Trends in Technical & Scientific Research
ISSN: 2641-8355

Juniper Publishers
key to the Researchers

Research Article
Volume 6 Issue 5 - October 2023
DOI: 10.19080/TTSR.2023.06.555697

Trends Tech Sci Res

Leveraging Social Media Data and NLP to Identify Key Leaders in AI and Digital Transformation in the Maritime Industry

Yuwei Hua and Carol Anne Hargreavesa*

National University of Singapore, Department of Statistics and Data Science, Singapore

Submission: September 11, 2023; **Published:** October 10, 2023

***Corresponding author:** Carol Anne Hargreavesa, National University of Singapore, Department of Statistics and Data Science, Singapore

Abstract

Social media has been increasingly popular amongst the last decades. Among many of these users on social media, there are people who have garnered attention and they are classified as influencers. LinkedIn was launched in 2003 and has been one of the top platforms for professionals in the industry to connect. It has also become a platform where it has allowed professionals to share topics relating to the industry. In this study, we identify influencers in the maritime industry in Singapore and Hong Kong and identify the discussion topics. Python packages such as Beautiful Soup and Selenium were used to scrape LinkedIn profiles of people in the maritime industry in Singapore and Hong Kong. For topic modelling, a Natural Language Processing technique, (NLP), Latent Dirichlet Allocation (LDA) was applied to identify topics posted on Linkedin was applied to derive the insights on the topics the leaders in the Maritime Industry were talking about. Recency, Frequency and Followers (RFF) modelling was performed to identify LinkedIn professionals that scored high in (1) how recently they have posted, (2) how frequently they have posted and (3) how many followers they have. Discussion topics in the maritime industry were identified and key leaders were identified by their top RFF scores. Fifteen discussion topics were identified. In conclusion, our study successfully identified key maritime leaders and the topics they posted. The identification of key maritime leaders is valuable as companies can invite these key leaders to attend events that require leaders to speak on topics that people will listen to and act on.

Keywords: Social Media; Natural Language Processing; NLP; AI; RFF Modelling; Latent Dirichlet Allocation; Influencers

Abbreviations: OSNs: Online Social Networks; AI: Artificial Intelligence; GDP: Gross Domestic Product; MPA: Maritime and Port Authority; LDA: Latent Dirichlet Allocation; NLP: Natural Language Processing

Introduction

Online social networks (OSNs) are very popular amongst people in this digital age, and there are platforms for people to connect with each other and to disseminate information. These platforms include Twitter, Facebook, YouTube, Instagram, LinkedIn and many more. Many of these platforms have professionals posting various topics. One special feature about LinkedIn is its density of professionals since LinkedIn was used as a job search and connection platform. Many people keep their profiles professional, and this is a good platform for companies who are actively looking for professionals in a certain industry. On LinkedIn, one can also post their activity, similar to that on other platforms such as Twitter or Instagram. However, the major difference between other platforms and LinkedIn is that people would keep their posts professional as it is a platform for networking between professionals (Utz and Breuer). This study aims to identify influencers for the purpose of Human Resource. With this study, questions such as who is the most suitable influencer that can be identified if there was an event or conference on key maritime discussion topics like artificial intelligence (AI) or digital transformation? As the purpose of this study is to identify influencers in the professional industry, LinkedIn is a suitable platform to retrieve data regarding professionals in the maritime industry and to identify AI and digital transformation influencers. This study aims to answer the following questions:

- Who are the AI/digital transformation influencers in the Maritime Industry?
- What are the AI/digital transformation influencers posting? What topics are they talking about?

The maritime sector is a crucial part of Singapore's industry and is essential for the nation's growth and success as the maritime sector accounts for an approximate 7% of the country's Gross Domestic Product (GDP). Singapore is a leading international maritime center and a global maritime hub that connects 600 ports

in over 120 countries. With the maritime industry playing such a big role in Singapore's economy, it is important that Singapore remains competitive and up to date. Singapore has to innovate to meet future needs in terms of digitalisation. With the Covid-19 pandemic, there was a need to accelerate the efforts of digitilisation in shipping because of the myriad of benefits that digitilisation brings. A webinar that took place on 8 October 2020, called the "Future of Shipping - Digitilisation" was jointly organised by the IMO and the Maritime and Port Authority of Singapore (MPA). The digitilisation efforts were communicated and emphasized at these events and webinars. People who are influential in AI in the maritime industry are the perfect choice for keynote speakers as they will influence others in the maritime industry to attend the event and to learn more about the digitilisation process and relevancy in the maritime industry.

The maritime industry in Hong Kong is also one of the busiest ports in the world, with it accounting for 1.1% of Hong Kong's GDP. With these 2 locations having a huge emphasis on Maritime, LinkedIn accounts from these 2 locations were investigated.

