Master of Research

Social Media Data Analytics for the NSW Construction Industry: A Study on Twitter

Shalini Mullassery Mohanan - 19419451

Supervisor: Dr Liyaning Tang & Mary Hardie

School of Built Environment

Statement of Authentication

The work presented in this thesis is, to the best of my knowledge and belief, original except as acknowledged in the text. I hereby declare that I have not submitted this material, either in full or in part, for a degree at this or any other institution.

[redacted]

Shalini Mullassery Mohanan

Abstract

The primary aim of this dissertation is to explore the social interaction and relationship of people within the NSW construction industry through social media data analytics. The research objective is to perform social media data analytics through Twitter and explore the social interactions between different stakeholders in the construction industry to understand the real-world situations better. The data analytics was performed on Twitter tweets, retweets, and hashtags that were collected from four clusters on construction stakeholders in NSW, namely construction workers, companies, media, and union. Tweets, retweets, and hashtags that were collected from four clusters on construction stakeholders in NSW, namely construction workers, companies, media, and unions. Social media data analytics is a rather modern data collection method that is fast gaining popularity over the traditional data collection methods, such as interviews, survey questionnaires, and opinion polls. The reason for this is because traditional data collection methods have a number of disadvantages, such as they are expensive to administer, have a limited population representation, are time-consuming, and sometimes contain dubious information. For example, when people are in interviews, some may not give genuine responses due to various factors like fear of judgment, intimidation, nervousness, or may be reserved.

Social media data, on the other hand, contains numerous advantages such as it is inexpensive to collect and analyze, has a wider population representation, has genuine information, and is faster to collect. It is for this reason that this thesis seeks to perform social media data analytics in order to explore and investigate the social interactions and links between the different stakeholders that are present in the construction industry. Investigating these interactions will help reveal a multitude of other related social aspects about the stakeholders, e.g., their genuine attitudes about the construction industry and how they feel being involved in

this field of work. In order to facilitate this research, a social media data analytics study was carried out to find out the links and associations that are present between the construction workers, companies, unions, and media group entities. Five types of analyses were performed, namely sentiment analysis, link analysis, topic modelling, geo-location analysis, and timeline analysis. The results indicated that there are minimal social interactions between the construction workers and the other three clusters (i.e., companies, unions, and the media). The main reason that has been attributed to this observation is the way workers operate in a rather informal and casual manner. The construction companies, unions, and the media define their behavior in a much more formal and corporate attitude, hence they tend to relate to one another more than they do with workers. A number of counteractive approaches may be enforced in an effort to restore healthy social relations between workers and the other three clusters. For example, the company management teams should endeavor to develop stronger interactions with the workers and improve the working conditions, in overall. There are several ways to achieve this, for example promote a healthy work-life balance, reward great work from the employees and provide constructive feedback, provide employee perks and benefits, as well as encourage job rotation and teamwork between various departments.

Contents

List of Figures

List of Tables

Chapter 1. Introduction

1.1 Background to the Study

The construction industry is a major contributor of the economy and the Gross Domestic Product (GDP) in many states today, yet there remains a dearth of literature regarding the industry's sociocultural dimensions. As Tang et al. (2017) state, there is minimal research that investigates the social interactions between the industry's stakeholders or even the cultural practices in construction companies. Majority of the present literature on the construction industry is, instead, limited to the safety risks and hazards that are likely to occur at construction sites as well as discussion of preventive measures for employees (Fang, Wu, & Wu, 2015). Others are concerned with identifying and assessing safety risks, maintaining safety awareness, supervisor-worker safety conduct, construction project tactics, and union participation (Hallowell, 2012). While this present literature about safety identification and management is important in the academia field, there still remains a crucial literature gap regarding the social interactions that exist between the stakeholders in this particular industry.

As Roffey (2016) states, there has been a growing need to learn the social interactions and relationships that stakeholders in an industry share amongst each other. However, learning the sociocultural dimensions of an organization has not always been at the centerfold of most industries' key interests in the past. As Mankins (2017) explains, for the longest time ever, organizations only seemed to focus on business processes such as profitability, prosperity, and shareholder value and worked hard to maximize their potentiality, believing that only these elements could determine the value of the business. For instance, a company would take several measures, e.g., cutting production costs, in order to increase its profitability, or even vamp up its promotional marketing so as to boost sales, dividends, and, eventually, the shareholder value of the company (DeFeo, 2018). Thus, these elements, i.e., profitability, prosperity, and shareholder value served as the chief performance indicators for the business excellence of a company, such that the higher each of these elements was, the more successful a business is (Hardjono & Marrewijk, 2001). For example, the higher the profit margins recorded, the more prosperous and successful that business is.

However, this trend is gradually changing to accommodate the sociocultural dimensions of organizations. While profitability, prosperity, and shareholder value are important aspects in a business, they are not the only performance indicators that determine the overall success and affluence of a company (Hawkins, 2011). Indeed, businesses are starting to acknowledge that sociocultural dimensions also count towards the overall success, performance and productivity, corporate image, brand reputation, business excellence, and performance of organizations (Haar et al., 2019). Learning the sociocultural dimensions of a company is, therefore, fundamental in order to understand the social dynamics between the stakeholders of a company, or organization, and model an organizational culture that best aligns itself to these dynamics (Di Fabio, 2017). Indeed, structuring the organizational culture to fit the stakeholders' sociocultural values is fast becoming a top strategy in creating great workplaces and healthy relationships between different departments within a company (Britt, 2020). De Waal (2018) emphasizes that establishing and maintaining healthy social interactions in an organization is key in ensuring that there is effective team work amongst the employees, trust and commitment, credibility amongst the members, and respect for each other. Strong social interactions also reinforce people's loyalty, communication, shared values and goals, as well as a sense of belonging and togetherness (Blustein et al., 2019). When people trust and respect each other, they are more likely to work together in unity and, thus, elevate the organization's performance, productivity, profitability, and prosperity.

1.2 Problem Statement

As earlier stated, there has been a growing need and concern to better understand the sociocultural dimensions in organizations today in order to appreciate the social interactions between an industry's stakeholders (Hardjono & Marrewijk, 2001). In response to this, numerous research studies and surveys have been conducted in order to investigate this topic area in various other fields of work, e.g., education, healthcare, etc. However, there still lacks substantial research on this topic area in the construction industry. As Tang et al. (2017) state, there is minimal literature that discusses the sociocultural dimensions in the construction industry, i.e., the work practices in construction companies, the relationships between different stakeholders in the companies, the organizational culture, and inclusivity and connectedness of stakeholders (or lack thereof) within this field of work.

As Su, Cong, and Liang (2019) emphasize, majority of the literature published on the construction industry mainly entails topics like safety risk identification and prevention. Studies such as Brandenbury, Jaas, and Byrom (2006), Sun et al. (2008), Hallowell (2012), and Fang et al. (2015) only cover topics like human resources in construction, safety risk identification and assessment, safety knowledge management, and supervisory worker safety behavior. Other research studies such as Fox (1998) as well as Behrens and Jacoby (2004) have covered topics in construction union memberships, construction union programs, and company union strategies, respectively. Brucker et al. (1999), on the other hand, has covered a topic in company project strategies. Thus, for the most part, the aforementioned research studies are only limited to workers' safety and preventive measures as well as construction business processes, such as project strategies etc.

