedings* (Vol. 2194, No. 1, p. 020093). AIP Publishing LLC.
Quijada, M., Munoz, E., Corrons, A., & Olmo-Arriaga, J. (2021). Engaging students through social media. Findings for the top five universities in the world, *Journal of Marketing for Higher Education*, DOI: 10.1080/08841241.2020.1841069
Rapp, A., Beitelspacher, L. S., Grewal, D., & Hughes, D. E. (2013). Understanding social media effects across seller, retailer, and consumer interactions. *Journal of the Academy of Marketing Science*, *41*(5), 547-566.
Ravenda, D., Valencia-Silva, M. M., Argiles-Bosch, J. M., & García-Blandón, J. (2022). The Strategic Usage of Facebook by Local Governments: A Structural Topic Modelling Analysis. *Information & Management*, https://doi.org/10.1016/j.im.2022.103704.
Rehs, A. (2020). A structural topic model approach to scientific reorientation of economics and chemistry after German reunification. *Scientometrics*, *125*(2), 1229-1251.
Reich, J., Stewart, B., Mavon, K., & Tingley, D. (2016, April). The civic mission of MOOCs: Measuring engagement across political differences in forums. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale* (pp. 1-10).
Reich, J., Tingley, D. H., Jetson, L., Roberts, M., Stewart, B. M. (September 22, 2014). Computer-Assisted Reading and Discovery for Student Generated Text in Massive Open Online Courses. HarvardX Working Paper Series Number 6, Available at SSRN: https://ssrn.com/abstract=2499725 or http://dx.doi.org/10.2139/ssrn.2499725
Reisinger, J., Waters, A., Silverthorn, B., & Mooney, R. J. (2010, January). Spherical topic models. In International Conference on Machine Learning (ICML).
Rekhter, N., & Hossler, D. (2020). Russian Students' Use of Social Network Sites for Selecting Universities Abroad: Case Study at the Russian State University for the Humanities. *Journal of International Students, 10*(3).
Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V. A., & Boyd-Graber, J. (2015). Beyond LDA: exploring supervised topic 230odelling for depression-related language in Twitter. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality* (pp. 99-107).
Retamosa, M., Millan, A., & Garcia, J. A. (2022). The journey towards finding your favourite university. A segmentation study based on selection criteria. *Journal of Marketing for Higher Education, 1-25.*
Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). Stm: An R package for structural topic models. *Journal of Statistical Software*, *91*, 1-40.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., ... & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American journal of political science*, *58*(4), 1064-1082.
Rodriguez, M. (2021, April 28). Facebook revenue rises 48%, driven by higher-priced ads. CNBC Company Research. https://www.cnbc.com/2021/04/28/facebook-fb-earnings-q1-2021.html
Romano, S., Fucci, D., Scanniello, G., Baldassarre, M. T., Turhan, B., & Juristo, N. (2020). Researcher Bias in Software Engineering Experiments: a Qualitative Investigation. 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), DOI: 10.1109/SEAA51224.2020.00054.
Rosenthal, M. (2016). Qualitative research methods: Why, when, and how to conduct interviews and focus groups in pharmacy research. *Currents in Pharmacy Teaching and Learning, 8*(4), 509-516.
Rosenzweig, R. M. (1999). What's for Sale These Days in Higher Education: Two Stories, Research and Occasional Paper Series: CSHE.4.99, UC Berkeley: Center for Studies in Higher Education.
Rossetto, K. L. (2014). Qualitative research interviews: Assessing the therapeutic value and challenges. *Journal of Social and Personal Relationships, 31*(4), 482-489.
Royo-Vela, M., & Hunermund, U. (2016). Effects of inbound marketing communications on HEIs' brand equity: the mediating role of the student's decision-making process. An exploratory research, *Journal of Marketing for Higher Education, 26(*2), 143-167.
Rutter, R., Roper, S., & Lettice, F. (2016). Social media interaction, the university brand and recruitment performance. *Journal of Business Research, 69*(8), 3096-3104. https://doi.org/10.1016/j.jbusres.2016.01.025
Ryoo, J., & Bendle, N. (2017). Understanding the social media strategies of US primary candidates. *Journal of Political Marketing*, *16*(3-4), 244-266.
Sá, C., Florax, R. J., & Rietveld, P. (2012). Living arrangement and university choice of Dutch prospective students. *Regional Studies*, *46*(5), 651-667.
Sadler, G. R., Lee, H. C., Lim, R. S. H., & Fullerton, J. (2010). Recruitment of hard-to-reach population subgroups via adaptations of the snowball sampling strategy. *Nursing & health sciences*, *12*(3), 369-374.
Sagadevan, S., Malim, N. H. A. H., & Husin, M. H. (2022). A Seed-Guided Latent Dirichlet Allocation Approach to Predict the Personality of Online Users Using the PEN Model. *Algorithms*, *15*(3), 87.
Sagynbekova, S. et al. (2020). Social media communication and higher education brand equity: The mediating role of eWOM. Journal of Public Affairs, 21:e2112. DOI: 10.1002/pa.2112
Salzberg, S. L. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data mining and knowledge discovery*, *1*(3), 317-328.
Sánchez-Franco, M. J., Arenas-Márquez, F. J., & Alonso-Dos-Santos, M. (2021). Using structural topic modelling to predict users' sentiment towards intelligent personal agents. An application for Amazon's echo and Google Home. *Journal of Retailing and Consumer Services*, *63*, 102658
Sandelowski, M. (1995). Sample size in qualitative research. *Research in Nursing & Health, 18*(2), 179-183.
Sandlin, J. K., & Pena, E. V. (2014). Building Authenticity in Social Media Tools to Recruit Postsecondary Students. *Innovation in Higher Education, 39*, 333-346.
Sandvig, J. C. (2016). The role of social media in college recruiting. *International Journal of Web Based Communities, 12*(1), 23-34.
Sarkar D. (2016) Semantic and Sentiment Analysis. In: Text Analytics with Python. Apress, Berkeley, CA.
Sashittal, H. C., Sriramachandramurthy, R. & Hodis, M. (2012). Targeting college students on Facebook? How to stop wasting your money. *Business Horizons, 55*(5), 495-507.
Saunders, B., Sim, J., Kingstone, T., Baker, S., Waterfield, J., et al. (2018). Saturation in qualitative research: exploring its conceptualization and operationalization. *Quality and Quantity, 52*(4), 1893-1907.
Savage, J. (2012). *Creating brand evangelists in the 21st century: Using brand engagement through social media to develop brand loyalty in teens*. University of Southern California.
Savin, I., Drews, S., Van Den Bergh, J., & Villamayor-Tomas, S. (2022). Public expectations about the impact of COVID-19 on climate action by citizens and government. *PloS one*, *17*(6), https://doi.org/10.1371/journal.pone.0266979.
Schetgen, L., Bogaert, M., & Van den Poel, D. (2021). Predicting donation behavior: Acquisition 231odelling in the nonprofit sector using Facebook data. *Decision Support Systems*, *141*, 113446.
Schivinski, B., & Dabrowski, D. (2016). The effect of social media communication on consumer perceptions of brands. *Journal of Marketing Communications*, *22*(2), 189-214.
Schmierbach, M., & Oeldorf-Hirsch, A. (2012). A Little Bird Told Me, So I Didn't Believe It: Twitter, Credibility, and Issue Perceptions. *Communication Quarterly, 60*(3), 317-337.
Schultz, C. D. (2016). Insights from consumer interactions on a social networking site: Findings from six apparel retail brands. *Electronic Markets*, *26*(3), 203-217.