Related Works

Luca et al. [1] built a "Micro-influencer classifier for the classification of micro influencers on two platforms, Twitter, and Instagram. According to Luca et al., the reason for studying influencers is that there could be two frameworks that could be discussed, firstly, the academic framework and secondly the economic framework. The academic framework makes up the first part of our study and focuses on data scraping that is important for academic research. There are Python libraries available to scrape data from the social networks, such as Twitter and Instagram. Data such as the follower count, the number of posts for each account, the number of likes, the number of reposts, the number of post shares, and the actual text of the post. The Python scraping tools have allowed researchers to see influencers beyond just their follower count but also to derive insights from the posts and to analyse the posts in great detail.

The other framework is the economic approach, where the work finds out what are the factors that affect the classification of the micro influencer. The model can then be used to classify new users if they are a micro influencer. LinkedIn is a platform for professional self-promotion and is used universally as a tool for career development. [2] LinkedIn provides the platform for professional communication and as of 2015 it has more than 400 million members. As of 2022, it has more than 800 million users [3]. It is now the most popular website for professional social networking.

This project was inspired by an article to identify potential potential YouTube influencers using python [4]. The python web scraping was used to identify potential influencers for collaboration on the online platform "YouTube". Youtubers that create content related to cryptocurrencies were scraped using the python package Selenium. The inspiration of this project was that in terms of the influencers, many companies were looking to influencers for brand marketing and other social media. If the maritime industry is interested in leaders, influencers are a great start. Identifying influencers are a hot topic and to date, there have been many projects on the topic of identifying influencers. Mehmet et al have used machine learning classification algorithms to find influencers on Twitter. Twitter consists of APIs (Application Programming Interface) to collect data. These machine learning algorithms used in the paper include Random Forest, BayesNet, K-nearest neighbours etc.

Latent Dirichlet Allocation was developed by David Blei et al in 2003 and became a standard tool for topic modelling. [5] Paper on "Topic modelling for Social Media content: A practical approach" have revealed that Latent Dirichlet Analysis are effective in detecting the topic facets and extracting their dynamics over time. This approach was experimented on 90,527 records in the project to detect topics in social media in the area of Aviation and Airport management. The results obtained by LDA was compared with the expert's own interpretation of the topic and it shows that LDA algorithm is effective in identifying topics in the domain.

Methodology

Introduction of the Study Analysis

The methods were divided into three phases. Figure 1 shows the flow of the project. First the data collection. Data collection methodology involves scraping the LinkedIn links of the accounts, afterwards by visiting each link, we would arrive at the profile and the profile consists of the number of followers, number of connections, the country of where the profile user is, the activity and their education and experience. The scraping of these data was done by Python using Beautiful Soup and Selenium. Followers was used instead of connections as LinkedIn would show that the profile has a maximum of 500 connections, profiles with more than 500 connections were all shown as "500+ connections". Using "Followers" would give a more exact number of audiences the account had. The second phase is Natural Language Processing, topic modelling using Latent Dirichlet Allocation (LDA) was used to identify the topics each account is posting. The third phase is using RFM modelling to identify profiles that score high in recency of posting, frequency of posting and number of followers.

Data Source and Cleaning

This project requires data on the profiles of maritime professionals with LinkedIn accounts. Due to the limited accessibility of maritime AI professionals' data, the data is scraped from the web by using Singapore and Hongkong's LinkedIn's API and by making use of Python's libraries: Selenium and Beautiful Soup. Chromedriver.exe was also a file used for the scraping process and was downloaded too. The Selenium package was imported by the Python script and it allowed us to interact with the pre-decided web browser, Chrome driver. Chrome driver was

How to cite this article: Yuwei H, Carol Anne H. Leveraging Social Media Data and NLP to Identify Key Leaders in AI and Digital Transformation in the Maritime Industry. Trends Tech Sci Res. 2023; 6(5): 555697. DOI:10.19080/TTSR.2023.06.555697

used to launch Google Chrome [6]. Chromedriver enabled us to navigate the LinkedIn webpage. As the LinkedIn webpage is in HTML, the python library Beautiful Soup was used to extract the content of the website in HTML ("Web Scraping: Crawling LinkedIn Profiles"). Both quantitative and qualitative data was scraped from the LinkedIn profiles. Quantitative data is data that is expressed in numbers or digits; an example of quantitative data is, the follower's count. Qualitative data is data that is expressed in words; an example would be the captions of posts. Figure 2 below describes the process and packages that had been used for the web scraping.