In light of the above, it is, therefore, clear that there is a current literature gap in the topics covered in the construction industry. Having seen the significance of the social interactions in industries today, it is becoming increasingly important for organizations to learn the social relationships and dynamics between their stakeholders in order to adjust their organizational cultures to fit the values and interests of the people. One of the best ways to do this is through conducting a research study on the same. A research study that covers the theoretical overview of this topic area as well as presents the results of a primary data analysis will greatly contribute to the current academia and help fill this literature gap.

1.3 The Aims and Purpose of the Research

The primary aim of this research was to explore and investigate the social interactions and links between the different stakeholders that are present in the construction industry. Investigating these interactions will help reveal a multitude of other related social aspects about the stakeholders, e.g., their genuine attitudes about the construction industry and how they feel being involved in this field of work. For instance, the research could help gain a better understanding of the employees' attitudes towards their fellow colleagues, their employers, and their commitment towards their jobs. The research could also help a better understanding of the corporate company managerial teams and their treatment towards their employees, suppliers, consumers, the government, and the community. In general, the facilitation of this research and thesis will contribute to a more defined understanding of the qualitative attributes of the construction industry, including the attitudes and sentiments of the stakeholders regarding the construction industry, the relationships between the different stakeholders, and even their geographical locations.

1.4 Scope of the Research

The scope of the research will encompass the qualitative attributes of the stakeholders within the construction industry in the NSW Company. The current research acknowledges that there are various categories of stakeholders in an industry and, therefore, categorizes the construction stakeholders into four distinct groups, namely construction workers, construction companies (the management teams), construction unions (e.g., labor unions), and the construction media. The study specifically chose to divide the stakeholders into these four clusters as they are the most common groups in the industry and, therefore, they are effectively representing the rest of the industry population (Tang et al., 2017).

1.5 The Chosen Methodology

The chosen methodology for this research was based on a rather new technique, known as the social media data analytics strategy. As Lawrence et al. (2010) explain, social media data analytics is a rather modern method, which entails the analysis of data that is extracted from social media networks, e.g., Facebook, Twitter, Instagram, LinkedIn, and Google +. According to Andryani, Negara, and Triadi (2019), social media data analytics is a new-fangled research methodology that is gaining fast momentum in the research realm today. The reason for its increasing popularity in modern-day research is because it enables researchers to collect big data, on a specific topic area, and conduct data analysis in a rather fast and inexpensive manner (Crimson Hexagon, n.d.). Social media data analytics is fast becoming a favorite research methodology and replacing the other traditional research techniques, such as survey questionnaires, interviews, opinion polls, case studies, participant, and non-participant observation, etc. (Dhawan & Zanini, 2014). As Tang et al. (2015) explain, methodologies such as participant and non-participant observation, questionnaires, interviews, and case studies are extremely time consuming as they require researchers to schedule time with human respondents and solicit their permission to partake in a research study. There are also other risks associated with these traditional techniques, such as dishonest responses or data, risk of physical or psychological harm, fear of prejudice or intimidation by participants especially during interviews, and the risk of participants dropping out of the research (Buntain et al., 2016). Wells and Thorson (2015) also state that the traditional techniques are quite costly, due to administration costs, transportation, and other overheads.

Social media data analytics, on the other hand, is a cheap and inexpensive method, since there are no overheads, such as transportation to the participants (Holsapple, Hsiao, & Pakath, 2014). The data is, instead, extracted right from social media content on Facebook, Twitter, Instagram, etc. (Bradbury, 2013). There are various tools available that can be used to extract this content from social media platforms, for example the Twitter APIs like Tweepy, API v2, and the RESTful API, which are all free and publicly available from the Twitter site (Spaiser, 2016). As Jurgens and Jungherr (2016) explain, these Twitter APIs can be imported into various programming languages, such as Python, Java, and R, and then run-in order to pull the necessary content from the Twitter platform. Facebook, too, has its own APIs, for example the Graph API and HTTP-based APIs, which can be used to pull data from Facebook into the programming languages for purposes of research and drawing meaningful insights and data patterns (Kumar, 2015).

For this particular research, we used the Tweety API to extract the most recent 401 tweets per each cluster, i.e., Construction companies, workers, media, and unions, from the Twitter social media platform. The reason why content from Twitter was used was because the content on Twitter is more geared towards industry matters, than is Facebook and other social media sites (Tang et al., 2017). Swanner (2016) asserts that the Facebook APIs are restricted to collecting only content associated with news and celebrities. The reason for this is because users on Facebook share more personal data than on Twitter, therefore, Facebook has stricter data privacy policies, compared to Twitter. In the contrary, users on Twitter tend to share more public information regarding their opinions about public matters, e.g., politicians, companies, celebrities etc. (Stack Overflow, 2012). Therefore, Twitter APIs are less restricted in the collection of data and, therefore, provide content that has more insights than Facebook and other sites.

1.6 Rationale of the Research

The rationale, or justification, of conducting this research is twofold. The first justification is based on the need to investigate and explore the sociocultural dimensions within the construction industry. Investigating these dimensions will help gain a better understanding of the social values of the construction workers, construction companies, construction unions, and construction media. The investigation will also help reveal the social relations that currently exist amongst the four clusters of stakeholders within the construction industry. In addition to this, the investigation shall also help to learn the attitudes and sentiments held by these four clusters about the construction industry they work in, the types of topics they discuss most often on Twitter, the timelines they share their opinions on Twitter, as well as their geographical locations. In general, these findings will greatly contribute to the academia field of sociocultural dimensions of the construction industry and help fill the current literature gap. Furthermore, learning these dimensions will also enable the construction companies to design workplace practices and cultures that are best aligned with the sociocultural values of their stakeholders.

The second rationale for conducting this research is make use of the new social media data analytics technique. As earlier stated, social media data analytics is a rather new research method that is fast gaining popularity over the other traditional research techniques, thanks to its cheap affordability, convenience, and extraction of big data from social media platforms. Integrating this new research method in this research will help me (the researcher) gain a practical hand-on skills and knowledge in this particular field. Furthermore, there is currently no research publication that has used social media data analytics to analyze the sociocultural dimensions in the construction industry. Therefore, this research will also fill this research gap and provide a source of reference for future researchers and scholars who may use this thesis as guidance in performing social media data analytics.

1.7 The Research Objectives

The following are the research objectives that will guide the process of this research:

i. To fill the literature gap in the lack of research on sociocultural dimensions in the construction industry

ii. To investigate the social interactions and relations between the four chosen stakeholder clusters in the construction industry, i.e., Construction workers, firms, unions, and the media in the construction industry.

iii. To use social media data analytics technique to conduct a sentiment analysis, topic modelling, link analysis, geo-location analysis, and time analysis on Twitter messages published by the four stakeholder clusters in the construction industry

1.8 Organisation of the Thesis

This thesis is divided into five chapters, namely the literature review, methodology, data analysis, results and findings, and finally discussion and conclusion. The literature review will provide a detailed and in-depth review of relevant topics, e.g., applications of social media data analytics as well as the benefits and significance of social media data analytics in the construction industry. There will also be a theoretical discussion of the various analysis techniques such as sentiment analysis, topic modelling, link analysis, geo-location analysis, and timeline analysis.

