Exploring Latent Topics in TVET using LDA Topic Modeling

Nur Hafazah Sharin
Mira Kartiwi
International Islamic University, Malaysia

Abstract

Social media contains vast amounts of textual data that can assist organizations in understanding their stakeholders better. To study public perceptions of Technical and Vocational Education and Training (TVET) in Malaysia, collecting data from social media is necessary. A total of 1,304 Facebook posts from the Ministry, news and media pages, and public groups were analyzed. Latent Dirichlet Allocation (LDA) topic modeling was utilized to uncover hidden themes by identifying the number of topics and associated keywords. With the highest coherence value of 0.4717, the analysis extracted eight (8) relevant themes regarding TVET. Ten (10) keywords from each topic help classify the TVET topic. The top three (3) topics that have been classified are skills/ competency, certification, and salary/ wage. By gaining the extracted topics, it would assist in decision-making and improve the TVET ecosystem.

Keywords: TVET, Topic Modeling, LDA, Coherence

INTRODUCTION

Technical and Vocational Education and Training (TVET) is often called as the "Game Changer" which aims to speed up the development of the country's human capital and create local talent that can fulfil market demand (Economic Planning Unit, 2015). In fact, TVET promotes the growth of human capital in preparation for industrialization. TVET helps a nation produce the highly skilled workers needed to spur economic growth (Saharudin et al., 2021).

Due to technological evolvement and growth of social media usage, people are intended to create a tendency to interact, share, and disseminate information ubiquitously. A report from the Digital 2022 January Global Overview stated that there are 4.62 billion social media users around the world. Millions of individuals voice their thoughts on services and products using social networking sites, blogs, and review sites. According to Razzaq et al. (2019), social media has become a highly rich source of information on people's behavioural states through reviews and comments. Active comments and feedback are extremely beneficial to businesses in terms of monitoring rivals, analyzing consumer satisfaction, as well as branding communication (Dhaoui et al., 2017; Salinca, 2016).

This study utilizes the baseline Latent Dirichlet Allocation (LDA) approach to uncover latent topics from Facebook postings pertaining to TVET issues in Malaysia. The outcome of the analysis is visually represented in order to observe the distribution of topics together with their corresponding terms inside each subject. Therefore, the identification of latent topics

can assist important stakeholders in understanding public attitudes and developing suitable policies and strategies to enhance the TVET ecosystem.

LITERATURE REVIEW

A vast amount of social media data collected from various platforms in different formats can now be efficiently and conveniently accessed while facilitating the acceleration of knowledge discovery (Kobayashi et al., 2018; Moe & Schweidel, 2017). Text mining is the process of finding or extracting useful information from text data by researching and identifying interesting patterns (Shinde & Govilkar, 2015). Topic modeling is a statistical tool for identifying the themes or topics in unstructured data. It helps in discovering the topic in a large collection of data. It is an unsupervised machine learning technique that classifies the documents based on the identified topics. Latent Dirichlet Allocation (LDA) is an unsupervised topic modeling method based on generative probabilistic modeling used to extract topics from a collection of unstructured data (Blei et al., 2003). Figure 1.1 illustrates the process of document generation in LDA. In this context, it can be observed that each individual node serves as a representation of a certain random variable. Furthermore, it is worth noting that the hyperparameters α and β are associated with the Dirichlet distribution of both ϕk and θd. The variable ϕk represents the distribution of words belonging to the kth topic, while θd corresponds to the distribution of words belonging to the dth document. The assignment of each word to its respective topic is determined by the parameter θd. The w is derived from the variables θd and z(d,n). The procedure for this generation is performed iteratively for a total of N words that are contained in M documents. The user is responsible for determining the values of α, β, and K in the LDA algorithm. The parameters ϕk and θd are learned by the LDA algorithm.