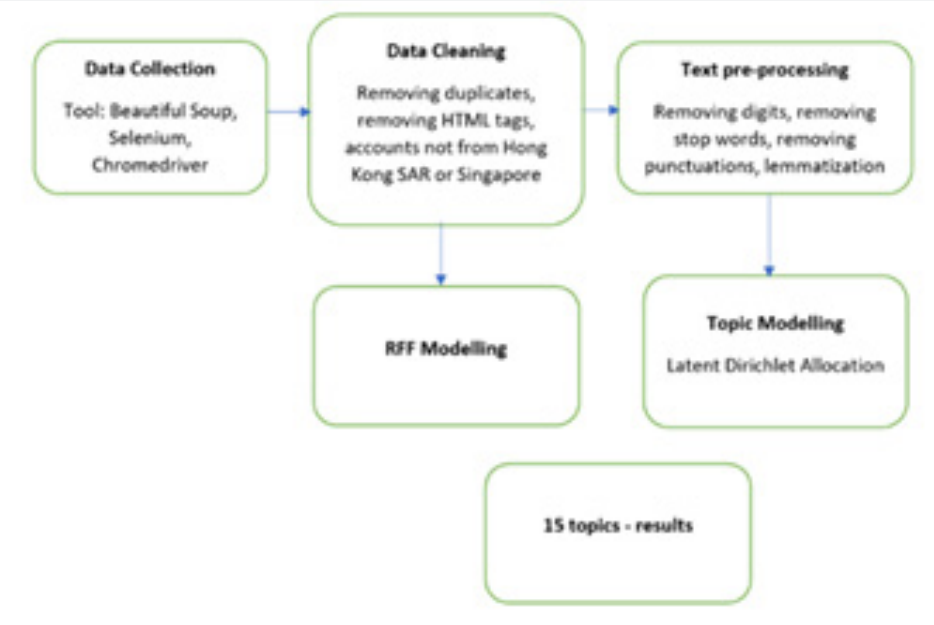

Figure 1: Overview of project.

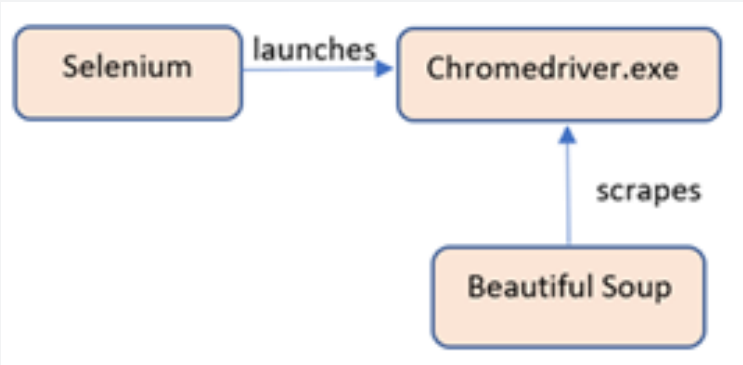

Figure 2: Web Scraping using Selenium, Chromedriver and Beautiful Soup.

Beautiful Soup

First, Selenium was used to launch the LinkedIn website on the web browser installed (Chromedriver), key in the log in credentials and login to LinkedIn, then key in the keywords of the relevant profiles to look for on LinkedIn in the Search bar. By limiting the search to only 'people' and filtering the search location to only 'Singapore' or 'Hong Kong', a list of profiles that were relevant to the maritime industry is obtained. As this project also looks into

How to cite this article: Yuwei H, Carol Anne H. Leveraging Social Media Data and NLP to Identify Key Leaders in AI and Digital Transformation in the Maritime Industry. Trends Tech Sci Res. 2023; 6(5): 555697. DOI:10.19080/TTSR.2023.06.555697

profiles from Hong Kong, the location filter is also used to filter maritime profiles from 'Hong Kong SAR'. Some keywords that are related to looking for profiles on LinkedIn relating to the maritime industry in Singapore and Hongkong are: 'maritime', 'maritime AI'. From this list of profiles shown in the web browser, it is crucial to obtain their profile links. To obtain the profile links, the package beautiful soup is used. We would obtain the result from a list of profile links. By inspecting elements from LinkedIn's HTML, we would obtain the class names of where the information we want to scrape were located.

By navigating with the 'element' tab in the web browser, we could see the information we want is in which class. By using this unique class name that contains the data needed, the class name was keyed into the Python script and Beautiful Soup was used to obtain the information from the elements of LinkedIn's HTML. This was done for all the profile links obtained and then the information was saved to a comma separated value (.CSV) file. The information collected from each profile were Name, Profile URL, Number of followers, Number of connections, Activity, Education, Experience, About, Title, Location.

- Name: Profile name of the account
- Location: Singapore or Hong Kong SAR
- Title: Headline section under the name in the LinkedIn Profile
- URL: LinkedIn page link
- Connections: number of connections
- Followers: number of followers
- About: Introduction of the profile
- Activity: Recent posts, comments
- Experience: Past job experiences
- Education: Past education records and the schools the user has attended.

A total of 800 profiles were scraped from LinkedIn, of which 56% of the profiles were from Singapore and 44% of the profiles were from Hongkong for the purpose of a balanced dataset. Once the data was obtained, data cleaning was done on the columns. Firstly, the duplicated names were checked to see if an account has been scraped more than once. The duplicated accounts were removed, leaving only one occurrence from each account.

Next, the location column was checked to see exactly where the profiles were from. Of which, there were profiles whose location was not identified as Singapore or Hong Kong have been scrapped. The profiles from these locations not identified as Singapore or Hong Kong were removed, leaving only profiles identifying them as Hong Kong or Singapore. Due to the scraping of the HTML or API of LinkedIn, the "followers" column was scraped, and it shows "526 followers 526 followers". While using python for data cleaning, only the number was needed to show 526 in integers.