Schultze, U., & Avital, M. (2011). Designing interviews to generate rich data for information systems research. *Information and Organization, 21*(1), 1-16.
Schwieger, D., & Ladwig, C. (2018). Reaching and Retaining the Next Generation: Adapting to the Expectations of Gen Z in the Classroom. *Information Systems Education Journal, 16(*3), 45-54.
Scott, J., & Marshall, G. (Eds.). (2009). *A dictionary of sociology*. Oxford University Press, USA.
Senadheera, V., Warren, M., & Leitch, S. (2017). Social media as an information system: improving the technological agility. *Enterprise Information Systems, 11*(4), 512-533.
Shang, Z., & Luo, J. M. (2022). Topic modelling for wildlife tourism online reviews: analysis of quality factors. *Current Issues in Tourism*, 1-15.
Shanka, T., Quintal, V., & Taylor, R. (2006). Factors influencing international students' choice of an education destination–A correspondence analysis. *Journal of Marketing for Higher education*, *15*(2), 31-46.
Shields, A. B., & Peruta, A. (2019). Social media and the university decision. Do prospective students really care? *Journal of Marketing for Higher Education, 29*(1), 67-83.
Shields, R. (2016). Following the leader? Network models of 'world-class' universities on Twitter. *Higher Education, 71*, 253–268.
Sidaoui, K., Jaakkola, M., & Burton, J. (2020). AI feel you: customer experience assessment via chatbot interviews. *Journal of Service Management, 31*(4), 745-766.
Sieber, S. D. (1973). The integration of fieldwork and survey methods. *American journal of sociology*, *78*(6), 1335-1359.
Siegfried, J., & Getz, M. (2006). Where do the children of professors attend college?. *Economics of Education Review*, *25*(2), 201-210.
Sietsma, A. J., Ford, J. D., Callaghan, M. W., & Minx, J. C. (2021). Progress in climate change adaptation research. *Environmental Research Letters*, *16*(5), 054038.
Silva, C. C., Galster, M., & Gilson, F. (2021). Topic 232odelling in software engineering research. *Empirical Software Engineering, 26*(6), 120.
Silva, N. F. F. D., Silva, N. R. D., Cassiano, K. K., & Cordeiro, D. F. (2021). Inference of topics with Latent Dirichlet Allocation for Open Government Data. *Perspectivas em Ciência da Informação*, *26*, 57-79.
Silva, S. C., Duarte, P. A O., & Almeida, S. R. (2020). How companies evaluate the ROI of social media marketing programmes: insights from B2B and B2C. *The Journal of business & industrial marketing, 35(*12), 2097-2110.
Simoes, C., & Soares, A. M. (2010). Applying to higher education: information sources and choice factors. *Studies in Higher Education*, *35*(4), 371-389.
Singh, A., & Glińska-Neweś, A. (2022). Modeling the public attitude towards organic foods: a big data and text mining approach. *Journal of big Data*, *9*(1), 1-21
Smit, A., Swartz, L., Bantjes, J., Roomaney, R., & Coetzee, B. (2021). Moving beyond text-and-talk in qualitative health research: Methodological considerations of using multiple media for data collection. *Qualitative Health Research*, *31*(3), 600-614.
Smith, A. E., & Humphreys, M. S. (2006). Evaluation of unsupervised semantic mapping of natural language with Leximancer concept mapping. *Behavior research methods*, *38*, 262-279.
Smith, C. S. (2022, March 25). 8 tips on how to handle negative Facebook comments. How to manage negative postings on Facebook? https://martech.org/8-tactics-handling-negative-facebook-comments/
Smith, N., & Graham, T. (2019). Mapping the anti-vaccination movement on Facebook. *Information, Communication & Society*, *22*(9), 1310-1327.
Smith, R. J., Crutchley, P., Schwartz, H. A., Ungar, L., Shofer, F., Padrez, K. A., & Merchant, R. M. (2017). Variations in Facebook posting patterns across validated patient health conditions: a prospective cohort study. *Journal of medical Internet research*, *19*(1), e6486.
Soutar, G. N., & Turner, J. P. (2002). Students' preferences for university: a conjoint analysis. *International Journal of Educational Management, 16(*1), 40-45.
Steele, K. (2008). Classifying universities: Institutional brands from the market's perspective. Retrieved from: http://eduvation.ca/wp-content/uploads/classifying-universities.pdf.
Steffes, E., & Burgee, L. E. (2009). Social ties and online word of mouth. *Internet Research, 19*(1), 42-59.
Sterrett, D., Malato, D., Benz, J., Kantor, L., Tompson, T., et al. (2019). Who Shared It?: Deciding What News to Trust on Social Media. Digital Journalism, 7(6), 783-801.
Stieglitz, S., Dang-Xuan, L., Bruns, A. et al. (2014). Social Media Analytics. *Business & Information Systems Engineering, 6*, 89–96.
Stokes, D. C., Purtle, J., Meisel, Z. F., & Agarwal, A. K. (2021). State legislators' divergent social media response to the opioid epidemic from 2014 to 2019: longitudinal topic 232odelling analysis. *Journal of general internal medicine*, *36*(11), 3373-3382
Sugihartati, R., Suyanto, B., & Sirry, M. I. (2020). The shift from consumers to prosumers: Susceptibility of young adults to radicalization. *Social Sciences*, *9*(4), 40.