The methodology chapter, on the other hand, will give a step-by-step procedure of how the social media data analytics technique was carried out. The topics that shall be discussed in this section will include the research design and research approach, the dataset used, the materials and tools used in the research, the sampling technique, and research ethics and guidelines. The next chapter after this will be data analysis, which will openly state the source code used during the social media data analytics process. The source code will be derived from the R programming language, where the Tweety API packages will be imported in order to extract the construction-relevant tweets from Twitter into R, for further data analysis purposes, e.g., sentiment analysis.

Once the data analysis process is completed, the results will be presented in the next chapter, which will be the findings and discussions chapter. This chapter will breakdown the meaning of the results that shall be obtained from the R program and discuss them in relation to the sociocultural dimensions of the construction industry. Finally, there shall be a conclusions chapter, which will give a brief recap of the aims and objectives of the study as well as the findings and their implications in the current topic area. There shall also be a mention of the limitations that may have been encountered during the study as well as a mention of the recommendations that the study makes to future researchers.

Chapter 2. Literature Review

2.1 Introduction to the Chapter

This chapter seeks to provide a detailed and in-depth review of relevant topics, e.g., applications of social media data analytics as well as the benefits and significance of social media data analytics in the construction industry. There will also be a theoretical discussion of the various analysis techniques such as sentiment analysis, topic modelling, link analysis, geo-location analysis, and timeline analysis. The purpose of all these discussions is to, at least, gain familiarity of other similar research studies that discuss the given topic area and make sense of their findings.

2.2 Significance of social media In Data Analytics

As Fan and Gordon (2014) assert, the advent of social media data analytics proves to be very beneficial for researchers and current world affair analysts. Social media data analytics has been made possible by the rapid advances in technology (BBC, 2013). Advances in technology have truly revolutionized so many aspects of life and given rise to digitization, at a global level. One of the fields of technology that has greatly advanced in terms of innovations and real-world applications is the information technology. Information technology has transformed the world into a global village through the introduction of the Internet and other online platforms, e.g., social media networks, video conferencing, etc. (Stieglitz & Linh, 2013).

According to IBM (2013), social media networks have particularly gained overwhelming popularity all over the globe, thanks to their instant communication channels, wide audience reach, direct connections with other peers, effective platforms for networking and partnerships, and real-time information sharing. Social media has disrupted numerous appliances and sectors of business, for example print media, television and radio, fax, email, voice calls and messaging, as well as traditional advertising, e.g., brochures, business catalogues, and flyers, etc. (Bradley, 2010). People, today, are using social media for numerous purposes, e.g., communication, peer-to-peer connections, news broadcasting, targeted marketing and advertising, information sharing, market research, education, entertainment, and brand awareness and promotion, just to mention a few (Alzahrani, 2016). Consumers, on the other hand, utilize social media to freely express their honest opinions and feedback about an organization or company's products and services, customer treatment and service, and overall experiences (Batrinca & Treleaven, 2015). Thus, social media serves as a platform for product reviews and feedback and public rating and approval of companies.

2.3 Use of Social Media by Construction Stakeholders

This section seeks to define the various ways and means that the four construction clusters, i.e., workers, company management, unions, and the media, use social media for their benefits.

2.3.1 Construction workers

Construction workers, or rather the employees, use social media mainly as a platform to freely express their opinions and sentiments about the corporate world, workplace issues, new organizational structures, workplace ethics, and their working conditions at their current jobs (Bizzi, 2018). As Adzovie, Nyieku, & Keku, (2017) state, workers always tend to turn to social media to share their true feelings about their workplaces, as most of them may shy away from approaching their bosses or supervisors in person. The reason for this is because in-person discussions may cause chaos at the workplace due to fear of judgment and intimidation (Jafar et al., 2019). However, the case is different for social media. On social media, there is the aspect of anonymity where workers can create pseudo profiles and freely share their experiences at their jobs. Here, they can talk about the treatment they receive from their bosses and supervisors, the status of their working conditions, and even whether they receive fair monetary compensation for their labor (Adzovie et al., 2017). They can also reveal whether there is job security at their workplaces, employee motivation, employee turnover, cultural diversity, and other shared organizational values, such as trust, respect, credibility, equality, commitment, and loyalty from the management, suppliers, customers, vendors, and amongst themselves.

Workers also use social media to create relations and bonds with their colleagues as well as their management (Deng et al., 2017). Jafar et al. (2019) state that employees use sites like Google Plus, Google Hangouts, Twitter, and Facebook to interact with their workmates, while in the office, and share ideas, concepts, and perspectives on various business-related aspects. Some of the ideas they mare share together include problem-solving solutions, new product service concepts, PowerPoint presentations, business memos, and even industry news (Bizzi, 2018). They can also communicate their attitudes about the organizational culture, goals and achievements, worker treatment, equality, and matters concerning fair compensation to their immediate management teams. According to Groen, Wouters, and Wilderom (2017), such interactions actually help to strengthen the relationships between the employees and the management, as well as the relations amongst themselves. For this reason, Deng et al. (2017) insist that those employers who believe that social media is a productivity killer during working hours and block access to social media, should stop doing so, and instead encourage it, so as to encourage employee relations.

2.3.2 Construction Companies

In the business corporate world, construction companies are increasingly using social media to fulfill their business process strategies, such as brand recognition, brand awareness and promotion, provide customer service and support, and marketing of their products and services online (Siricharoen, 2012). Organizations are now creating corporate profiles and even business pages on sites like Facebook, Twitter, and Instagram in a concerted effort to increase their online presence and establish solid connections with their customers, investors, suppliers, vendors, and the community at large (Agbaimoni & Bullock, 2013). As Abuhashesh (2014) states, it is no longer an option but a core necessity for companies to have an online presence, especially on social media sites. This is in order to increase their online traffic and get more eyeballs on their products and services, strengthen their corporate image and brand identity, generate leads for their sales, and communicate about critical issues, e.g., company news, new products and services, corporate events, price discounts, new projects, etc. (Castronovo & Huang, 2012). They also use social media to conduct social listening, i.e., monitoring the conversations that are relevant to their brands, all to identify the sentiments of their stakeholders and adjust accordingly (Bruhn, Schoenmueller, & Schäfer, 2012; Castronovo & Huang, 2012). Social listening also enables corporate entities to conduct market research, stay atop of industry news, as well as keep an eye on the level of competition from rivalries.