The "Activity" column shows the recent 3 months or equivalent to 90 days of activities of the user, this includes the posts the user had posted, the reshares of another post and expressing their viewpoints, their comments on another post or their likes on another post. The "Activity" column shows their posts and interactions. From searching for keywords in the "Activity" column, one can see how many posts, reshares and comments a user has made. These posts and interactions are made into a new column. When a user posts, the user can use hashtags to identify certain topics that they would like to relate to, on a LinkedIn post, this could appear as "#maritime", by counting the number times the symbol "#" appeared in the post, one can obtain the number of hashtags the user has used.

After sieving out information from the "Activity" column, there were new columns created and appended to the table to form a new table. The new columns include post_count, hashtag_count etc. New columns are created from data cleaning and these columns include:

- hashtag_count: number of hashtags used.
- post_count: number of posts posted in the 3 months.
- Followers_count: numerical digit of the number of followers
- Activity_clean: cleaned activity column that does not contain html words.

As the profiles scraped may include undergraduates, which for this study were not the group to look into as they are not leaders in the field. One key point of this project was to look into the leaders of the maritime industry, hence these undergraduates will be filtered out when building the model for natural language processing for topic modelling. These undergraduates have indicated under the "Title" column with keywords such as "Undergraduate", "Final Year" or "Penultimate. Hence accounts with these keywords under the "Title" column were filtered, leaving accounts that are non-undergraduates. The median of the followers of all the accounts is 598 followers. With this, accounts that have followers smaller than or equal to 598 are considered as non-influencers in the project. Accounts that consists of followers count that are greater than 598 followers are considered as influencers in the project.

NLP Process - Latent Dirichlet Allocation (LDA)

Natural language processing (NLP) is the process of analyzing human language with the help of computer science and artificial intelligence. There are various forms of Natural Language processing and topic modelling is one of a NLP technique. Topic modelling was used to look into insights about the topics the leaders in the Maritime Industry were talking about (Figure 3). Topic modelling is recognizing the words from the topics present

How to cite this article: Yuwei H, Carol Anne H. Leveraging Social Media Data and NLP to Identify Key Leaders in AI and Digital Transformation in the Maritime Industry. Trends Tech Sci Res. 2023; 6(5): 555697. DOI:10.19080/TTSR.2023.06.555697

in the document or the corpus of the data, instead of extracting a word in a document [6]. In this context, we will be using Latent Dirichlet Allocation (LDA), one of the methods of topic modelling in NLP to identify topics in a document. The data where LDA is performed will be on the 'Activity' column of non-undergraduates. The 'Activity' column was where the users post their posts and shows their recent activity. Text pre-processing was done to make text processing easier in natural language processing as it helps to clean the data, gets rid of noise and unhelpful parts in the data [7]. The steps for text pre-processing for LDA includes lowering characters, removing stop words, removing digits and punctuations, irrelevant character removal, tokenization and lemmatization [8].

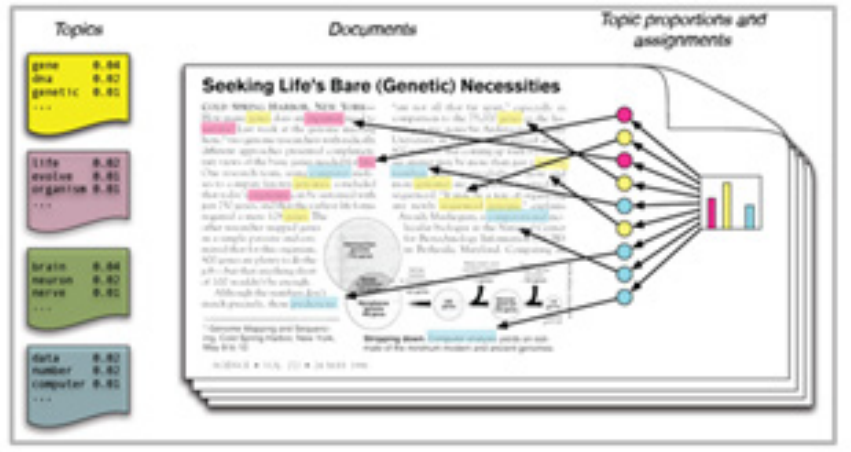

Figure 3: Topic modelling.

Figure 4: Word cloud.

After the text pre-processing, a word cloud, also known as tag cloud, was formed. A word cloud represents the text data in a visual manner. The words visualized in the word cloud are usually single words. The words will be depicted in various sizes. The size of the words portrays how frequently the word appears in the data [9]. Figure 4 below shows the image of the word cloud and the results of the word cloud. From the final word cloud, we could see that the largest words are 'ship', 'vessel', 'safety' and 'event'. The words 'ship' and 'vessel' appear big due to these keywords being closely related to the maritime industry, and hence it was frequently used. Other words that are related to recent hot topics in the maritime industry are: 'decarbonisation', 'security', 'energy'.