Sullivan, G. M. (2013). Analyzing and Interpreting Data From Likert-Type Scales. *Journal of Graduate Medical Education, 5*(4), 541–542.
Suresh, K. P. (2011). An overview of randomization techniques: an unbiased assessment of outcome in clinical research. *Journal of human reproductive sciences*, *4*(1), 8.
Syed, J., Johnson, A. T., & Ferguson, S. (2021). College choice & the consumer: the impact of gender on higher education 233odelling. *Journal of Marketing for Higher Education*, 1-21.
Szekeres, J. (2010). Sustaining student numbers in the competitive marketplace. *Journal of Higher Education Policy and Management*, *32*(5), 429-439.
Tafesse, W., & Wien, A. (2018a). Implementing social media marketing strategically: an empirical assessment. *Journal of Marketing Management, 34*(9-10), 732-749.
Tafesse, W., & Wien, A. (2018b). Using message strategy to drive consumer behavioral engagement on social media. *Journal of Consumer Marketing, 35*(3), 241-253.
Tai, C. H., Tan, Z. H., Lin, Y. S., & Chang, Y. S. (2015, October). Mental disorder detection and measurement using latent Dirichlet allocation and SentiWordNet. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*(pp. 1215-1220). IEEE.
Tan, A. (2015). College Choice Behaviors of International Students. SAGE Open, 5(4). Doi:10.1177/2158244015618995
Tang, C., Zhou, L., Plasek, J., Rozenblum, R., & Bates, D. (2017). Comment topic evolution on a cancer institution's Facebook page. *Applied clinical informatics*, *8*(03), 854-865.
Tas, A., & Ergin, E. A. (2012). Key factors for student recruitment: The issue of university branding. *International Business Research, 5*(10), 146.
Taylor, Z. W., & Bicak, I. (2020). Buying search, buying students: how elite U.S. institutions employ paid search to practice academic capitalism online. *Journal of Marketing for Higher Education, 30*(2), 271-296.
I, P. L., Piao, S., Almansour, M., Ong, H. F., & Ahad, A. (2022). Analysis of Popular Social Media Topics Regarding Plastic Pollution. *Sustainability*, *14*(3), 1709.
I, Y., Jordan, M., Beal, M., Blei, D. (2006) Hierarchical Dirichlet processes. *Journal of American Statistical Association, 101*(476), 1566–1581.
Teng, S., & Khong, K. W. (2015). An Exploratory Investigation of Study- Abroad Online Information Cues, *Journal of Teaching in International Business, 26*(3), 177-196. DOI: 10.1080/08975930.2015.1078268
Teng, S., Khong, K. W., & Chong, A. Y. (2015). Study abroad information in the new media, *Journal of Marketing for Higher Education, 25*(2), 263-286.
Terho, H., Giovannetti, M., & Cardinali, S. (2022). Measuring B2B social selling: Key activities, antecedents and performance outcomes. *Industrial Marketing Management, 101*, 208-222.
Thorson, K., Medeiros, M., Cotter, K., Chen, Y., Rodgers, K., Bae, A., & Baykaldi, S. (2020). Platform civics: Facebook in the local information infrastructure. *Digital Journalism*, *8*(10), 1231-1257
Tian, K., Revelle, M., & Poshyvanyk, D. (2009). Using Latent Dirichlet Allocation for automatic categorization of software, 2009 6th IEEE International Working Conference on Mining Software Repositories, 163-166.
To, W.M., J.W. Lung, Lai, L.S., & Lai. T.M. (2014). Destination choice of cross-border Chinese students: An importance-performance analysis. *Educational Studies. 40*(1): 63-80.
Tong, X., Li, Y., Li, J., Bei, R., & Zhang, L. (2022). What are People Talking about in# BackLivesMatter and# StopAsianHate? Exploring and Categorizing Twitter Topics Emerging in Online Social Movements through the Latent Dirichlet Allocation Model. *arXiv preprint arXiv:2205.14725*
Toubia, O., Iyengar, G., Bunnell, R., & Lemaire, A. (2019). Extracting features of entertainment products: A guided latent 233odelling allocation approach informed by the psychology of media consumption. *Journal of Marketing Research*, *56*(1), 18-36.
Toubia, O., Iyengar, G., Bunnell, R., & Lemaire, A. (2019). Extracting features of entertainment products: A guided latent 233odelling allocation approach informed by the psychology of media consumption. *Journal of Marketing Research*, *56*(1), 18-36.
Trow, M. (1957). Comment on" Participant Observation and Interviewing: A Comparison". *Human organization*, *16*(3), 33-35.
Turner, D. W. (2010). Qualitative Interview Design: A Practical Guide for Novice Investigators. *The Qualitative Report, 15*(3), 754-760.
Tuten, T. L., & Solomon, M. R. (2017). Social Media Marketing. SAGE.
Tvinnereim, E., & Flottum, K. (2015). Explaining topic prevalence in answers to open-ended survey questions about climate change. *Nature Climate Change*, *5*(8), 744-747.
Van Asperen, M., De Rooij, P., & Dijkmans, C. (2018). Engagement-based loyalty: The effects of social media engagement on customer loyalty in the travel industry. *International Journal of Hospitality & Tourism Administration*, *19*(1), 78-94.