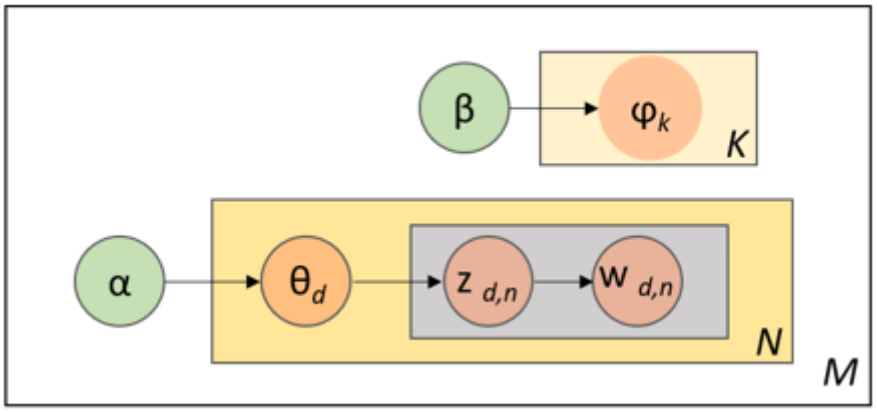

Figure 1.1: LDA graphical model

In LDA, the coherence value is primarily considered as a divided topic. The coherence of each topic is measured to assess whether the words in a topic are actually related to each other. The higher the coherence value, the better the interpretability of the model (Yin & Yuan, 2022). LDA is considered the most widely used topic modeling method.

Previous studies have employed LDA to uncover latent themes within social media posts across several domains, including healthcare, education, and disaster management. In their study, Phang et al. (2021) conducted an analysis of discussions pertaining to thalassemia inside Facebook groups. The researchers identified a total of eight (8) distinct topics that were discussed and the challenges encountered by thalassemia patients got the highest attention and engagement. In addition, the exploration of tweets on the topic COVID -19 was

investigated by Mathayomchan et al. (2022) using LDA. The study examined public discourse about countries during the COVID -19 pandemic and found twelve (12) topics. The top three topics discovered are military, political and economic. Nurmawiya & Harvian (2021) examined public opinion on education in the context of face-to-face events. Using the LDA model, relevant topics emerged, three (3) of which are vaccination, public preferences and reopening of schools. Meanwhile, Durham et al. (2023) revealed the key themes of disaster-related tweets by using LDA to understand the contextual information in social media posts during the disaster. Using social media data can help improve emergency management responses.

METHODOLOGY

This study gathered Facebook comments from news and media and ministries pages as well as public groups that discussed TVET Malaysia. It is conducted using Python programming language. The steps of this study consist of data collection, text preparation, text pre-processing, feature extraction, topic modeling with LDA and visualisation.

1. Data Collection

The data was collected from Facebook pages and public groups from April 2021 to March 2023. There are twelve (12) news and media pages, four (4) ministries, and three (3) public groups selected for crawling. The Malay and English data are collected from this platform.

2. Text Preparation

This step selects the crawled data by removing irrelevant posts such as politics, advertising, and promotional posts. Some corrections of spelling mistakes, short forms, and slang are made before the text are translated into English.

3. Text Pre-processing

The purpose of pre-processing is to streamline the data in preparation for subsequent stages. The process encompasses case folding, tokenization, removal of stop words, and lemmatization. The process of case folding entails transforming all text into lowercase format. Tokenization is a process employed to segment the given text into individual words. The process of eliminating stop words involves the removal of less significant words that are not directly relevant to a certain subject, such as pronouns (e.g., I, me, us, your, he, she). Lemmatization is the process through which a given word is transformed into its base root form.

4. Feature Extraction

In the process of feature extraction, the techniques employed include Term Frequency-Inverse Document Frequency (TF-IDF) and N-grams. The TF-IDF method is commonly employed as a weighting factor to assess the importance of a word inside a document in a given collection. An n-gram refers to a set of n successive words extracted from a certain text or sentence. The process of segmenting sentences into uni-, bi-, and tri-grams involves the use of N-grams, which are either individual words or phrases. An n-gram can be referred to as a unigram when it consists of a single word, a bigram when it consists of two words, and a trigram when it consists of three words.

5. Topic Modeling

The topic modeling step is conducted via the baseline LDA technique. The Gensim package was utilised to apply the LDA algorithm in order to identify potential latent topics throughout the entire dataset. The computation of coherence score is used to obtain the total number of themes. The coherence score is utilised by training various LDA models with varying numbers of topics chosen, and afterwards selecting the model that attains the highest coherence score.

