Topic Mining-based Fintech Program Analysis

ABSTRACT

FinTech, or "financial technology," describes how contemporary technologies are applied to financial services and products in an effort to enhance, innovate, and broaden the provision of financial services. This evolving field is disrupting and reshaping how traditional financial institutions operate and serve their customers. In order to understand the existing pros and cons of the Fintech program, we are using topic mining to analyze the collected curriculum dataset and utilize the findings to further support the design of a similar Fintech program as a guideline. Firstly, we collected the Fintech curriculum dataset from different universities. Then we cleaned and prepared the dataset. Lastly, we used topic modeling to analyze the dataset. The findings of the analysis are presented as the results. Through this research, we can have a deeper understanding of the Fintech program.

Keywords: Fintech, Topic Mining, Keyword Extraction, Word Cloud

INTRODUCTION

Financial technology, also called FinTech, uses information technology to deliver and enhance financial services (Leong & Sung, 2018). It combines digital innovation with established financial services. With technology support, companies can advance, simplify, and even completely alter conventional financial operations including transaction processing, accounting, and investing, turning them into innovative digital products and online applications (Broby, 2021). These innovations include a wide range of services, such as online banking, mobile stock trading, digital contracts, and intelligent payment solutions.

Such FinTech innovation needs a combination of knowledge from different fields, including banking, software engineering, data science, and cybersecurity (Kaur et al., 2021). This cross-disciplinary cooperation is essential to the development of the fintech industry since it fosters innovation and guarantees the dependability of these electronic financial solutions.

However, this communication among numerous stakeholders is also mandated by the collaboration among professionals. Future development of the fintech sector depends on the ability to integrate different points of view and insights from these many sectors.
Therefore, the future of fintech could demand people with in-depth knowledge across several different fields (Sharma et al., 2023). This dynamic environment presents both challenges and opportunities for higher education institutions. They must adapt to nurture and equip students with the interdisciplinary skills needed by the developing fintech sector (Abdeldayem & Al Dulaimi, 2020). To respond to such demands and challenges, many universities are now beginning to provide courses, certificates, and degree programs in fintech. In this research-in-progress study, we plan to present an overview of the fintech education programs at the university level. Additionally, we want to include the courses, credentials, and degree programs that are available, along with the skills that are taught in each.

RELATED WORK

FinTech Development

FinTech is the nexus of finance and technology, covering a wide range of innovative digital financial services and products (Allen et al., 2019). Online banking, payment services, blockchain technology, and cryptocurrencies are just a few of the services and products that are impacted by this technology

breakthrough, which also boosts sales and creates new business prospects (De Filippi & Loveluck, 2016). Fintech apps, for instance, enable banks to convert from traditional services to solutions that are less costly (Vives, 2017). Banks can reach more marginalized communities by offering mobile banking and digital wallets (Dabrowski & Janikowski, 2018). In addition to the advantages and possibilities, FinTech also confronts several dangers and problems, including worries about data privacy, the need to comply with laws and regulations, and security threats (Barefoot, 2020). Another issue affecting the growth of the fintech sector is how consumers adapt to and trust new technology (Meyliana & Fernando, 2019).

FinTech Education

Due to the change in the finance industry, companies need employees with new knowledge and skills in FinTech to support future development. Students have the chance to get ready for this dynamic business through education in fintech (Molnár et al., 2020). A crucial job for FinTech education is developing courses or curricula at universities, particularly in business schools (Shino et al., 2022). These courses should include digital skills like data analytics, programming, and cybersecurity in addition to diverse themes like blockchain, cryptocurrencies, or laws (Doherty & Stephens, 2023). In addition to teaching students about finance and information technology, regulatory and ethical material is crucial because it helps students understand compliance requirements, data privacy concerns, and ethical issues related to FinTech innovation (Prastyanti et al., 2023). Along with the subjects and abilities listed above, the collaboration between academia and business is essential for providing students with access to real-world projects, internship opportunities, and hands-on experiences (Ivascu et al., 2016). Education may therefore enable students to access and use financial services more skillfully or it may even encourage students to start their own fintech businesses (Purwanto, 2020). However, it is challenging for education programs to integrate involvement from the sector due to the quick-paced nature of the fintech industry, the need for an updated curriculum, and the dearth of qualified faculty members (Panos & Wilson, 2020). In order to help students prepare for future challenges, scholars are looking into what courses and other elements should be included in FinTech education programs (Abdeldayem & Al Dulaimi, 2020). In particular, business schools should design their programs to include courses from various disciplines, such as finance, technology, and entrepreneurship, and they should also incorporate real-world cases into their classroom instruction(Nach et al., 2021). Previous studies used structured interviews to review the skill needed from industry perspectives (Doherty & Stephens, 2023), fewer studies use resources from education and industry, implement text-mining analysis, and compare the similarities and differences between education objectives and industry requirements (Jiang & Chen, 2022).