Van der Vegt, I., Mozes, M., Gill, P., & Kleinberg, B. (2021). Online influence, offline violence: language use on YouTube surrounding the 'Unite the Right'rally. *Journal of computational social science*, *4*(1), 333-354.
Vasileiou, K., Barnett, J., Thorpe, S., & Young, T. (2018). Characterising and justifying sample size sufficiency in interview-based studies: systematic analysis of qualitative health research over a 15-year period. *BMC Medical Research Methodology, 18*, 148.
Vauterin, J. J., Linnanen, L., & Marttila, E. (2011). Issues of delivering quality customer service in a higher education environment. *International Journal of Quality and Service Sciences, 3*(2), 181-198.
Veletsianos, G., Kimmons, R., Shaw, A., Pasquini, L., & Woodward, S. (2017). Selective openness, branding, broadcasting, and promotion: Twitter use in Canada's public universities. *Educational Media International, 54*(1), 1-19.
Veloutsou, C., Lewis, J. W., & Paton, R. A. (2004). University selection: information requirements and importance. *International Journal of Educational Management*, *18*(3), 160-171
Villanueva, J., Yoo, S., & Hanssens, D. M. (2008). The Impact of Marketing-Induced versus Word-of-Mouth Customer Acquisition on Customer Equity Growth. *Journal of Marketing Research, 45*(1), 48-59.
Vrontis, D., El Nemar, S., Ouwaida, A., & Shams, S. R. (2018). The impact of social media on international student recruitment: the case of Lebanon. *Journal of International Education in Business, 11*(1), 79-103.
Wallace, L., Wilson, J., & Miloch, K. (2011). Sporting Facebook: A content analysis of NCAA organizational sport pages and Big 12 Conference Athletic department pages. *International Journal of Sport Communication, 4*, 422–444.
Wallach, H. M. (2006, June). Topic 234odelling: beyond bag-of-words. In Proceedings of the 23rd international conference on Machine learning (pp. 977-984).
Walsh, C., Moorhouse, J., Dunnett, A., & Barry, C. (2015). University choice: which attributes matter when you are paying the full price?. *International Journal of Consumer Studies, 39*(6), 670-681.
Walsh, G., & Mitchell, VW. (2010). The effect of consumer confusion proneness on word of mouth, trust, and customer satisfaction. European *Journal of Marketing, 44*(6), 838-859.
Wang, C., Blei, D. (2009). Decoupling sparsity and smoothness in the discrete hierarchical Dirichlet process. *Advances in Neural Information Processing Systems, 22*.
Wardati, N. K., & Mahendrawathi, E. R. (2019). The impact of social media usage on the sales process in small and medium enterprises (SMEs): A systematic literature review. *Procedia Computer Science*, *161*, 976-983.
Watermeyer, R., Crick, T., Knight, C. et al. (2021). COVID-19 and digital disruption in UK universities: afflictions and affordances of emergency online migration. *Higher Education, 81*, 623–641.
Wawrzuta, D., Jaworski, M., Gotlib, J., & Panczyk, M. (2021). Social Media Sharing of Articles About Measles in a European Context: Text Analysis Study. *Journal of medical Internet research*, *23*(11), e30150
Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences*. Chicago: Rand McNally.
Weeks, B. E., Ardevol-Abreu, A., & Gil de Zuniga, H. (2017). Online influence? Social media use, opinion leadership, and political persuasion. *International Journal of Public Opinion Research*, *29*(2), 214-239.
Weis, D., Willems, H. (2017). Aggregation, Validation, and Generalization of Qualitative Data – Methodological and Practical Research Strategies Illustrated by the Research Process of an empirically Based Typology. *Integrative Psychological and Behavioral Science, 51*, 223–243.
West, C., (2016). Going Social. International Educator: *International Enrollment Supplement*, 2-12.
Westbrook, R. A. (1987). Product/consumption-based affective responses and post-purchase processes *Journal of Marketing Research, 24*(3), 258-270.
Whiting, A., & Deshpande, A. (2016). Towards Greater Understanding of Social Media Marketing: A Review. *The Journal of Applied Business and Economics, 18*(4), 82-91.
Whitler, K. A. (2014, July 17). Why Word Of Mouth Marketing Is The Most Important Social Media. Forbes. https://www.forbes.com/sites/kimberlywhitler/2014/07/17/why-word-of-mouth-marketing-is-the-most-important-social-media/?sh=5203332b54a8
Wijenayake, P., Alahakoon, D., De Silva, D., & Kirigeeganage, S. (2021). Deep LSTM for Generating Brand Personalities Using Social Media: A Case Study from Higher Education Institutions. *International Journal of Computer and Communication Engineering*, *10*(1), 17-27.
Wilkins, S., & Huisman, J. (2011). International student destination choice: The influence of home campus experience on the decision to consider branch campuses. *Journal of Marketing for Higher education*, *21*(1), 61-83.
Wolf, A., & Jenkins, A. (2018). What's in a name? The impact of reputation and rankings on the teaching income of English universities. *Higher Education Quarterly*, *72*(4), 286-303.