2.3.3 Construction Unions

Construction union's also increasingly using social media to engage in conservations with the members of these unions, particularly the employees of construction companies (Pasquier & Wood, 2018). The conversations can be either public or private personal discussions on Twitter, Facebook, LinkedIn, etc. The reason why unions are increasingly preferring social media to the traditional media strategies is because the former allows for unmediated communications and dialogue across vast geographical distances, are inexpensive, and enables faster, easier interactions (Hodder & Houghton, 2015). Unions can now use social media sites to campaign their motives and agendas on a global level as well as raise awareness on certain matters to their members (Carneiro & Costa, 2020). Social media sites, such as Facebook, allow unions to create fan pages where they can directly publish updates to their fans home page and build user interactions with their members (Hodder & Houghton, 2019).

2.3.4 Construction Media

The media, on the other hand, use social media to publish press releases concerning various business news, such as new product launches, upcoming company events, project strategies, business startups, etc. (Hanley, 2014). Press releases published by the media tend to have a wider audience reach because their social media profiles have many followers and fans, which makes their communication more effective (Barasa, 2012). For this reason, the news media firms are often approached by many corporate organizations to help promote their business appearance, brand recognition and awareness, corporate image and brand reputation, company official statements, and public relations strategies, etc. (Hanley, 2014). Construction companies, for example, could request the media to publish press releases to the public about new project strategies, real estate investments, construction safety summits, building expos and exhibitions, innovation conferences, new architectural designs, etc. (Tang et al., 2017).

2.4 Benefits of Social Media Data Analytics

As previously discussed, there are numerous ways how people use social media. The previous section discussed how the media, workers, unions, and companies in the construction industry use social media to pass information and share their experiences. This, therefore, shows that there is a widespread use of social media, which, in turn, makes social media an ideal source of information for performing any worthwhile research on the construction industry. As Leung, Yu, and Liang (2013) state, social media data analytics provides numerous benefits over the traditional research techniques, as discussed below:

2.4.1 Cheap and Inexpensive

Social media data analytics is truly a cheap and inexpensive process compared to other traditional research methods. Working with traditional methods like interviews and questionnaires would include a variety of expenditures, including travel charges if in-person interviews are conducted, participant incentives, transcribing costs, and so on (Nguyen, 2014). Other methods, such as participant observation, would require the researcher to incur additional resources such as lodging, food, and personal utilities, just to mention a few (Comcowich, 2017). On the contrary, social media data analytics does not require one to incur any expenses (Staff et al., 2016). One can simply extract data from any social media platform using free APIs (for example Facebook, Twitter, Instagram, YouTube, Reddit, Snapchat, Pinterest, Tumblr, and Viber APIs) and use it for further analysis purposes. The only expenses that may be incurred when using social media data analytics is if one wants to use sophisticated highly advanced APIs, which have a higher level of data access and reliability (Spaiser, 2016).

2.4.2 Wider population representation

Social media data analytics has a wider population representation due to the big volumes of data that are extracted from social media. According to Kalil (2012), there is a lot of big data available on social media, which means that users are continuously sharing information on social media every second and every minute. The amount of content on social media is increasingly high in volume and yet continues to grow exponentially with high velocity every passing minute (Hoit, 2013). For example, Facebook generates at least 4 petabytes of data daily, which mainly constitutes of 136,000 picture uploads, 510,000 comments, and 293,000 status updates every 60 seconds (Sahitreddy, 2020). Twitter, on the other hand, receives at least 400 million tweets on a daily basis (Brantner & Pfeffer, 2018), while Instagram receives 95 million photos daily

(Gousios, 2020). These numbers only mean that there are millions of people who share their experiences on social media, and not just a few selected people. Therefore, the data that is retrieved from social media is representative of a wider population, and, therefore, the results are more reliable and generalizable (Mayer-Schönberger & Cukier, 2013).

2.4.3 Faster and More Comprehensive Data Collection and Analysis

Social media data analytics also enables faster data collection and analysis. Retrieving data from social media platforms using the API features takes only a matter of minutes and can extract big volumes of data (Kallback, 2019). The situation is however different when working with other traditional research methods. Conducting interviews, for example, can take days, weeks, or even months, especially if the interviewees are on a tight schedule and so unavailable (Icha & Agwu, 2015). Even then, one will have only conducted a limited number of interviews with interviewers who are readily available. In the case of survey questions, the situation is identical. Due to time limits and other variables, only a few hundred surveys may be completed (Anavizio, 2019). Social media data analytics is, however, different as there are no constraints in time, administration costs, geographical distance, or space. For example, this research study collected at least 401 construction-relevant tweets from each of the four aforementioned clusters, i.e., workers, companies, unions, and the media. In total, there were at least 1,604 tweets that were extracted from Twitter alone in a few minutes. None of the other traditional research methods could ever manage to collect all these tweets within a day, let alone hours. Also, data analysis is fast since it is done using programming languages.

2.4.4 Genuine Content

The content found on social media tends to be more genuine and truly representative of people's feelings and experiences, than the data collected using the traditional research methods

(Woldesenbet, Jeong, & Park, 2016). As Das and Lall (2016) explain, when people are in interviews, some may not give genuine responses due to various factors like fear of judgment, intimidation, nervousness, or may be reserved. The situation, however, changes when it comes to social media content. While on social media, people tend to drop their guards and share their true feelings, opinions, views, and beliefs about a given phenomenon (Jiang, Lin, & Qiang, 2016). Therefore, analyzing content from social media may, in fact, give results that are more genuine, accurate, and authentic.

2.5 Types of Social Media Data Analytics

As Tang et al. (2017) assert, there are several types of social media data analytics, for example sentiment analysis, topic modelling, link analysis, geo-location analysis, and timeline analysis, amongst many others. This research seeks to conduct these five types of social media analytics to establish the social relations between the four aforementioned clusters within the construction industry. Therefore, the subsections below first describe these five social media analytics in detail to facilitate an understanding of their theoretical frameworks.

2.5.1 Sentiment Analysis

2.5.1.1 Background into sentiment analysis

Sentiment analysis refers to the computational evaluation of natural human language to determine the opinions, emotions, and attitudes that inherent within that given text (Medhat, Hassan, & Korashy, 2014). As Farhadloo and Rolland (2016) state, sentiment analysis is a subfield that falls under text classification and natural language processing (NLP). NLP, on the other hand, is a blend of machine learning, artificial intelligence, and linguistics. The central role of NLP is to enable computers to read, analyze, decipher, understand, and modify any form of natural language that is only readable to humans (Kharde & Sonawane, 2016). As is known, computers are programmed to only read and understand data in form of binary digits, i.e., 0 and 1. Therefore, all forms of data must first be converted into binary codes by the programming languages, before any computer can read the data, process it, and produce the required results.