Text Visualization and Analysis

Text visualization is a branch of data visualization focused on representing and making sense of textual data. Given the unstructured nature of text, extracting insights can be challenging (Cao & Cui, 2016). Visualization techniques can aid in summarizing, understanding, and interacting with large volumes of text. The ultimate goal is to comprehend the pattern in the dataset. Text analysis is the process of extracting meaningful information from textual data (Talib et al., 2016). It involves transforming unstructured text data into structured or semi-structured formats to facilitate various forms of analysis (Raminhos & Moura-Pires, 2007). The purpose of text analysis is mainly used for information retrieval.

This study aims to present an overview of the university-level fintech education programs through text visualization and analysis. The tools we used include word cloud, which is used for visualization, topic modeling that is for categorize the dataset into different categories. Word cloud is a graphic representation of text data in which each word's size and color reflect how important or frequently it appears (Heimerl et al., 2014). In a word cloud, words that appear more frequently in the source text

are displayed more prominently (usually size indicates the frequency of the words), making it easy to discern which words are mentioned most often. The layout is typically random, with words positioned closely together, forming a cluster or cloud-like pattern.

METHODOLOGY

The framework of the proposed method is illustrated in Fig. 1, which comprises three stages: dataset preprocessing, dataset cleaning, topic modeling, visualization, and analysis of the results. The solid line represents our current work, while the dash line will be our future work. In this work, we focus only on the left side of the framework. Our goal is to provide an overview of the fintech master program education landscape. In order to get a detailed view of collected dataset, this work highlights the key topics covered in courses in fintech through different text mining techniques.

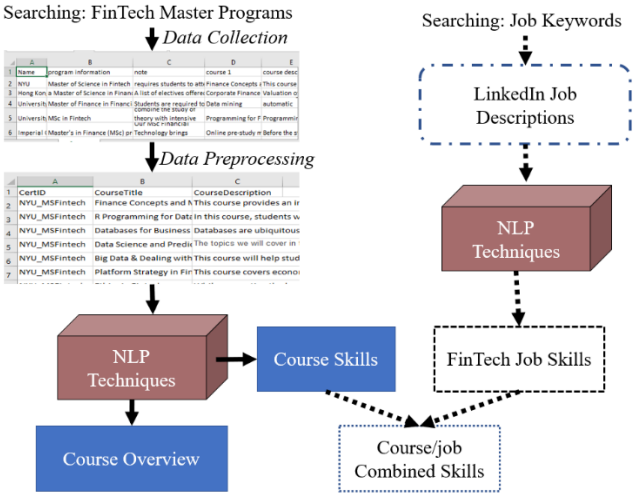

Fig. 1. The framework of the proposed method.

Data Collection

During this preliminary phase of data collection and analysis, we conduct a Google search using the terms "fintech" and "university." We then utilize the pre-existing university list featuring fintech majors or courses from a previous study (Al Hudithi & Siddiqui, 2021) as supplementary sources of information to generate our own list for analysis. In order to conduct the initial analysis, we compiled a sample list of US, UK, and Hong Kong universities offering fintech master's programs in this work-in-progress study.

We examined the eight universities that offered 131 fintech courses. The variables of the list include university name, program information, notes, course 1, course1 description, course 2, and so on. We organized the dataset into a structured format. There are three attributes in each record. Certificate ID is the key id for each record, Course Title is the title for the course, and Course Description is like an abstract of the course. The detailed structure of the dataset is in Fig. 2.