How to cite this article: Yuwei H, Carol Anne H. Leveraging Social Media Data and NLP to Identify Key Leaders in AI and Digital Transformation in the Maritime Industry. Trends Tech Sci Res. 2023; 6(5): 555697. DOI:10.19080/TTSR.2023.06.555697

The keyword 'decarbonisation' was one of the words that appears in the word cloud. It is also one of the hot topics in the maritime industry where leaders in the maritime industry encourages to build a sustainable Maritime Singapore. Some keywords that were relevant to the maritime AI industry and appeared in the word cloud includes: 'data', 'technology'.

Next, Latent Dirichlet Allocation (LDA) was performed on the cleaned dataset, the 'Activity' column, using the python genism package. The genism package uses unsupervised machine learning algorithms to process texts [5]. The genism is a Natural Language Processing package that does topic modelling. The results of the NLP LDA were visualized using the python LDAvis package. The LDAvis package helps to create the visualization of the results of the topic modelling. The visualization is an interactive platform shown in Figure 5.

Figure 5: LDA visualization.

	C	D	L	M	N	O	P	Q	R	S	T
1	Location	Title	Followers	hashtag_c	post_cour	Activity_clean	influencer F		Fol	Recency	Topics
2	Singapore	Lead Shipping Analyst | S&	13161	35	4	posted this 1mo1mo Thanks Foreign Policy for sharing my view. , the lead ship	1	3	3	3	5
3	Singapore	Chief Executive Officer at t	7895	20	4	posted this 1w1w This is what a typical afternoon of project scoping and hash	1	3	3	3	11
4	Singapore	Principal Consultant - Conv	9176	11	3	posted this 1mo1mo Partnering with a global dry bulk vessel owner for this po	1	3	3	3	6
5	Singapore	Head of Communications ,	2344	23	4	A"hope for the best" approach will not help to prevent cyber security attacks. M	1	3	3	3	7
6	Singapore	Empowering maritime con	3751	16	4	posted this 13h13h Following on from the examples of Maritime Data Benefits	1	3	3	3	3
7	Kowloon,	Chartering & Operations p	9726	14	3	posted this 1d1d There is absolutely no reason to be around anyone who mak	1	3	3	3	9
8	Hong Kon	Group Operations Head at	1459	0	3	posted this 2d2d "What differentiates bold leaders and leadership teams isnt	1	3	3	3	3
9	Hong Kon	MICS, MIMarEST, MRINA ,	17292	6	3	reshared a post 1mo1mo Wish you all the best. Welcome to Team Adamar2	1	3	3	3	11
10	Hong Kon	Senior Consultant at Hays	2208	24	3	reshared a post 21h21h FINALL CALL for our Webinar tomorrow on hashtag#	1	3	3	3	4
11	Hong Kon	Senior Advisor to McKinse	1236	27	3	reshared a post 4d4d Interesting post Zachary Davis. Amazing statistics and n	1	3	2	3	6
12	Hong Kon	Hong Kong International h	634	2	4	was at the docking of this stunning lady today, what beautiful lines.... hashtag#	0	3	1	3	12
13	Singapore	CEO at Mare Maritime	2770	1	1	posted this 1mo1mo We need edible sunflower oil 15000 MT / month. CIF Djib	1	2	3	3	12
14	Singapore	Maritime Recruitment Con	16023	1	1	posted this 2d2d Caliber8 is hiring! Join us to be part of a dynamic and fun wo	1	2	3	3	1
15	Singapore	Executive Search | Talent /	13031	4	1	commented on a post 5d5d that\'s a great perspective, thanks Greg32 2 c	1	2	3	3	4
16	Singapore	Senior Fellow, Maritime Se	1750	0	1	commented on a post 1mo1mo Alex, You did a great job and are going to do a	1	2	3	3	13
17	Singapore	Experienced General Man	1428	2	1	#kudos The pride you take in your work is truly inspiring hashtag#GoingAboveA	1	2	3	3	13
18	Singapore	Managing Director -BLU M	2817	10	2	reshared a post 1w1w BMC is expanding its horizons beyond any boundaries	1	2	3	3	6
19	Singapore	Specialist Recruiter for Shi	1370	4	1	commented on a post 2w2w Congratulations! Happy to work with you my frie	1	2	3	3	12
20	Singapore	General Manager | Maritir	2529	1	1	commented on a post 1w1w Thanks for efforts, this was much needed46	1	2	3	3	11

Figure 6: Topic allocation for accounts.
The results were sorted according to high RFF scores.