However, the advent of information technology in the modern day is generating big volumes of human-readable data, which is also known as natural language (Taboada et al., 2011). Platforms like social media networks, web forums, review sites, political debates, stock market exchanges, microblogging sites, and the media have become hubs for natural language. People are increasingly using these platforms to post their honest views, opinions, attitudes, and sentiments about various phenomena around them, hence contributing to an already-existing large pool of human-readable data (Khan et al., 2016). This large pool of data has developed into now what is known as big data, whereby there is big volumes of data that is constantly posted on these platforms at an increasingly high velocity (Jaafar, Al-Jaadan, & Alnutaifi, 2015). New streams of data keep coming in every minute, every second. For example, as previously highlighted, Facebook generates at least 4 petabytes of data daily, which mainly constitutes of 136,000 picture uploads, 510,000 comments, and 293,000 status updates every 60 seconds (Sahitreddy, 2020). Twitter, on the other hand, receives at least 400 million tweets on a daily basis (Brantner & Pfeffer, 2018), while Instagram receives 95 million photos daily (Gousios, 2020).

The emergence of big data in the form of natural language creates the need for new techniques that can read, decipher, and process this type of data, rather than simply binary digits (Guilel & Boukhalfa, 2015). So far, there have been tremendous efforts to process natural language through development of machine learning and artificial intelligence. Efforts in artificial intelligence have enabled machines to simulate human behavior and perform tasks like a human being, while development initiatives have enabled computer systems to build applications that can perform complex tasks on their own, without being programmed to do so, by simply learning from past data (Jaafar et al., 2015). Progresses in machine learning and artificial intelligence have given birth to the subfield of natural language processing (NLP) that makes it possible for computer systems to read and understand big data documents in form of natural language data and, later, extract meaningful phrases and insights from them (Cambria & White, 2014). There have been many applications of NLP so far, for example machine translation, chat bots, optical character recognition (OCR), speech recognition, speech synthesis, text processing and document classification, topic segmentation or topic modelling, named entity recognition, and sentiment analysis.

All these applications rely on NLP technologies to process and analyze natural human language, like English, Spanish, Arabic, Russian, etc. Speech recognition, for example, uses NLP algorithms, such as Hidden Markov Model (HMM), to capture speech, convert it into a digital format using a sound card, preprocess it, and output a response (Cambria & White, 2014). Document classification, on the other hand, involves processing human-readable documents or texts in order to classify them into different categories, such as technology, sports, entertainment, etc.; the end goal of classifying documents in this way is so as to make it easy to sort and manage a big pool of documents that have varied topics (Khan et al., 2016). Similarly, sentiment analysis analyzes natural language data to establish whether the document has a positive or negative score (Kharde & Sonawane, 2016). As Farhadloo and Rolland (2016) claim, sentiment analysis is becoming increasingly popular in the modern-day corporate world as companies are using this technique to evaluate their customers' feedback and draw insights on how best to improve their product and service delivery. Sentiment analysis is, however, not limited to customer feedback and reviews. It is can also be used to learn the opinions, attitudes, emotions, and the positions employees, suppliers, vendors, investors, and even the managerial teams have towards a given organization in general, an industry, or entity (Medhat et al., 2014).

This thesis employs sentiment analysis to gain an in-depth understanding of the opinions and attitudes the construction workers, companies, unions, and the media have towards the construction industry in general. The application of sentiment analysis, in the present case, is to, therefore, make known whether these four clusters have a positive or negative outlook on the construction industry, or not. A positive outlook is desirable as it lays the foundation of commitment, hard work, and total devotion to work and be productive within the industry, and vice versa. A positive mentality also begets favorable working conditions, enthusiasm, employee motivation, and healthy interactions. A negative mentality and attitude, on the other hand, can only mean that there is unhealthy working conditions, toxic relations which may even be non-existent, non-realistic expectations and targets, and dysfunctional organizational cultures at the workplace.

2.5.1.2 Sentiment analysis algorithms and techniques

Sentiment analysis can be facilitated using several algorithms and techniques (see Figure 1 below).

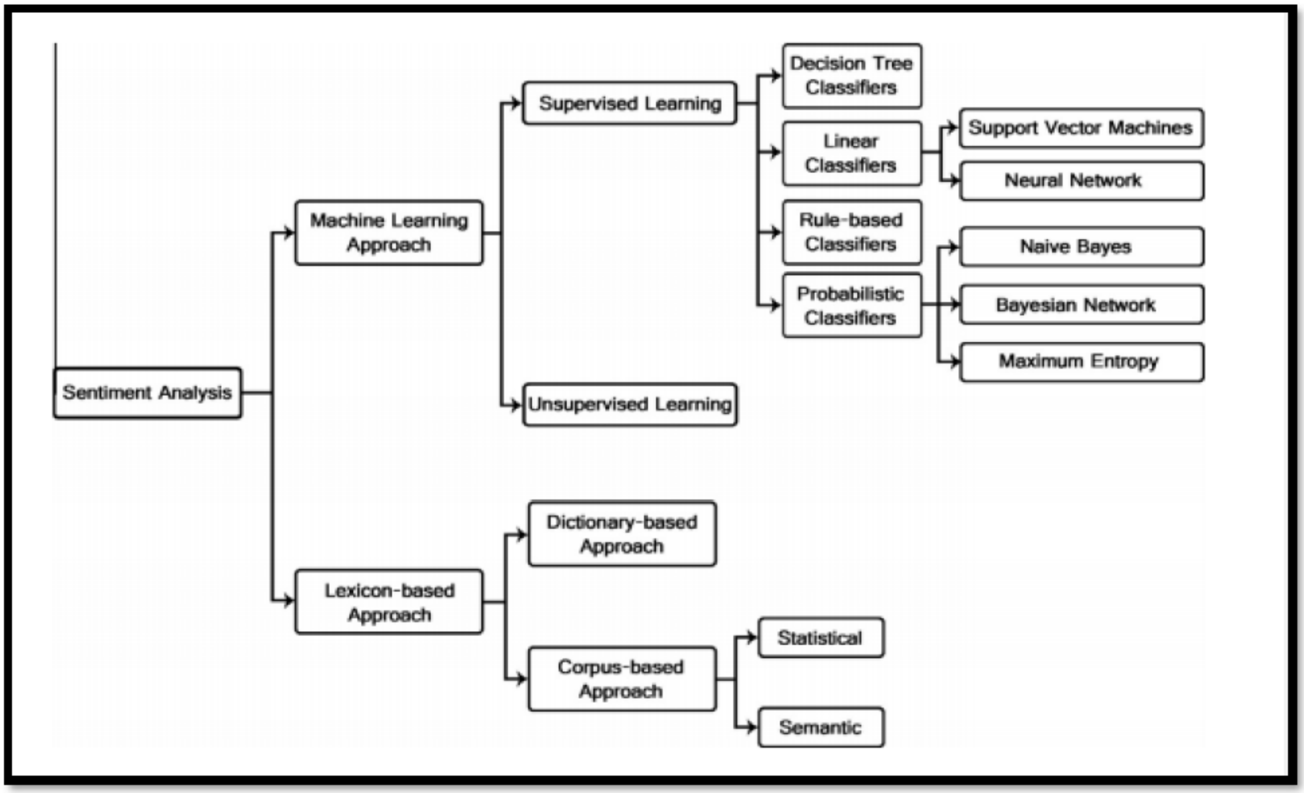


Figure 1: Sentiment Analysis Algorithms and Techniques

Source: Medhat, Hassan, and Korashy (2014)

As can be seen in Figure 1 above, sentiment analysis algorithms can be generally divided into two broad categories, namely the machine learning approach and the lexicon-based approach. As Feldman (2013) explains, the machine-learning approach uses both syntactic as well as linguistic features to analyze the sentiments within different texts and documents. Generally, there are two classifications of machine-learning sentiment analysis, namely the supervised and unsupervised learning. The difference between these two classifications is that in supervised learning, the sentiment analysis process is done dependent on a large number of labelled training documents (Kharde & Sonawane, 2016). However, there are often times when these labelled training documents are not easily available or also when the data being processed is not organized in a structured manner, e.g. Twitter tweets, Facebook posts, etc. In such cases, unsupervised learning is performed (Medhat et al., 2014). There are many examples of supervised learning algorithms, including decision tree classifiers, linear classifiers (e.g., support vector machines and neural networks), rule-based classifiers, and probabilistic classifiers (e.g., naives bayes, Bayesian network, and maximum entropy) (Farhadloo & Rolland, 2016).