	CertID	CourseTitle	CourseDescription
0	NYU_MSFintech	Finance Concepts and Math (In-Person)	This course provides an introduction to the ke...
1	NYU_MSFintech	R Programming for Data	In this course, students will learn how to pro...
2	NYU_MSFintech	Databases for Business Analytics	Databases are ubiquitous in all businesses and...
3	NYU_MSFintech	Data Science and Predictive Analytics	The topics we will cover in this course begin ...
4	NYU_MSFintech	Big Data & Dealing with Data in Finance	This course will help students learn how to pr...

Fig. 2. Samples of the collected dataset.

Before the preprocessing, we generate the word cloud to visualize both the course title and course description in Fig. 3. From the word cloud we can see that the data contains a lot of interference text. So, the next step is to utilize the data preprocessing to handle and clean the dataset.

(a) (b)

Fig. 3. Word cloud of both course title (a) and course description (b)

Dataset Preprocessing

The next step is to use NLTK (Natural Language Toolkit) to do the text cleaning. NLTK is used for text cleaning. NLTK is a useful platform for creating Python applications that interact with data in human languages. It provides tools to work with text data and perform tasks in natural language processing (NLP) and computational linguistics. It contains professional text manipulation like Tokenization, Stemming, Lemmatization, and so on. It works on the assumption that words that appear frequently and in association with other frequent words are most likely representative of the content. First, we eliminated punctuation, lowercased the text, removed text enclosed in square brackets, and eliminated words that contained numbers using regular expressions. Then packages from NLTK are adopted to lemmatize and tokenize the text. During this process, we also removed the stops words. Due to the benefits from keywords extraction from previous step, we customized the stops words to add more specific stops words for the dataset. After that, the vectorization is used to converting a collection of text into a matrix of token (typically word) counts. With this structured format, we can then work on it with the topic modeling model.

Topic Modeling for Course Overview

The topic modeling model we used is Latent Dirichlet Allocation (LDA) (Blei et al., 2003). LDA is a widely used unsupervised machine learning algorithm that is commonly used for topic modeling. It is a generative probabilistic model that assumes that there is an underlying, unobserved probability distribution that generates the observed data. The goal of LDA is to identify the latent variables that govern this distribution. Specifically, LDA aims to identify the distribution of topics in a corpus and the distribution of words within each topic. These distributions are represented as probability distributions over a fixed vocabulary.

The LDA algorithm works as follows. First, it randomly assigns each word in the corpus to one of K topics. Next, for each document in the corpus, LDA generates a topic distribution by randomly sampling a mixture of K topics. Finally, for each word in each document, LDA generates a word distribution by randomly selecting a word from the assigned topic. During training, the algorithm updates the assignment of topics to words based on the likelihood of the observed data. This update involves iteratively estimating the posterior distribution of the latent variables given the observed data, and then re-estimating the parameters of the model using this distribution. This process continues until convergence is achieved. LDA analyzes each document d to determine the probability p of each topic t given the words w in the document, denoted as $p(\text{topic } t \mid \text{document } d)$. The algorithm also calculates the probability of each word w being associated with topic t, represented as $p(\text{word } w \mid \text{topic } t)$, by

examining the frequency of topic assignments to that word across all documents. These probabilities are then utilized to update the probability of the word *w* being associated with topic *t*, indicated by *p*(word *w* with topic *t*) = *p*(topic *t* | document *d*) * *p*(word *w* | topic *t*). This process is iterated until the algorithm converges to a steady state where topic assignments no longer change. Finally, based on the final topic assignments, the proportion of topics in each document can be determined.

From the previous step, the dataset is organized in sparse matrix format which represents documents as sparse matrices where each row might represent a document and each column represents a term (word or n-gram) from the entire corpus vocabulary (Most of these values are zeros (hence, sparse matrix)). In order to apply LDA, the dataset is converted from sparse matrix to corpus which represents the documents as a list. Each document is a list of tuples. Each tuple contains an index (representing the term) and a value (like a count or a TF-IDF weight). We then generate topics via LDA based on the keywords.