The layout of the visualization is that there is a global topic view on the left, and the bar charts on the right. There were 15 topics as chosen to separate from. This topic number was a result of taking the UMass coherence score. The UMass coherence score calculates the frequency of how two words wi and wj appear together in a corpus. The UMass coherence score for 15 topics was -12.979. Hence, the topic modelling was used to separate into 15 topics. The various topics and some of their words would be shown in table 1 below where the topics and their various keywords were used to decipher what the topic each profile was allocated was talking more about. The output of the LDA is the probability of how much an account had mentioned the words in the topic. The probability out of the highest of the topics would be allocated to the profile. Hence, a new column was appended to the profiles to identify the topics these users were posting.

How to cite this article: Yuwei H, Carol Anne H. Leveraging Social Media Data and NLP to Identify Key Leaders in AI and Digital Transformation in the Maritime Industry. Trends Tech Sci Res. 2023; 6(5): 555697. DOI:10.19080/TTSR.2023.06.555697

Table 1: LDA Topics.

Topic	Keywords	Topics
1	company, recruitment, bulk, hiring, leader, leadership, vessel, change, data, crew	Company Leadership & Crew
2	safety, market, ship, change, talent, data, asia, benefit, fuel, digitaltransformation, port	Data, Digital transformation & Change
3	grain, export, global, safety, Ukraine, port, analyst, modernization, market, security	Global export & Safety
4	technology, professor, success, journal, talent, associate, group, award, collaboration, event	Technology, Collaboration, Professor & Journal
5	candidate, interview, salary, market, recruitment, port, company, experience, hiring, technical	Candidate Recruitment, Interview & Salary
6	vessel, peace, award, event, market, change, bunkering, transparency, security, carbon, reduce	Bunkering Transparency & Carbon Reduce
7	grain, global, export, Ukraine, safety, analyst, future, intelligence, drybulk, food	Grain, Drybulk, Food & Intelligence, Analyst, Future
8	oil, marine, vessel, safety, engineer, project, professional, discussion, machine, resolve, system	Engineering System Discussions & Professional Resolutions
9	investment, event, LinkedIn, forecast, report, launch, candidate, report, hiring, recruitment	Investment Forecast and Event Report
10	school, group, specialist, risk, energy, warranty, professor, technology, cyber, shipbuilding	Energy, Technology, Cyber, Shipbuilding & Risk
11	event, safety, covid, ship, contact, pandemic, marine, humanfactors, social, contact	Covid Pandemic, Social Contact and Human Factors
12	green, project, opportunity, decarbonisation, future, data, technology, sustainability, ecosystem, plan	Sustainability, decarbonisation, Green Ecosystem
13	conference, project, meeting, trading, convention, trader, blog, petrochemical, role, register	Conferences, Conventions, Meetings, & Blogs
14	risk, director, engineer, energy, challenge, learning, leadership, experience	Risks and Challenges, Director, Leadership and Learning
15	safety, strategy, science, article, support, safe, insightful, institute, tech	Safety, Science Strategy, Insightful – Institute, Tech

RFF Process

After finding out the topics, we used the RFM modelling method to classify influencers. RFM modelling uses Recency, Frequency and Monetary value to assign a firm's customer base on a particular trait [10]. RFM modelling was used to analyse customer purchase behaviour by segmenting customers according to their purchase behaviour. [11] RFM is most commonly used in marketing analysis

How to cite this article: Yuwei H, Carol Anne H. Leveraging Social Media Data and NLP to Identify Key Leaders in AI and Digital Transformation in the Maritime Industry. Trends Tech Sci Res. 2023; 6(5): 555697. DOI:10.19080/TTSR.2023.06.555697

to segment customers. In this context, the RFM modelling used in this project in 3 aspects are Recency - how recent the profile had posted their last post, Frequency - how frequent an account had posted their posts within a period of 3 months, and their Followers count - how many followers they have. Each category has a ranking of 3, for example in Recency with 1 being the least recent and 3 being the most recent. Frequency with 1 being the least frequent and 3 being the most frequent. Followers are also classified into 3 rankings, with 1 having the least number of followers and 3 with accounts having the greatest number of followers. Because we use 'number of followers' instead of 'monetary value', we will replace RFM with RFF. Table 2 below indicates how the RFF scores were assigned to the LinkedIn professionals.

Table 2: Adapted RFF model for social media analysis.

Recency	Frequency	Followers
R-3 (Most Recent at posting, 1 month, 1 week etc, earlier than 2 months)	F-3 (Most frequency of posts in 3 months, >= 3 posts)	Fol-3 (More than the 75th percentile)
R-2 – (Last post to be 2 months ago to 3 months ago)	F-2 (1-2 posts in 3 months)	Fol-2 (Between 50th percentile to 75th percentile)
R-1 (Least Recent at posting – have not posted in the 3 months)	F-1 (Least frequency of posts in 3 months, 0 posts in 3 months)	Fol-1 (Up to 50th percentile)