6. Visualisation

The pyLDAvis library in Python is used to represent topics visually. The intertopic distance map is used to visually represent the arrangement of topics in a two-dimensional space. The area of these topic circles is proportional to the amount of words that belong to each topic across the dictionary.

DISCUSSION OF ANALYSIS AND FINDINGS

The determination of the number of topics is based on the acquisition of the highest coherence score. The concept of coherence is used to assess the quality of the data by examining the degree of semantic similarity between frequently occurring terms within a given topic. The coherence score of 0.4717 indicates the discovery of eight (8) topics, as depicted in Figure 1.2.

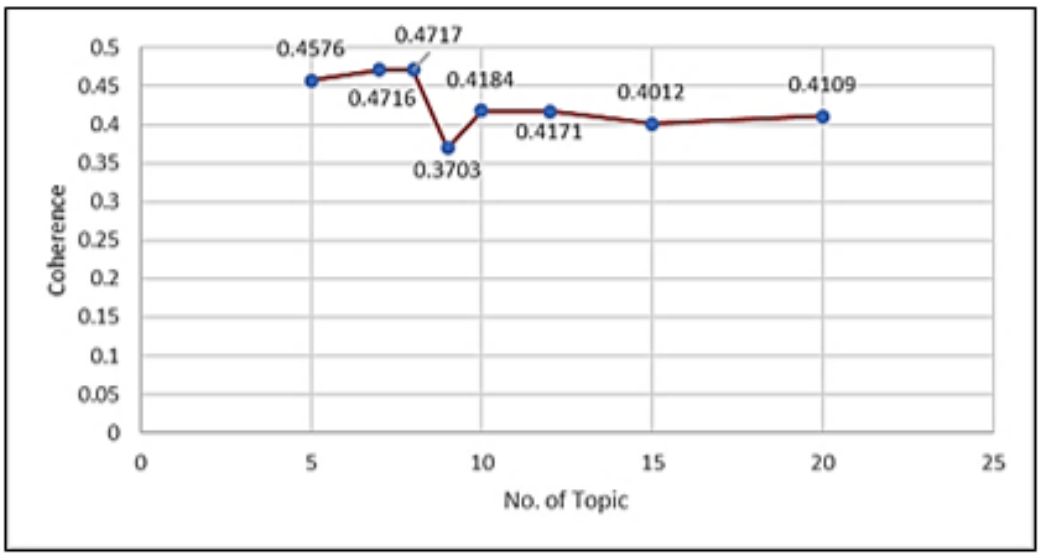

Figure 1.2: Coherence score with the number of topics

The LDA model identified the primary ten keywords for each topic, along with the corresponding distribution percentages, as presented in Table 1.1.

Table 1.1 : Top 10 keywords for each topic

Topic	Keywords
0	'0.148*"science" + 0.106*"agricultural" + 0.055*"student" + ' '0.042*"principal" + 0.028*"think" + 0.027*"interested" + 0.027*"right" + ' '0.025*"effort" + 0.025*"teacher" + 0.023*"put"'
1	'0.077*"people" + 0.064*"enter" + 0.040*"want" + 0.038*"know" + 0.033*"come" ' '+ 0.032*"many" + 0.032*"thing" + 0.032*"really" + 0.030*"ask" + ' '0.029*"lecturer"'
2	'0.122*"graduate" + 0.103*"salary" + 0.036*"university" + 0.035*"get" + ' '0.035*"pay" + 0.035*"well" + 0.033*"say" + 0.029*"set" + 0.025*"year" + ' '0.023*"time"'
3	'0.101*"tvet" + 0.081*"skill" + 0.046*"need" + 0.030*"school" + ' '0.029*"skilled" + 0.025*"empower" + 0.021*"minimum" + 0.020*"training" + ' '0.019*"level" + 0.018*"institution"'
4	'0.083*"student" + 0.032*"certificate" + 0.029*"accord" + 0.028*"study" + ' '0.027*"high" + 0.026*"area" + 0.025*"expensive" + 0.025*"university" + ' '0.025*"former" + 0.024*"mostly"'
5	'0.082*"employ" + 0.032*"various" + 0.029*"share" + 0.024*"involve" + ' '0.024*"true" + 0.020*"agency" + 0.017*"way" + 0.014*"exam" + 0.012*"party" ' '+ 0.012*"parent"'
6	'0.102*"good" + 0.079*"employer" + 0.040*"apply" + 0.038*"practice" + ' '0.037*"take" + 0.037*"interest" + 0.035*"stream" + 0.034*"maybe" + ' '0.033*"equivalent" + 0.021*"experience"'
7	'0.148*"class" + 0.125*"learn" + 0.059*"old" + 0.048*"site" + 0.048*"root" + ' '0.045*"lose" + 0.018*"expertise" + 0.016*"market" + 0.015*"provide" + ' '0.015*"day"'