Keyword Extraction for Course Skills

To get the skill keywords among the collected dataset, we used three classic keyword extraction approaches which are plus TF-IDF (term frequency-inverse document frequency), RAKE (Rapid Automatic Keyword Extraction), and SpaCy to extract the keywords. TF-IDF is a numerical statistic that aims to represent a word's significance to a document within a corpus or collection. It's one of the most popular techniques for converting text data into a meaningful vector format, and it's widely used in information retrieval and text mining. It first calculates two components: TF(*TF(t) = Number of times term t appears in a document / Total number of terms in the document*) and IDF (*IDF(t) = log(Total number of the document / Number of the document with term t in it*)). Then the TF-IDF is computed based on the expression TF-IDF(t)=TF(t)×IDF(t). A high TF-IDF value indicates that the term is important for that specific document, relative to the entire corpus.

RAKE is an unsupervised, domain-independent, and language-independent method for extracting keywords from text. It's designed to determine key phrases in a document by analyzing the frequency of word appearance and its co-occur patterns. SpaCy is a Python natural language processing (NLP) library. It's designed for production use and widely used as NLP libraries in both academia and industry.

EXPERIMENTAL RESULTS

This section demonstrates the NLP based data analysis results on the collected dataset to effectively provide an overview of fintech master program curriculum. The experiments were performed on a laptop equipped with an Intel(R) i7-Core CPU, using the Python programming language and relevant packages, including nltk, sklearn, genism, and wordcloud.

LDA based Fintech Master Curriculum Overview

The aim of this section is to provide an overview of the current Fintech master curriculum. We use LDA based topic modeling technique to extract topics from course description. Similar to previous work (Zhou et al., 2023), we set the number of topics to be five. The first 15 words of the discovered topics are listed in Table 1.

Table 1. Topics Generated by LDA with topic number = 5

Topic ID	First 15 words of each topic

Topic 0	'data', 'financial', 'analysis', 'topic', 'application', 'used', 'technology', 'mining', 'skill', 'focus', 'cover', 'finance', 'technique', 'time', 'capital'
Topic 1	'data', 'model', 'fintech', 'analytics', 'learning', 'business', 'financial', 'system', 'technique', 'ml', 'technology', 'analysis', 'machine', 'payment', 'assessment'
Topic 2	'finance', 'market', 'data', 'fintech', 'strategy', 'financial', 'investment', 'programming', 'trading', 'equity', 'theory', 'asset', 'behavioral', 'bond', 'learning'
Topic 3	'financial', 'data', 'risk', 'technology', 'management', 'fintech', 'investment', 'portfolio', 'service', 'industry', 'asset', 'application', 'strategy', 'concept', 'blockchain'
Topic 4	'financial', 'learning', 'finance', 'market', 'application', 'project', 'topic', 'technology', 'part', 'data', 'basic', 'concept', 'technique', 'fintech', 'assessment'

As can be observed from Table 1, almost all of current Fintech master program focuses on “data”, “model”, “finance”, “analysis”, “market”, “risk” related information. Specifically, Topic 0 focuses on financial analysis based on data and application; Topic 1 emphasizes on the model and analytics of fintech, it mentioned ML and assessment; Topic 2 has the market, investment and behavior related perspective of the curriculum contents; Topic 3 is relevant to risk, management, investment, and asset perspectives of fintech; Topic 4 focuses on market, project, and application aspects of fintech.

For all the collected courses, we assigned each course in the dataset to its corresponding topic ID. For instance, for the program “HKUST_MSFintech”, the first course is “Finance Concepts and Math (In-Person)”. The course descriptions are shown below. This course can be classified to the Topic_4. The resulting distribution of topics within the dataset is presented in Fig. 4.