Wright, L., Paul, E., Steptoe, A., & Fancourt, D. (2022). Facilitators and barriers to compliance with COVID-19 guidelines: a structural topic modelling analysis of free-text data from 17,500 UK adults. *BMC public health*, *22*(1), 1-22.
Wu, J., Fan, S., & Zhao, J. L. (2018). Community engagement and online word of mouth: An empirical investigation. *Information & Management, 55*(2), 258-270.
Wu, Y. Q., & Cheong, C. Y. M. (2021). Corporate branding of academic institutions: semiotic communication of logos and names. *Journal of Marketing for Higher Education,* DOI: 10.1080/08841241.2021.1989544
Wut, T. M., Xu, J., & Lee, S. W. (2022). Does University Ranking Matter? Choosing a University in the Digital Era. *Education Sciences*, *12*(4), 229.
Xie, W., Zhu, F., Jiang, J., Lim, E. P., & Wang, K. (2016). Topicsketch: Real-time bursty topic detection from twitter. *IEEE Transactions on Knowledge and Data Engineering*, *28*(8), 2216-2229.
Xue, J., Chen, J., Chen, C., Zheng, C., Li, S., & Zhu, T. (2020). Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic 235odelling on Twitter. *PloS one*, *15*(9), e0239441.
Yahia, I. B., Al-Neama, N., & Kerbache, L. (2018). Investigating the drivers for social commerce in social media platforms: Importance of trust, social support and the platform perceived usage. *Journal of Retailing and Consumer Services, 41*, 11-19.
Yan, Q., & Wu, S. (2018). How differences in eWOM platforms impact consumers' perceptions and decision-making. *Journal of Organizational Computing and Electronic Commerce, 28*(4), 315-333.
Yang, M. and Han, C. (2021), "Revealing industry challenge and business response to Covid-19: a text mining approach", *International Journal of Contemporary Hospitality Management, 33*(4), 1230-1248.
Yang, M. C., & Rim, H. C. (2014). Identifying interesting Twitter contents using topical analysis. *Expert Systems with Applications*, *41*(9), 4330-4336.
Yang, S., & Zhang, H. (2018). Text mining of Twitter data using a latent Dirichlet allocation topic model and sentiment analysis. *International Journal of Computer and Information Engineering*, *12*(7), 525-529.
Yoon, G., Li, C., Ji, Y., North, M., Hong, C., & Liu, J. (2018). Attracting Comments: Digital Engagement Metrics on Facebook and Financial Performance. *Journal of Advertising, 47*(1), 24-37.
Zailskaite-Jakste, L., & Kuvykaite, R. (2012). Implementation of Communication in Social Media by Promoting Studies at Higher Education Institutions. *Engineering Economics, 23*(2).
Zarra, T., Chiheb, R., Faizi, R., & El Afia, A. (2016). Topic Modeling and Sentiment Analysis in Facebook to Enhance Students' Learning. *Journal of Theoretical and Applied Information Technology*, *94*(1), 188.
Zelditch Jr, M. (1962). Some methodological problems of field studies. *American journal of Sociology*, *67*(5), 566-576.
Zeyu, L. (2019, October). Towards an Understanding of Online Extremism in Japan. In *IEEE/WIC/ACM International Conference on Web Intelligence-Companion Volume* (pp. 7-13).
Zhang, M., Guo, L., Hu, M., & Liu, W. (2017). Influence of customer engagement with company social networks on stickiness: Mediating effect of customer value creation. *International Journal of Information Management, 37*(3), 229-240.
Zhao, Y., Zhang, J., & Wu, M. (2019). Finding users' voice on social media: An investigation of online support groups for autism-affected users on Facebook. *International Journal of Environmental Research and Public Health*, *16*(23), 4804.
Zhou, S., Kan, P., Huang, Q., & Silbernagel, J. (2021). A guided latent Dirichlet allocation approach to investigate real-time latent topics of Twitter data during Hurricane Laura. *Journal of Information Science*, https://doi.org/10.1177/01655515211007724
Zhu, Y. (2019). Social media engagement and Chinese international student recruitment: understanding how UK HEIs use Weibo and WeChat, *Journal of Marketing for Higher Education, 29*(2), 173-190.
Ziang, X., Zhou, M., Liao, Q., Mark, G., Chi, C., Chen, W., & Yang, H. (2020). Tell Me About Yourself: Using an AI-Powered Chatbot to Conduct Conversational Surveys with Open-ended Questions. *ACM transactions on computer-human interaction, 27*(3), 1-37.
Zuorba, H. D., Olan, C. L. O., & Cantara, A. D. (2017, December). A framework for identifying excessive sadness in students through twitter and facebook in the Philippines. In *Proceedings of the International Conference on Bioinformatics Research and Applications 2017* (pp. 52-56).