The intertopic distance map effectively represented the distribution of topics and the prominent terms associated with the selected topics.

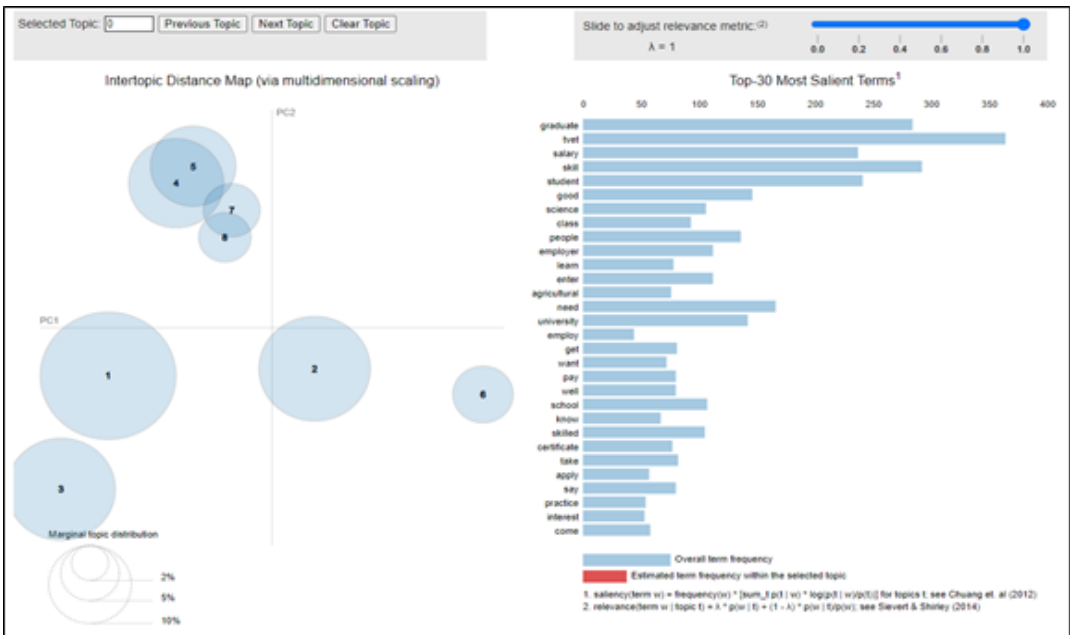

Figure 1.3: LDA Intertopic distance map (overall)

The size of a circle in the visual representation indicated the number of documents on a particular topic, with larger circles indicating a higher number of reviews within that topic.