Finance Concepts and Math (In-Person): *"This course provides an introduction to the key concepts and the associated analytical tools that are the basic building blocks for all financial analysis. These concepts and tools are essential in order to understand the material in later courses, but they are also interesting and important in their own right. Fintech tests the laws of finance, ……"*

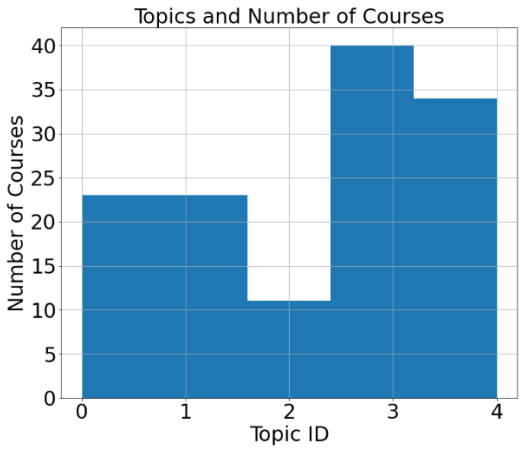

Fig. 4. Histogram of course description and topic_ID distribution.

As illustrated in Fig. 4, Topic_3 has about 40 courses, indicating that most courses cover information about risk, management, investment, and asset perspectives of fintech. Topic_2 has around 11 courses, which means current fintech master program curriculum have less emphasis on market, investment and behavior related contents.

Table 2. Topic Distribution in terms of Fintech Master Program Institution Name

Institution Program Name	Topic_0	Topic_1	Topic_2	Topic_3	Topic_4
NYU_MSFintech	2	6	4	7	4
HKUST_ MSFintech	2	1	0	4	3
UHK_ MSFintech	2	9	0	10	10
Ustrathclyde_ MSFintech	5	2	2	7	3
Imperial_College_ MSFintech	5	0	3	4	7
NewJerseyCityU_ MSFintech	3	2	1	1	1
UoEssex_ MSFintech	2	1	0	0	3
Coventry_University_ MSFintech	2	2	1	7	4

By utilizing the topics identified through topic modeling methods, we can annotate each program record in the collected dataset with relevant topics. Table 2 shows the topic ID for each program. Table 2 and Table 3, we can clearly see the difference between the focus of each program.

Keyword Extraction on Course Skills

To get the set of skills of fintech, we employ keyword extraction to extract keywords from the collected dataset. To ensure we get a well analyzed set of course skills, we employed three different keyword extraction techniques, i.e., TF-IDF based NLTK, Rake, and SpaCy. Fig. 5 shows the wordcloud result of the keyword extraction.

Then we combine all three keywords lists together. After removing duplicated keywords, we get a list of course skills, as shown in Table 3.

Table 3. Fintech Master Program Course Skills

data, python, analysis, data mining, neural network, predictive, cyber risk management, online privacy, data protection, machine learning, predictive tools, predictions analysis, visualizing, data graphics, visual representation methods, tableau, statistics, data mining, data protection, personal data privacy ordinance, R, SQL query, SQL, database systems

(a)TF-IDF (b) rake (c) SpaCy

Fig. 5. WordCloud visualizing keywords extraction results.

CONCLUSION

Fintech is an area that received attention due to its uniqueness of combining both financial and technology into education. In this paper, we analyzed the Curriculum dataset collected from Fintech programs. Topic modeling is used to categorize the courses in the curriculum into different classes. We used LDA and generated 5 topics from the dataset. Further analysis shows the topic modeling can correctly divide the dataset into distinct categories. From the topic distribution, we can certainly see

the insights of the curriculum design of Fintech program. For future work, we will collect more datasets and apply text analysis and text visualization techniques on both fintech programs and job market.

REFERENCE

Abdeldayem, M. M., & Al Dulaimi, S. H. (2020). Trends ofglobal fintech education practices and the gcc perspective. *Int. J. Adv. Sci. Technol*, *29*, 7150–7163.

Al Hudithi, F., & Siddiqui, K. A. (2021). Designing the guidelines for FinTech curriculum. *Entrepreneurship and Sustainability Issues*, *9*(1), 633.

Allen, F., Brealey, R., & Myers, S. (2019). *Principles of Corporate Finance, 13th edition*. McGraw-Hill.