Appendix A

Participant Information Statement

Form URL: https://forms.gle/NUv9sWD9Ztzy8G1N6

Research Study: [Understanding human decision-making via Social Media eWOM with an AI-led Chatbot] - Ethics Approval Project No: 2022/566

Professor Uri Gal (Responsible Researcher)
Department of Business Information Systems |Usyd Business School
Email: uri.gal@sydney.edu.au

Ilker Cingillioglu
Department of Business Information Systems |Usyd Business School
Email: ilker.cingillioglu@sydney.edu.au

Professor Artem Prokhorov
Department of Business Analytics |Usyd Business School
Email: artem.prokhorov@sydney.edu.au

---

1. What is this study about?

We are conducting a research study about decision-making factors in university selection via social media electronic word-of-mouth with an AI-led chatbot. Taking part in this study is voluntary.

Please read this sheet carefully and ask questions about anything that you don't understand or want to know more about.

2. Who is running the study?

The study is being carried out by the following researchers:

- Ilker Cingillioglu
  Department of Business Information Systems |Usyd Business School
- Professor Uri Gal
  Department of Business Information Systems |Usyd Business School
- Professor Artem Prokhorov
  Department of Business Analytics |Usyd Business School

Ilker Cingillioglu is conducting this study as the basis for the degree of PhD at The University of Sydney.

3. **Who can take part in the study?**
We are seeking current or former university students to take part in this study due to their experience in choosing a higher education institution (HEI). There is no limitation as to where or at which HEI a participant studied in the past or currently studies.

4. **What will the study involve for me?**

If you decide to take part in this study, you will be asked to interact with an AI-led chatbot (Sydn-e) in an interview-like chatbot (chat) survey format and provide information regarding what factors may shape your decision to choose a university.

Your interaction with Sydn-e is estimated to last for about 7-10 minutes. You can complete this chat-survey anytime at your convenience by no later than 15/5/2023. Sydn-e will ask you a few open-ended and mostly multiple-choice questions. You can use any device with Internet connection to take the chat-survey. There is no video or audio recording, so all your responses will be recorded on cloud in text form. You can provide a response to each question and complete the survey only once.

5. **Can I withdraw once I've started?**

Being in this study is completely voluntary and you do not have to take part.

Your decision will not affect your current or future relationship with the researchers or anyone else at The University of Sydney.

If you decide to take part in the study and then change your mind you can withdraw before or while interacting with Sydn-e by closing the chat window. By submitting your chat survey, you consent to take part in the study. You can withdraw any time before you submit however once your responses are submitted, they cannot be withdrawn. This is because they are anonymous, and we will not be able to tell which one yours is. If you decide to withdraw, we will not collect any more information from you. Any information that we have already collected will be kept in our study records due to real-time data flow into cloud, and they may be included in the study results.

6. **Are there any risks or costs?**

Aside from giving up your time, we do not expect that there will be any risks or costs associated with taking part in this study.

7. **Are there any benefits?**

You will experience chatting with a sophisticated AI-led chatbot and contribute to knowledge with your opinions on higher education.

8. **What will happen to information that is collected?**

By providing your consent, you are agreeing to us collecting information about you for the purposes of this study.

Any information you provide us will be stored securely and we will only disclose it with your permission, unless we are required by law to release information. We are planning for the study findings to be published. You will not be individually identifiable in these publications.

Study materials will be stored on Sydney University's Research Data Store (RDS) during and upon completion of the project. The RDS is a central networked drive maintained by the University of Sydney specifically for research data. It is a secure, enterprise-grade Network Attached Storage device. Only the responsible researcher will have access to it. After the approved retention period, the study materials will be disposed of via deletion of any electronic documents or data related to the research. The IT department will be involved to ensure the deletion is complete and files are not retrievable from any source. Any hard-copy paper materials will be shredded.

9. **Will I be told the results of the study?**

You have a right to receive feedback about the overall results of this study. If you choose to confirm in the Participant Consent Form (PCF) that you "would like feedback on the overall results of this study", you will need to provide your email address. This feedback will be in the form of a brief lay summary.

10. **What if I would like further information?**

When you have read this information, the following researcher/s will be available to discuss it with you further and answer any questions you may have:

Ilker Cingillioglu
Department of Business Information Systems |Usyd Business School
Email: ilker.cingillioglu@sydney.edu.au

11. **What if I have a complaint or any concerns?**

The ethical aspects of this study have been approved by the Human Research Ethics Committee (HREC) of The University of Sydney Project No: **2022/566** according to the *National Statement on Ethical Conduct in Human Research (2007).*

If you are concerned about the way this study is being conducted or you wish to make a complaint to someone independent from the study, please contact the University:

Human Ethics Manager
human.ethics@sydney.edu.au

+61 2 8627 8176

This information sheet is for you to keep

Appendix B

Participant Consent Form

Form URL: https://forms.gle/67UqPgghBGnWVd1k8

Research Study: [Understanding human decision-making via Social Media eWOM with an AI-led Chatbot] – Ethics Approval Project No: 2022/566

Ilker Cingillioglu
Department of Business Information Systems |Usyd Business School
Email: ilker.cingillioglu@sydney.edu.au

Professor Uri Gal (Responsible Researcher)
Department of Business Information Systems |Usyd Business School
Email: uri.gal@sydney.edu.au