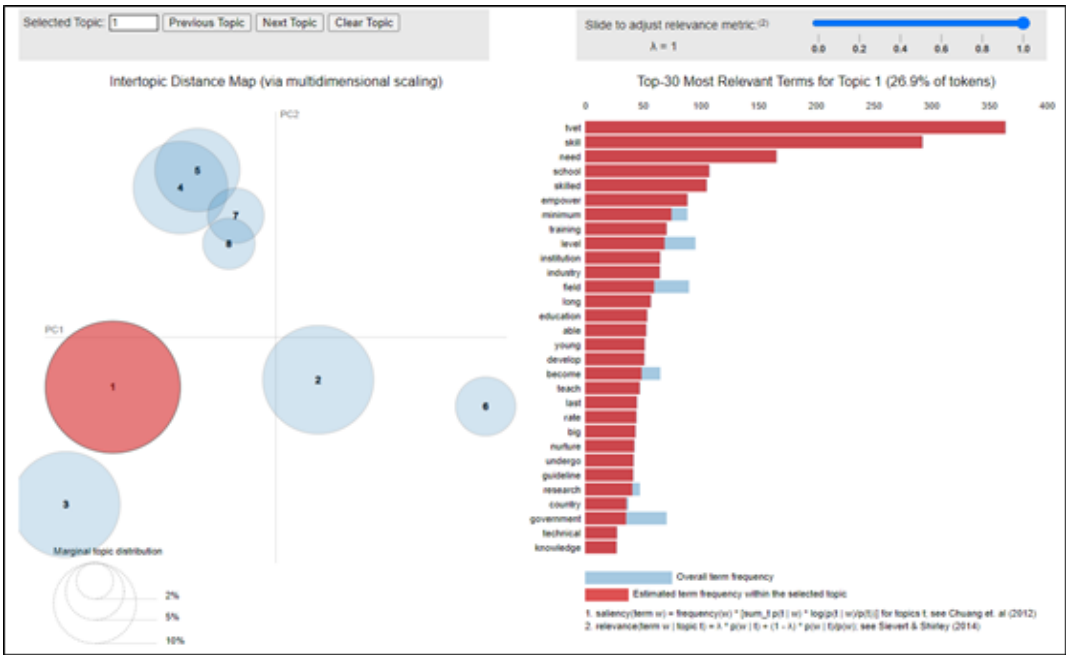

Figure 1.4: LDA Intertopic distance map (Topic 1)

Figure 1.3 illustrates the comprehensive distribution of topics and terms, whereas Figure 1.4 presents the pertinent terms associated with the selected topic. In the context of topic 1, the prevalent terms observed are ‘tvet’, ‘skill’, ‘need’, and ‘school’. The appropriate topic that encompasses those terms is skills or competency. People always refers TVET as skill that must be acquired in order to pursue technical studies and get employment opportunities. The most terms describe in topic 2 are ‘student’, ‘certificate’, and ‘study’. There is a discussion among Facebook users that some certifications in higher secondary level like vocational college are not being recognized to continue study to tertiary level. Only some of those institution apply accreditation from Malaysia Qualification Agency (MQA). The suitable topic is certificate.

Topic 3 encompasses of ‘graduate’, ‘salary’, ‘get’, and ‘pay’ terms which pertain to the topic of compensation for employment. There is considerable debate and discussion that the government plan to propose wage of RM3,000 for TVET graduates. However, both TVET and non-TVET graduates exhibit a preference for higher salaries in their work. In relation to topic 4, the prevalent terms are ‘people’, ‘enter’, ‘want’, and ‘come’. People have been discussing and exchanging information regarding the process of enrolling in TVET institutions.
This includes inquiries about the qualifications necessary for enrolment and deliberations on which institution provides the best services and support. This term can be commonly associated with the subject matter of intake and admission.

Topic 5 represents the terms of ‘good’, ‘employer’, ‘apply’ and ’practice’. Employers are engaging in the practice of expressing their perspectives and opinions on TVET students who are participating in internships and securing employment inside their respective organizations. These terms describe the perspective of employer towards the TVET students and graduates. Other terms that could be considered for topic 6 are ‘science’, ‘agriculture’ and ‘student’. There

is an ongoing discourse about the attraction of studying agriculture to students, even those studying science. In addition, they advised secondary school graduates who are not interested in higher education to consider acquiring technical and vocational skills through the many courses available. The appropriate topic for the terms is related to courses.

Topic 7 comprises the terms of 'class', 'learn', and 'old'. These terms have the potential to encompass the term of stigma. In this particular context, the term 'class' pertains to the social stratum that is categorized as second class or lower class. Individuals have recommended that parents and older individuals refrain from comparing TVET occupations with other professional employment, while also refraining from categorizing them as being of lower social standing. This observation indicates that the societal stigma remains prevalent. The final topic being examined encompasses the term of 'employ', 'various', 'share'' and 'involve'. There is a strong encouragement for industries and their respective agencies to engage in sharing and collaborative efforts with TVET institutions in order to increase the knowledge and capacities within the TVET sector. These terms can be appropriately described as industry involvement.