The lexicon-based approach, on the other hand, uses lexicons or corpora that contain a collection of pre-selected and pre-compiled sentiment words and phrases (Maks & Vossen, 2012). As Jalayer Academy (2017) explains, there are lexicons that contain purely negative terms, e.g., bad, ugly, disrespectful, annoying, uncomfortable, unpleasant, etc. Other lexicons contain only positive terms, e.g., beautiful, pretty, pleasant, welcoming, refreshing, comforting, wonderful, etc. Therefore, when one wants to analyze the sentimental score of as given document or text, they would just need to compare that document with either a positive or negative lexicon, or even if need be. The purpose of this comparison is to assess the number of words or phrases that given document contains that may be similar or synonymous to the sentimental terms in the lexicons (Jalayer Academy, 2017). If the document has a higher number of negative terms than positive terms, then it is considered generally negative, and vice versa. If it is negative, the sentiment score will be given as -1, while if it is positive, the sentiment score will be given as 1 (Jalayer Academy, 2017). However, if it is neutral (i.e., does not contain either positive or negative terms), then the sentiment score will be given as 0.

As can be seen in Figure 1 above, there are two broad categories of the lexicon-based approach, namely the dictionary-based approach and the corpus-based approach (statistical and semantic). The dictionary-based approach entails a manual process of collecting opinion words from well-known entities, such as WordNet, and thesaurus (Sadia, Khan, & Bashir, 2018). Any synonyms for these opinion words are also collected and added to the dictionary to form a seed list. As Aung and Myo (2017) expound, the process is quite iterative and continues nonstop until all the relevant words are added to the list and there are no more new words to be found. Afterwards, a manual inspection is conducted to see whether the seed list contains any errors or omissions. While the dictionary-based approach is comprehensive and contains numerous opinion words, Khoo and Johnkhan (2017) state that it is rather limited to detect opinion within single words, and not within an entire sentence. This means that this approach is incapable of detecting the opinion or sentiment of a sentence and, therefore, cannot decipher the contextual meaning of a document (Medhat et al., 2014). This limitation is quite disadvantageous because natural language data is often provided in form of sentences, phrases, paragraphs, and documents. People rarely post reviews and give feedback in single words, which makes the dictionary-based approach unsuitable to use when analyzing sentiments within reviews, personal stories, news articles, user feedback, etc.

The corpus-based approach, on the contrary, is not limited to single words (Medhat et al., 2014). Rather, the approach is designed in such a way that it can detect an opinion, sentiment, or emotion within whole phrases, sentences, and documents. Therefore, the approach is capable of deciphering sentiments within context-specific orientations (Sadia et al., 2018). According to Aung and Myo (2017), the corpus-based approach uses syntactic patterns to determine sentiment in phrases that frequently occur with a seed list of opinion terms. Therefore, the corpus-based approach can be considered an upgrade of the dictionary-based approach, as the former technique adds peripheral terms to the basic seed lists that are developed by the dictionary-based approach. This thesis seeks to use the corpus-based approach to analyze the sentiment scores of the Twitter tweets of the four clusters in the construction industry.

2.5.2 Topic Modelling

Topic modelling is yet another subfield of natural language processing and machine learning. Kherwa and Bansal (2018) define topic modelling as a computerized technique that identifies the topics available in each collection of documents. As Tang et al. (2017) explain, it is generally assumed that all natural language data contains certain topics, or themes, that people seem to place a great emphasis upon as they express their feelings either verbally or in written form. For instance, a man expressing his love for a woman will most likely frame his discussion around topics such as ‘love’, ‘lust’, ‘admiration’, ‘affection’, ‘relationship’, ‘marriage’, and ‘children’. On the other hand, a woman talking about her experiences as a mother will most likely frame her speech around topics such as ‘motherhood’, ‘perseverance’, ‘financial responsibilities’, ‘unconditional affection for her children’, and ‘childbirth’, just to mention a few. In short, it is generally expected that human beings will always formulate their conversations or written texts around certain topics or themes of interest. Also, different people have different themes of interest and, therefore, their natural languages will also have different topics. The purpose of topic modelling is, therefore, to harvest the topics that may be present within a given corpus of documents, texts, articles, and any other form of natural language data.

The identification of these topics greatly helps data scientists to sort through large corpora of data and understand the content and concepts of this data (Alghamdi & Alfalqi, 2015). At first, one may not really grasp the objective of topic modelling if they are dealing with a small proportion of data, such as a word, phrase, a single sentence, or even a paragraph. This is because one can easily tell the topic or thesis statement of a sentence or paragraph in a matter of minutes (Tang et al., 2017). For example, in the sentence “Yesterday, I read a book about the gender discrimination of women in Africa and the Middle East” – one can easily tell that the speaker was focusing on the topic of ‘gender discrimination of women’. However, when faced with big data from social media platforms, web forums, review sites, and news articles, it becomes increasingly difficult to read all the texts and identify the topics therein (Liu et al., 2016). It is for this reason that topic modelling exists to determine the topics present within big data texts.

This thesis will apply the technique of topic modelling to pick out the unique topics that are present within the Twitter tweets posted online by the construction workers, companies, unions, and the media. Identifying these topics will build upon the efforts of the sentiment analysis process. This way, we shall be able to know the topics that are most often spoken about online, on Twitter, by each of the four clusters and know how each cluster feels about each topic. For example, it is expected that construction workers, or employees, are most likely to center their discussions on topics such as ‘salaries’, ‘fair compensation’, ‘working conditions’, ‘employee treatment’, ‘employee motivation’, ‘perks and benefits’, ‘team work’, and ‘job contracts’, just to mention a few. Therefore, performing topic modelling will help confirm whether the topics are present within the construction workers’ tweets or not. On top of that, we shall also be able to know the sentiments the workers have towards each of these topics. For instance, we shall know whether the workers are happy about the salaries they are offered or not. We shall also know whether the job contracts given to the workers are satisfactory or not.