Professor Artem Prokhorov
Department of Business Analytics |Usyd Business School
Email: artem.prokhorov@sydney.edu.au

Participant Name ____

I agree to take part in this research study. In giving my consent, I confirm that:

- The details of my involvement have been explained to me, and I have been provided with a written Participant Information Statement to keep.
- I understand the purpose of the study is to investigate decision-making factors in university selection via social media electronic word-of-mouth with an AI-led chatbot.
- I acknowledge that the risks and benefits of participating in this study have been explained to me to my satisfaction.
- I understand that in this study I will be required to interact with an AI-led chatbot [Sydn-e] and provide information regarding the factors that are associated with choosing a university.
- I understand that being in this study is completely voluntary.
- I am assured that my decision to participate will not have any impact on my relationship with the research team or the University of Sydney.

- I understand that I am free to withdraw from this study and that I can choose to withdraw any information I have already provided (unless the data has already been de-identified or published).

- I have been informed that the confidentiality of the information I provide will be protected and will only be used for purposes that I have agreed to. I understand that information about me will only be told to others with my permission, except as required by law.

- I understand that the results of this study may be published, and that publications will not contain my name or any identifiable information about me.

- Please confirm the following:

I consent to being contacted for future studies Yes ☐ No ☐

I consent to my data being used in future research Yes ☐ No ☐

I would like feedback on the overall results of this study Yes ☐ No ☐

If you answered **yes**, please provide your email address:

____

____

____

____

- I understand that after I sign and return this consent form it will be retained by the researcher, and that I may request a copy at any time.

Participant Name ____

Signature ____

Date ____

Appendix C

Chat Dialogue Items (including the Constant and IV Interventions)

Chatbot (C): Hi, my name is Sydn-E. I am an AI-powered chatbot developed by researchers at the University of Sydney Business School.
In case you haven't already, please check these out: Participant Information Statement and Participant Consent Form.
Thank you so much for taking part in this study.
I'm excited to chat with you.
What's your name?
P: … #to be recorded under $person; ## Only the name of the user is recognized and recorded even if the user enters a sentence (e.g. "my parents call me …") regardless of language (e.g. "Ich heisse …"); ### If no pretrained name is recognized, entity will be left blank.

C: Nice to meet you, $person.
Please enter your unique Prolific ID.
P: …

C: Are you currently studying at a higher education institution?
Yes, I am
No, but I intend to enrol in one
No, I was never enrolled in one and I don't intend to
No, but I was enrolled in one in the last 5 years
No, but I was enrolled in one more than 5 years ago

C: Great! Where is it? # Except for:
[No, but I intend to enrol in one][5] or
[No, I was never enrolled in one and I don't intend to][6]
Africa Asia Australia/NZ Canada Europe South America UK USA

C: When did you start studying there?
in 2023 in 2022 in 2021 in 2020 in 2019 in 2018 before 2018

C: Ok, I know some people use social media more than others, but I'm interested to know which social media platforms you normally use.

 # 3-second Pause

C: What is the social media platform you most frequently use?
P: … #to be recorded under $smp1 ## Only the name of a social media platform is recognized and recorded even if the user enters a sentence. ### Capable of fuzzy matching and disambiguation. #### If no pretrained social platform name is recognized, a retype prompt will follow within a while loop: "I am not familiar with this social media platform, please retype your most favorite social media platform!"

C: Got it. What is your second most frequently used social media platform?
P: … #to be recorded under $smp2 ## Only the name of a social media platform is recognized and recorded even if the user enters a sentence. ### Capable of fuzzy matching and disambiguation. #### If no pretrained social platform name is recognized, a retype prompt will follow within a while loop: "I am not familiar with this social media platform, please retype your second most favorite social media platform!"

C: Ok, now, please type this code: … #One of the 9 RA (Random allocation) codes (See Table 11)
P: …

[5] Dialogue skips "When did you start studying there?", then prompts: "Where do you intend to study?", then connects to the social media node (Ok, I know some people use social media more than others, but I'm interested to know which social media platforms you normally use).
[6] Dialogue continues in this case as: Ok, I see... I think there's been some sort of mistake.
This study is targeted at current, future or former higher education students.
I am sorry about this confusion.
Would you like to add anything before we say our Goodbyes? # Connect to final node

If an invalid code gets entered, the dialogue loops back to "Please enter the code: …"
When the code is entered correctly (not case or whitespace sensitive and capable of disambiguating & fuzzy matching and hence handling typos)

C: Ok. Great! Now, let me give you a scenario:
Imagine you are seeking information before enrolling in university and you've come across the following statements on a university's website:

If participant is in CTRL, then comes [[CONSTANT]],
Else, come [[CONSTANT]] + one of the IVs [[Intervention Dialogue of an EG – See Table 11]]

Constant:

"*Our university offers a range of opportunities for personal growth and professional development. We combine rich history and tradition with innovative scholarship and cutting-edge research. Our students create and apply knowledge by thinking and doing, preparing for leadership in a rapidly changing world."*

#10-second Pause

"Courses, taught by our esteemed faculty and enhanced by our unparalleled libraries and resources, will take you as far as your imagination allows. Here, you're going to be part of a community - one where everybody works hard, but that also takes a breather every now and then. In fact, the students who do best here already have some kind of outlet, such as theater, athletics or the arts."

#10-second Pause

Table 11: Group ids, random allocation codes, module (Constant and IV code), IV factors and intervention dialogues.