CONCLUSION AND FUTURE RESEARCH

This study investigated latent topic in TVET share on social media using base line LDA topic modeling. A number of processes were carried out including data collection, data preparation and pre-processing, feature extraction, topic modeling, and visualization. This approach revealed eight (8) topics including skills/competencies, certification, salary/wage, intake/enrolment, employer views, course, stigma and industry involvement. It can be concluded that TVET-related data on Facebook can be analyzed and modelled to provide useful information. Therefore, the extracted topics can serve as a guide for relevant stakeholders to consider public opinion on social media platforms when developing recommendations, strategies and policies. Further studies will explore different approaches including Latent Semantic Analysis (LSA), Non-negative Matrix Factorisation (NMF) and BERTopic to conduct a comparative analysis of the identified topics.

REFERENCES

Blei, D. M., Ng, A. Y., & Edu, J. B. (2003). Latent Dirichlet Allocation Michael I. Jordan. In Journal of Machine Learning Research (Vol. 3)

Dhaoui, C., Webster, C. M., & Tan, L. P. (2017). Social media sentiment analysis: lexicon versus machine learning. Journal of Consumer Marketing, 34(6), 480–488. https://doi.org/10.1108/JCM-03-2017-2141

Durham, J., Chowdhury, S., & Alzarrad, A. (2023). Unveiling Key Themes and Establishing a Hierarchical Taxonomy of Disaster-Related Tweets: A Text Mining Approach for Enhanced Emergency Management Planning. Information (Switzerland), 14(7). https://doi.org/10.3390/info14070385

Economic Planning Unit. (2015). Eleventh Malaysia Plan 2016-2020. https://doi.org/10.1109/CLUSTR.2004.1392655

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018). Text Mining in Organizational Research. Organizational Research Methods, 21(3), 733–765. https://doi.org/10.1177/1094428117722619

Mathayomchan, B., Taecharungroj, V., & Wattanacharoensil, W. (2022). Evolution of COVID-19 tweets about Southeast Asian Countries: topic modelling and sentiment analyses. Place Branding and Public Diplomacy. https://doi.org/10.1057/s41254-022-00271-5

Moe, W. W., & Schweidel, D. A. (2017). Opportunities for Innovation in Social Media Analytics. Journal of Product Innovation Management, 34(5), 697–702. https://doi.org/10.1111/jpim.12405

Nurmawiya, & Harvian, K. A. (2021). Public sentiment towards face-to-face activities during the COVID-19 pandemic in Indonesia. Procedia Computer Science, 197, 529–537. https://doi.org/10.1016/j.procs.2021.12.170

Phang, Y. C., Kassim, A. M., & Mangantig, E. (2021). Concerns of thalassemia patients, carriers, and their caregivers in malaysia: Text mining information shared on social media. Healthcare Informatics Research, 27(3), 200–213. https://doi.org/10.4258/HIR.2021.27.3.200

Razzaq, A., Asim, M., Ali, Z., Qadri, S., Mumtaz, I., Khan, D. M., & Niaz, Q. (2019). Text sentiment analysis using frequency-based vigorous features. China Communications, 16(12), 145–153. https://doi.org/10.23919/JCC.2019.12.011

Saharudin, M. S., Ali Nahran, S., Che Nasir, N. A., Mohamad Nor Azli, M. S., & Wan Muhamad, W. M. (2021). Prospect, Issues, and Challenges in Malaysia TVET-Based Education. In International Conference on Advancing and Redesigning Educationn: Thriving in Times of Global Change

Salinca, A. (2016). Business Reviews Classification Using Sentiment Analysis. Proceedings - 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2015, 247–250. https://doi.org/10.1109/SYNASC.2015.46

Shinde, P., & Govilkar, S. (2015). A Systematic study of Text Mining Techniques. International Journal on Natural Language Computing, 4(4), 54–62. https://doi.org/10.5121/ijnlc.2015.4405

Yin, B., & Yuan, C. H. (2022). Detecting latent topics and trends in blended learning using LDA topic modeling. Education and Information Technologies, 27(9), 12689–12712. https://doi.org/10.1007/s10639-022-11118-0