Barefoot, J. A. (2020). Digital technology risks for finance: Dangers embedded in Fintech and Regtech. *M-RCBG Associate Working Paper Series*, *151*.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Broby, D. (2021). Financial technology and the future of banking. *Financial Innovation*, *7*(1), 1–19.

Cao, N., & Cui, W. (2016). *Introduction to text visualization* (Vol. 1). Springer.

Dabrowski, M., & Janikowski, L. (2018). Virtual currencies and their potential impact on financial markets and monetary policy. *CASE Research Paper*, *495*.

De Filippi, P., & Loveluck, B. (2016). The invisible politics of bitcoin: Governance crisis of a decentralized infrastructure. *Internet Policy Review*, *5*(4).

Doherty, O., & Stephens, S. (2023). Hard and soft skill needs: Higher education and the Fintech sector. *Journal of Education and Work*, *36*(3), 186–201.

Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014). *Word cloud explorer: Text analytics based on word clouds*. 1833–1842.

Ivascu, L., Cirjaliu, B., & Draghici, A. (2016). Business model for the university-industry collaboration in open innovation. *Procedia Economics and Finance*, *39*, 674–678.

Jiang, H., & Chen, C. (2022). Data science skills and graduate certificates: A quantitative text analysis. *Journal of Computer Information Systems*, *62*(3), 463–479.

Kaur, G., Lashkari, Z. H., & Lashkari, A. H. (2021). *Understanding Cybersecurity Management in FinTech*. Springer.

Leong, K., & Sung, A. (2018). FinTech (Financial Technology): What is it and how to use technologies to create business value in fintech way? *International Journal of Innovation, Management and Technology*, *9*(2), 74–78.

Meyliana, M., & Fernando, E. (2019). The influence of perceived risk and trust in adoption of fintech services in Indonesia. *CommIT (Communication and Information Technology) Journal*, *13*(1), 31–37.

Molnár, B., Tarcsi, Á., Baude, F., Pisoni, G., Ngo, C. N., & Massacci, F. (2020). Curriculum guidelines for new Fintech Master's Programmes. *2020 18th International Conference on Emerging eLearning Technologies and Applications (ICETA)*, 470–474.

Nach, H., Salampasis, D., Vulkan, N., Zachariadis, M., & Lenz, J. (2021). *Financial Technologies on Campus: The Rise of FinTech in Higher Education*.

Panos, G. A., & Wilson, J. O. (2020). Financial literacy and responsible finance in the FinTech era: Capabilities and challenges. In *The European Journal of Finance* (Vol. 26, Issues 4–5, pp. 297–301). Taylor & Francis.

Prastyanti, R. A., Rezi, R., & Rahayu, I. (2023). Ethical Fintech is a New Way of Banking. *Kontigensi: Jurnal Ilmiah Manajemen*, *11*(1), 255–260.

Purwanto, A. (2020). Revealing the Intervening Effect of Fintech Lending between Entrepreneurial Education and Entrepreneurial Intentions. *International Journal of Advanced Science and Technology*.

Raminhos, R. F., & Moura-Pires, J. (2007). *Extraction and transformation of data from semi-structured text files using a declarative approach*. Ninth International Conference on Enterprise Information Systems, Madeira, Portugal.

Sharma, S. K., Ilavarasan, P. V., & Karanasios, S. (2023). Small businesses and FinTech: A systematic review and future directions. *Electronic Commerce Research*, 1–41.

Shino, Y., Lukita, C., Rii, K. B., & Nabila, E. A. (2022). The Emergence of Fintech in Higher Education Curriculum. *Startupreneur Business Digital (SABDA Journal)*, *1*(1), 11–22.

Talib, R., Hanif, M. K., Ayesha, S., & Fatima, F. (2016). Text mining: Techniques, applications and issues. *International Journal of Advanced Computer Science and Applications*, *7*(11).

Vives, X. (2017). The impact of FinTech on banking. *European Economy*, *2*, 97–105.

Zhou, J., Tao, Y., Duan, Z., & Gautam, S. (2023). Topic Mining Based Literature Analysis for Face Recognition. *Data Science and Algorithms in Systems: Proceedings of 6th Computational Methods in Systems and Software 2022, Vol. 2*, 807–818.