Group	RA Code	Module	IV Factor	Intervention Dialogue
CTRL	Ptc45	Constant	NA	NA
EG1	Mnk19	Constant + IV1	Reputation, image and ranking	In addition, imagine you read the following post about this University on social media: "The Times Higher Education ranked this University among the top universities in the world for a range of disciplines." *#10-second Pause* Moreover, you read this message about the same University on social media: "The University's faculty and research are world-renowned, as it has excellent reputation and image both nationally and internationally…"
EG2	Knr24	Constant + IV2	Living and study costs, and availability of scholarships	In addition, imagine you read the following post about this University on social media: "This University is quite affordable, and also known for its extensive scholarship program…" *#10-second Pause* Moreover, you read this message about the same University on social media: "I lived on and off campus whilst studying at this University and I must say it was much more affordable than many other places …"
EG3	Hpm38	Constant + IV3	Work and internship placements during study and job prospects upon graduation	In addition, imagine you read the following post about this University on social media: "This University helped me find a good internship while studying which led to my first full-time job at a reputable firm after graduation…" *#10-second Pause* Moreover, you read this message about the same University on social media: "I know for a fact that this University has a great career network, plenty of opportunities …"
EG4	Gwn42	Constant + IV4	Ease of admission, entrance requirements and open communication with admissions staff	In addition, imagine you read the following post about this University on social media: "My admission process at this University was fast and easy, and the entrance requirements were not hard to meet at all …" *#10-second Pause* Moreover, you read this message about the same University on social media: "I had a pleasant experience with the University's admissions staff: they were responsive and quick to guide me through the whole process…"
EG5	Bmr57	Constant + IV5	Campus location (proximity to home, convenience and comfort), safety and	In addition, imagine you read the following post about this University on social media: "This University is centrally located which is important to me because I can visit my parents anytime I want since home is not far away…" *#10-second Pause*

			physical appeal, and vibe of the city	Moreover, you read this message about the same University on social media: "I love the city and the campus because it is safe, conveniently located, vibrant and close to many attractions…"
EG6	Mha68	Constant + IV6	Availability, flexibility and attractiveness of the course and on-campus support services	In addition, imagine you read the following post about this University on social media: "The flexibility of the program I'm currently studying at this University suits my work-study-life balance, it is also quite relevant to my career aspirations …" *#10-second Pause* Moreover, you read this message about the same University on social media: "I am really happy with the availability of the courses and on-campus support I've received at this University…"
EG7	Ghw71	Constant + IV7	Prior knowledge of the study destination	In addition, imagine you read the following post about this University on social media: "It was a relief to be familiar with the city and I enjoy the benefits of knowing the place before I even started studying at this University …" *#10-second Pause* Moreover, you read this message about the same University on social media: "I could quickly adjust to the city because I'd lived there for a while before I enrolled …"
EG8	Yrk86	Constant + IV8	Collaboration with other universities	In addition, imagine you read the following post about this University on social media: "This University has research collaborations with many other universities all around the world." *#10-second Pause* Moreover, you read this message about the same University on social media: "Thanks to the University's student exchange arrangements, I can choose to study a whole year at a university in another country …"

CTRL: Control Group; EG: Experimental Group; IV: Independent Variable; RA: Random Allocation

C: Based on what you've read so far $person, how likely would you enroll in this Uni?
5: Very likely 4: Likely 3: Neutral 2: Unlikely 1: Very unlikely

C: Good to know. Thanks.
What other information about a university you read on $smp1 or $smp2 do you think would influence your choice to study there?
P: …

C: What information, coming from other sources than social media, would you consider important when choosing a university?
P: …

C: When you choose a university how important do you think its reputation, image and global ranking are?
5: Very important 4: Important 3: Neutral 2: Unimportant 1: Very unimportant

C: What about living and study costs, availability of scholarships etc.?
Which Uni would you prefer in terms of overall costs?
5: Highest priced. 4: Higher than average. 3: Average. 2: Lower than average. 1: Lowest priced

C: What about work and internship placements during study, job opportunities and potential work-related benefits after graduation? When you choose a university how important do you think this is?
5: Very important 4: Important 3: Neutral 2: Unimportant 1: Very unimportant

C: Alright, $person. I'm just curious to know some more …
When you choose a university how important is its ease of admission, considering its entrance requirements and open communication with admissions staff?
5: Very important 4: Important 3: Neutral 2: Unimportant 1: Very unimportant

C: What about the location of the campus?
When you choose a university how important are factors such as: its proximity to home, convenience and comfort, safety and physical appeal, and vibe of the city?
5: Very important 4: Important 3: Neutral 2: Unimportant 1: Very unimportant

C: How important to you are the flexibility and attractiveness of the course/program of study (in line with your career aspirations and earning potential) and on-campus support services?
5: Very important 4: Important 3: Neutral 2: Unimportant 1: Very unimportant

C: When you choose a university how important do you think whether you have prior knowledge of the study destination (city)?
5: Very important   4: Important   3: Neutral   2: Unimportant   1: Very unimportant

C: And finally, $person, when you choose a university how important to you is its collaboration with other universities?
5: Very important   4: Important   3: Neutral   2: Unimportant   1: Very unimportant

C: That's very helpful. Thanks a lot, $person.
By the way, how old are you?
P: …

C: Would you consider yourself a domestic or international student?
Domestic   International

C: Please click here [custom completion link to Prolific] and I'll let Prolific know that you've completed this study.
Would you like to add anything else before we say our goodbyes?
P: …

C: OK, then.
I really appreciate your help and learned a lot from our conversation.
Thank you!
Take care, $person bye …
P: …