Next, we explore the high RFF score influencers to better understand their LinkedIn behavior and the topics they post. For example, we found a key opinion leader with a high RFF score of 333, who posted recently (R), the post counts 4 times in 3 months (F) and had 3751 followers. Looking into this LinkedIn account, the key opinion leader mostly talked about topic 4, which includes, 'Technology', 'Collaboration' & 'Professor and Journal' as allocated by the LDA. This key opinion leader was a Data Orchestration and empowers maritime companies to accelerate their digital transformation journey. The first post discusses three benefits of achieving superior data quality in the maritime industry, including a decrease in demurrage charges, a time reduction in claims collection, and an increase in the bottom line through the prioritization of high-scoring charterers. The second post discusses a deal between Starlink and Royal Caribbean to provide high-speed internet on their ships. The third post talks about the importance of understanding the benefits of maritime data and gave an example of how the outcome of reaching the next port of call can be turned into a measurable benefit by optimizing fuel consumption and CII. The post encourages maritime leaders to think seriously about their maritime data strategy and offers to chat with anyone unsure about their own data outcomes and benefits. Additionally, the fourth post discusses the challenges of Management of Change (MOC) in the maritime industry and offers a method to make employees feel appreciated rather than threatened by technological changes to achieve successful digital transformation. Some of the hashtags that the posts included were #digitaltransformation and #dataculture.

Another key opinion leader, who had a RFF score of 3 3 3 had 2,208 followers and was allocated Topic 5 by the LDA approach, which includes topics such as 'Candidate', 'Recruitment', 'Interview' and 'Salary'. This opinion leader posted about a webinar, calling for people in tech to do a survey to identify trends for recruitment in the region. The post also highlighted the use of AI tools in various products and industries, including Netflix, PayPal, Salesforce.com and Facebook. The post was about the recent Revive Tech Asia event and Panel discussions on hot topics in the tech market, including data, AI, UX design, customer experience, digital, insurtech, and smart city. This key opinion leader had used hashtags such as #DataScience, #AI, #DataAnalytics and #MachineLearning. The results of the RFF modelling and the topics allocated to the key opinion leaders is shown in figure 6 below. The results were sorted according to high RFF scores.

Limitations

One limitation is the information from LinkedIn profiles. LinkedIn limits the number of profiles one can scrape in a day. It is recommended that with a LinkedIn premium account, one is encouraged to scrape a maximum of 150 profiles a day. Thus, for this project, the data obtained is scraped from LinkedIn over several days from 25 August 2022 to 7 September 2022. On LinkedIn, a user can set their profile privacy settings to not be seen by other users. Hence, when obtaining the LinkedIn URLs of LinkedIn profiles, their LinkedIn URL would be masked and not be able to obtain. Hence, the data from these profiles were not scraped as one was unable to access their profiles on LinkedIn. The Activity section indicates the user's recent posts, re-shares of other posts, comments on other posts or likes on other posts. The activity section scraped from these profiles only shows the recent 3 months / 90 days. If a user did not do any of the actions, the Activity section will be shown as "(Name) has not posted lately". Thus, from scraping the LinkedIn's profile page, one is able to access a user's activity of up to 3 months ("Your Recent Activity on LinkedIn"). One limitation of the word cloud was that the words that appear are the frequency it appears in the data and not exactly its importance. Word clouds as it shows only 1 word, it does not provide context so the meaning of individual words may be lost.

How to cite this article: Yuwei H, Carol Anne H. Leveraging Social Media Data and NLP to Identify Key Leaders in AI and Digital Transformation in the Maritime Industry. Trends Tech Sci Res. 2023; 6(5): 555697. DOI:10.19080/TTSR.2023.06.555697

Challenges

One of the challenges faced during the project is the lack of data. Data needs to be scraped from LinkedIn. However, there were limitations to the accounts that one can access on LinkedIn. For example, only 100 pages of profiles will be shown, with each page showing 10 accounts. Accounts after 100 pages cannot be accessed by a normal LinkedIn account. Thus, there were limited profiles one with a normal LinkedIn account can scrape. In this project, a total of 800 accounts were scraped. The number of accounts that were visible for scraping would also depend on the connections of the logged in account to scrape the data. Some profiles are inaccessible as LinkedIn consists of 1st degree, 2nd degree and 3rd+ degree connections. Certain accounts have set their settings to be only visible to various degrees of connections, hence their accounts are not visible to everyone. Accounts that are too distant from the degree of connections would be blocked from scrapping. Hence, a challenge of data scrapping is that not all accounts could be scrapped.

Additionally, the scraping of the posts was only done to captions and words. There are posts that also include the posting of images and videos. Hence posts in the form of images or videos were not able to be scraped to identify what these images and videos are about. The data collected was only limited to words and captions. Thus, when analyzing the topics of the user, only posts with captions were used [12-16].

Future Work

In the future of this work, ways to scrape more accounts could be made possible by having a premium LinkedIn account where there is no limit to how many profiles a premium LinkedIn account can access. LinkedIn also provides a sales navigator account. Being able to view profiles and having access to view more accounts would make it easier to obtain a greater dataset. With more data, one can deliver better results with a wider dataset. As influencers change overtime, due to various reasons such as a change in the content they have posted, a real time scraping of the posts of the accounts could be built, thus this will have the profiles posts to be updated. More posts could be scrapped from each account for a greater dataset of the posts posted by each account. As the current project only focused on the analysis of captions, in the future work, scraping of images and videos and then classifying these images and videos into words would be able to analyze more posts not limiting to only the captions of the posts.