2.5.3 Link Analysis

Link analysis, on the other hand, is a subset of data mining that is designed to establish the connections and associations between the nodes present within a given ecosystem (Olson & Lauhoff, 2019). This thesis will conduct a link analysis in order to learn whether there are any forms of associations or connections between construction workers, companies, unions, and the media. The results of the link analysis will reveal whether there are truly any social interconnections between, and amongst, these four clusters, or whether their relations are non-existent and strained. The link analysis experiment will also tell the strength of the associations between, and amongst, these four clusters – i.e., whether the relations are strong and healthy or are weak and forced (Hevey, 2018). Performing a link analysis in this study is fundamental as it will help to unearth the status of social interactions of the stakeholders within the construction industry. The execution of this type of analysis will generate a graph or chart with several nodes (or vertices) at the peripherals and arrows that connect these nodes (Fouss, Saerens, & Shimbo, 2016). Apart from showcasing the connections between the different clusters, the link analysis network graph will also demonstrate the relationships that exist within each cluster. This means that we shall be able to learn the intra-relations amongst the workers themselves, within company managerial teams, within the unions, and between different media firms. This information is very essential as it will throw light on the types of interactions that people within the same levels have.

For example, knowing about the status of the intra-relations amongst the construction workers will demonstrate whether there are any positive employee interactions at the workplace. If not, then that means that more effort should be put into cultivating more positive employee interactions, since healthy interpersonal relationships are important for improved performance outcomes, increased organizational productivity, and even knowledge transfer amongst the employees (Houston, 2020). Rosales (2015) emphasizes that employees who practice healthy social interactions amongst themselves tend to experience a positive impact on their mental health, mortality, and health well-being. Indeed, certain physiological processes (e.g., immune responses, cardiovascular reactivity, hormonal patterns, and neuroendocrine systems) perform better when one has positive social interactions with other people, especially at their places of work. The end results are higher possibilities of employee engagement, minimal absenteeism, higher dedication, and commitment to work, and lower staff turnover (Tumen & Zeydanli, 2015). At the same, such workers are also motivated to collaborate with each other in teamwork, engage in knowledge transfer and productivity spillover as senior and junior staffs work in unison, and develop strong emotional support for each other.

2.5.4 Geo-location Analysis

According to Bo, Cook, and Baldwin (2012), geo-location analysis is a data mining technique that is used to reveal the geographical position, locality, or venue of a person, thing, or any other perceived entity. As Bo, Cook, and Baldwin (2012) explain, there may be several reasons as to why one may need to make known their physical location or venue, e.g., social media updates, finding local businesses, etc. Companies may also use geo-location information for marketing purposes, i.e., by learning their customers' localities and then modifying their product and service offerings to fit the expectations of the customer segments within that geographic locality. Geo-location analysis is also used by web mapping services and apps such as Google Maps to track users' physical locations and provide them with satellite imageries of where they currently are, aerial photography, route planning, public transportation, and update business review sites (Gidofalvi, 2017).

As Mahmud, Nichols, and Drews (2012) expound, geo-location analysis is made possible by the recent innovations that have been accomplished in the GPS technology and the Internet of Things (IoT). The field of IoT has significantly advanced to have onboard sophisticated computing devices, e.g., Location-Enabled Devices (LEDs) that are able to send and receive data to each other over the internet. For example, the past few recent years have seen the development of various types of LEDs, e.g., Personal Digital Assistants (PDAs) and smartphones, which are fitted with web mapping applications that pick up the user’s past, current, and even future locations (Gidofalvi, 2017). As Roller et al. (2012) further explain, these devices are built with positioning technologies, for example GPS-based positioning, cellular network-based positioning, geo-referenced sensor based positioning, and geo-referenced user entry, just to mention a few. Once activated, these technologies can accumulate and update on the user’s location data as they contain spatio-temporal dimensions.

Since everybody, if not all, owns a smart phone that is built with these features, it is, thus, expected that there is a lot of geo-location information of mobile users, on the internet. It is for this reason that this thesis makes use of geo-location analysis to learn of the users’ localities. The users’ (i.e., construction workers, companies, unions, and the media) usage of Twitter to post their tweets, retweets, and replies makes it even more possible to note their geographic locations, since social media platforms have the function of sharing the users’ geo-location data. Therefore, a collection of the users’ geo-location data will help the readers understand where construction workers, companies, unions, and the media are located around the NSW Company. Also, it will help identify which issues or topics affect which the location the most. This type of information will make it easier to zero in on the problems the construction stakeholders are facing and provide feasible solutions for them, based on the most affected locations.

2.5.5 Timeline Analysis

Just like the geo-location data, social media platforms also have a function for showing the time and date when users post any content online (Gidofalvi, 2017). In addition to this, smart phones, PDAs, tablets, etc. also show the timeline of any data that is added or updated on the internet, thanks to the advancements in the field of IoT (Roller et al., 2012). Performing a time analysis in this thesis is, therefore, fundamental to reveal the time frames during which the construction workers, companies, unions, and the associated media is most active online, particularly on Twitter. Knowing this type of information may be of great use as it will enable the appropriate parties to look out for those periods when these users are actively engaged on Twitter and respond appropriately to their posts, tweets, retweets, and replies. This will increase engagement and social interactions between the four clusters of construction stakeholders online and ease the flow of communication amongst everyone. A timeline analysis will also help evaluate the number of times, or frequency, each cluster tends to post content on Twitter. If it is observed that the clusters post too frequently and are persistent about a particular topic, or have a constant negative sentiment, then this means that that user has a very pressing issue that needs to be addressed immediately. At other times, it could mean that that user may be experiencing a mental health problem that requires immediate treatment or therapy (Tang et al., 2017).

Chapter 3. Methodology

3.1 Introduction to the Chapter

This third chapter of the thesis aims to present a step-by-step technique for the social media data analytics processes used in this research, to investigate the 1,604 most recent Twitter tweets posted online by the construction workers, companies, unions, and the associated media agencies. The chapter is divided into several subheadings, ranging from research approach and design, sampling technique, data collection process, the dataset used, tools and instrumentation, the 5 social media data analytics (i.e., topic modelling, sentiment analysis, link analysis, geo-location analysis, and timeline analysis), and, finally, the ethical considerations of the research.

3.2 Research Approach

Generally, there are two broad categories of research approach, namely qualitative and quantitative. The qualitative approach entails the collection and analysis of data that is narrative in nature (Hameed, 2020). This means that qualitative data is presented as text, which is descriptive in nature, and basically defines people's subjective thoughts, opinions, views, outlooks, experiences, attitudes, and cognitive processes (Yin, 2015). The quantitative research approach, on the other hand, encompasses the collection and analysis of data that is statistical in form, which, therefore, means that quantitative data is numerical in nature and rather objective (Addo & Eboh, (2014). As Hameed (2020) explains, qualitative data is collected and analyzed in cases where a researcher seeks to formulate new theories and answer 'why' questions, while quantitative is ideal for studies that seek to test hypotheses.