Conclusion

Social media analysis has become a topic of study and increasingly more data is generated everyday with users on these social media platforms. This study narrowed down the influencers that are talking about relevant AI topics and hence provide a list of the top influencers that can be used by company's who are in search of influencers to help them advertise their company brand on social media or to help the company increase the number of participants at a conference. For young graduates who are interested in analytics and AI mentors, the results of this study will be very useful for them.

References

1. Luca, Supervisor, et al. (2022) POLITECNICO DI TORINO Micro Influencer Classifier: An Academic and Economic Approach Candidate Paolo FIORIO PLA. 2022.
2. Bridgstock, Ruth (2019) Employability and Career Development Learning through Social Media: Exploring the Potential of LinkedIn.
3. Shepherd, Jack. "40 Essential LinkedIn Statistics You Need to Know in 2022." The Social Shepherd, 3 Jan. 2023, thesocialshepherd.com/blog/linkedin-statistics.
4. Sharkbite. "Identifying Potential YouTube Influencers Using Python." Analytics Vidhya, 18 July 2022, www.analyticsvidhya.com/blog/2022/07/identifying-potential-youtube-influencers-using-python/. Accessed 21 Feb. 2023.
5. Blei, David, et al. "Latent Dirichlet Allocation." Journal of Machine Learning Research, vol. 3, 2003, pp. 993–1022, www.jmlr.org/papers/volume3/blei03a/blei03a.pdf?ref=githubhelp.com. (Blei et al.)
6. Vaidya, Neha. "Selenium Certification Training Course." Edureka, 29 Apr. 2019, www.edureka.co/blog/selenium-chromedriver-and-geckodriver/. (Vaidya)
7. Singhal, Gaurav. "Importance of Text Pre-Processing | Pluralsight." Www.pluralsight.com, 5 Oct. 2020, www.pluralsight.com/guides/importance-of-text-pre-processing.
8. Rastogi, Kashish. "Text Cleaning Methods in NLP." Analytics Vidhya, 31 Jan. 2022 www.analyticsvidhya.com/blog/2022/01/text-cleaning-methods-in-nlp/#:~:text=Removing%20Punctuations&text=The%20punctuation%20removal%20process%20will. Accessed 2 Mar. 2023.
9. Sciences, Waikoloa, HI, USA, 2014, pp. 1833-1842, doi: 10.1109/HICSS.2014.231. (Heimerl et al.)
10. Wei, Jo-Ting, et al. "A Review of the Application of RFM Model." African Journal of Business Management, vol. 4, no. 19, 2010, pp. 4199–4206, academicjournals.org/article/article1380555001_Wei%20et%20al.pdf. (Wei et al.)
11. Murphy, Casey. "Recency, Frequency, Monetary Value (RFM) Definition." Investopedia, 19 Nov. 2022, www.investopedia.com/terms/r/rfm-recency-frequency-monetary-value.asp#:~:text=The%20Recency%2C%20Frequency%2C%20Monetary%20Value%20(RFM)%20Model%20assigns.
12. Utz Sonja, Johannes Breuer (2019) The Relationship between Networking, LinkedIn Use, and Retrieving Informational Benefits." Cyberpsychology, Behavior, and Social Networking 22(3): 180-185.
13. "Web Scraping: Crawling Linkedin Profiles." GitHub, 28 Dec. 2022, github.com/boringPpl/Linkedin-profiles-scraping/.
14. "Your Recent Activity on LinkedIn." LinkedIn Help, www.linkedin.com/help/linkedin/answer/a546122/your-recent-activity-on-linkedin?lang=en#:~:text=If%20you%20haven. ("Your Recent Activity on LinkedIn")
15. "Topic Modelling | Topic Modelling in Natural Language Processing." Analytics Vidhya, 1 May 2021, www.analyticsvidhya.com/blog/2021/05/topic-modelling-in-natural-language-processing/. (Vidhya)
16. F. Heimerl, S. Lohmann, S. Lange and T. Ertl, "Word Cloud Explorer: Text Analytics Based on Word Clouds," 2014, 47th Hawaii International Conference on System.

How to cite this article: Yuwei H, Carol Anne H. Leveraging Social Media Data and NLP to Identify Key Leaders in AI and Digital Transformation in the Maritime Industry. Trends Tech Sci Res. 2023; 6(5): 555697. DOI:10.19080/TTSR.2023.06.555697

DOI: 10.19080/TTSR.2023.06.555697

How to cite this article: Yuwei H, Carol Anne H. Leveraging Social Media Data and NLP to Identify Key Leaders in AI and Digital Transformation in the Maritime Industry. Trends Tech Sci Res. 2023; 6(5): 555697. DOI:10.19080/TTSR.2023.06.555697