Since the current study does not seek to test any hypotheses, then the quantitative methodology is not the ideal approach for this thesis. Instead, the study seeks to understand the socio-cultural dimensions and dynamics within the construction industry. The topic of socio-cultural dimensions is a rather subjective matter as it focuses on people's personal experiences in the industry and their social interactions with each other. Also, the data that is about to be collected will be retrieved from Twitter tweets and will entail the users' subjective attitudes towards the construction industry, their opinions about various topics, e.g., working conditions, organizational performance, and other related affairs. Because of this, the study undertakes a qualitative research approach and design, and not qualitative, to analyze the tweets and, thus, fulfill the aims and objectives of the study.

3.3 Data Set Used for the Study

The data set that was used for this study was three-fold, as elaborated in further details below:

3.3.1 Tweets

The first category of data collected was a batch of 1604 tweets that had been posted on the Twitter social media platform by the four stakeholder clusters of the construction industry. These clusters were namely the construction workers, construction companies, construction union organizations, and construction media. The accumulation of all these tweets (i.e., the 1604 batch) was made up of four distinct categories, each representing 401 tweets for each of the four stakeholder clusters. The tweets consisted of various types of content from the users, such as normal tweets, replies, and retweets. Tang et al. (2017) define normal tweets as any message that is originally posted on Twitter by a user. This message must have at least 140 characters and cannot be a reply to another prior tweet, or even a retweet. According to Tang et al. (2017), normal tweets consist of 3 subtypes, which are solo tweets (original messages that are plain text and no other users have interacted with so far), favorites (original messages that have other users have interacted with by liking them), and hashtags (messages that other users have interacted with by inserting the # symbol). Replies, on the other hand, comprise of those messages that are posted as a response or comment to another prior tweet, by a specific user. These replies usually contain the '@username' symbol. The third category of tweets is the 'retweets', which basically encompasses those messages that have been forwarded by a specific user. As Tang et al. (2017) clarify, in order to generate a retweet, a user has to use the retweet feature on Twitter and also include in it the RT @ symbol.

3.3.2 Users' Graphs

Besides the tweets, the researcher also collected some additional information from Twitter, precisely the users' graphs as well as retweet graphs. According to Tang et al. (2017), a user's graph is a data point that shows the connections one user has with other users on a given social media platform. On Twitter, a user's graph could be evaluated to show the number of followers (those people that follow one person for regular updates on their walls) and followees (those people that a particular person follows in return) that User A has. For example, if User A follows User B, and User C follows User A; then User A could be considered as a follower of User B and a followee of User C. The 'follower and followee' concept, generally, identifies the interconnections of different users on a social media platform, and can, therefore, be used to calculate a ratio known as the follower-followee ratio (Tang et al., 2017). The higher this ratio is, the more popular a user is on that platform, and vice versa.

3.3.3 Retweet Graph

The third category of data collected was the retweet graph from Twitter. According to Tang et al. (2017), the retweet graph is similar to the user graph that is elaborated in the past subsection. The retweet graph is also a set of data points that show how users retweet other users' original/ normal tweets, and, just like the user graph, the retweet graph can also be used to calculate a ratio known as the tweet (normal tweets)-retweet ratio. Also, the higher this ratio is, the more popular a user is on the Twitter platform, and vice versa. This is because people tend to retweet original messages, from popular people and influencers, more often than they would retweet an unknown person. Both the user graph and retweet graph are very significant in this thesis as they will be used in the link analysis process to demonstrate the interrelationships that people in the construction industry have with each other. The more interconnections a person has (as shown from the two graphs), the healthier his/ her social relations are, and vice versa.

3.4 Participants Involved

As earlier highlighted, there were four clusters of stakeholders involved in the study, namely the construction workers, companies, unions, and the media in the construction industry. The preponderant reason for choosing these four clusters was because these categories represent the majority of the stakeholders/ players in the construction industry. The workers constitute of the employees who work within the NSW Company and are given a monetary compensation for their labor. The construction companies, on the other hand, encompass the top management teams of the NSW, whose work is to guide and oversee the development, maintenance, and allocation of resources that enable the employees achieve the organizational objectives and goals (Siricharoen, 2012).

The construction unions are third-party organizations that are formed to campaign for workers' labor rights and regulate the cooperation between the employers and employees (McAllister, 2015). Every worker has a constitutional right to organize, join, or engage in the activities or affairs of a union corporation. Once a worker joins a union organization, the latter assists the former to negotiate for better remuneration from his/ her employer(s), fairer employee treatment and better working conditions, collective bargaining agreements, and refined dispute resolution strategies (Moetti-Lysson & Ongori, 2011). Trade unions also help workers deliberate for procedural hiring and firing, workplace safety and policies (especially in the construction industry which is a labor-intensive industry and exposure to potentially harmful construction sites), and complaint procedures (Carneiro & Costa, 2020). Thus, it is expected that such trade union members, as well as their employees, regularly post their own content on Twitter to address certain issues that they may have about various construction companies, e.g., poor employee treatment, unfair worker dismissal, delayed promotions, member mobilization, trade union member fees, etc.

Lastly, the media was also selected as a suitable participant category, or cluster, to represent those institutions or agencies that publish news about construction-related topics or issues. The inclusion of the media as a suitable participant in this study was because the media have a big client base on all social media platforms, including Twitter. As Hanley (2014) states, the media commands a lot of attention from their clients and followers, due to their news-publishing content that tends to seize the audience’s interests. They, therefore, have a huge following on Twitter as they post plenty of news-worthy content on their Twitter accounts. For example, news on planned corporate events, new product launches, management structuring and restructuring, company scandals, proposed merger, and acquisition (M&A) initiatives, and other construction-related subjects will be published in the construction media (Barasa, 2012). This study can use this media data to get more insight about the construction companies and even the union organizations.

3.5 Tools and Instrumentation

Several tools were used for the analysis of the Twitter tweets. These included the Tweety, Twitter API package, the Python programming language, and a newly developed web-based data crawler. These are elaborated below in further detail:

3.5.1 Tweety

Roesslein (2018) defines Tweepy as an open-source package that is compiled in the Python language and is used to invoke several HTTP endpoints. Invoking HTTP endpoints simply means that the user can easily surpass low-level tasks, such as data encoding and decoding from scratch, conducting data serialization, OAuth authentication, HTTP requests, results pagination, and creating rate limit parameters from scratch (Garcia, 2020). Without Tweepy, one would have to perform all these tasks from the ground up. With Tweepy, however, one can easily make such requests in automated fashion. In this study, Tweepy was used to specifically invoke the Twitter API, from the Python programming environment, which, in turn, gives access to the Twitter platform using the using the OAuth authentication feature.

3.5.2 Twitter API

The Twitter API, on the other hand, acts at the middle point between the Twitter platform and the Tweepy open-source package. The API receives an invoke request from Tweepy and, thereafter, makes an application to the OAuth authentication feature – which is basically also an open-source authorization protocol – to authenticate the invoke request from Tweepy (Garcia, 2020). Once approved, the API can now siphon various Twitter entities, such as tweets, retweets, direct messages, favorites, trends, users, likes, and hashtags. For this study, the API was used to extract only three types of content from Twitter, i.e., tweets, user graphs and retweet graphs.

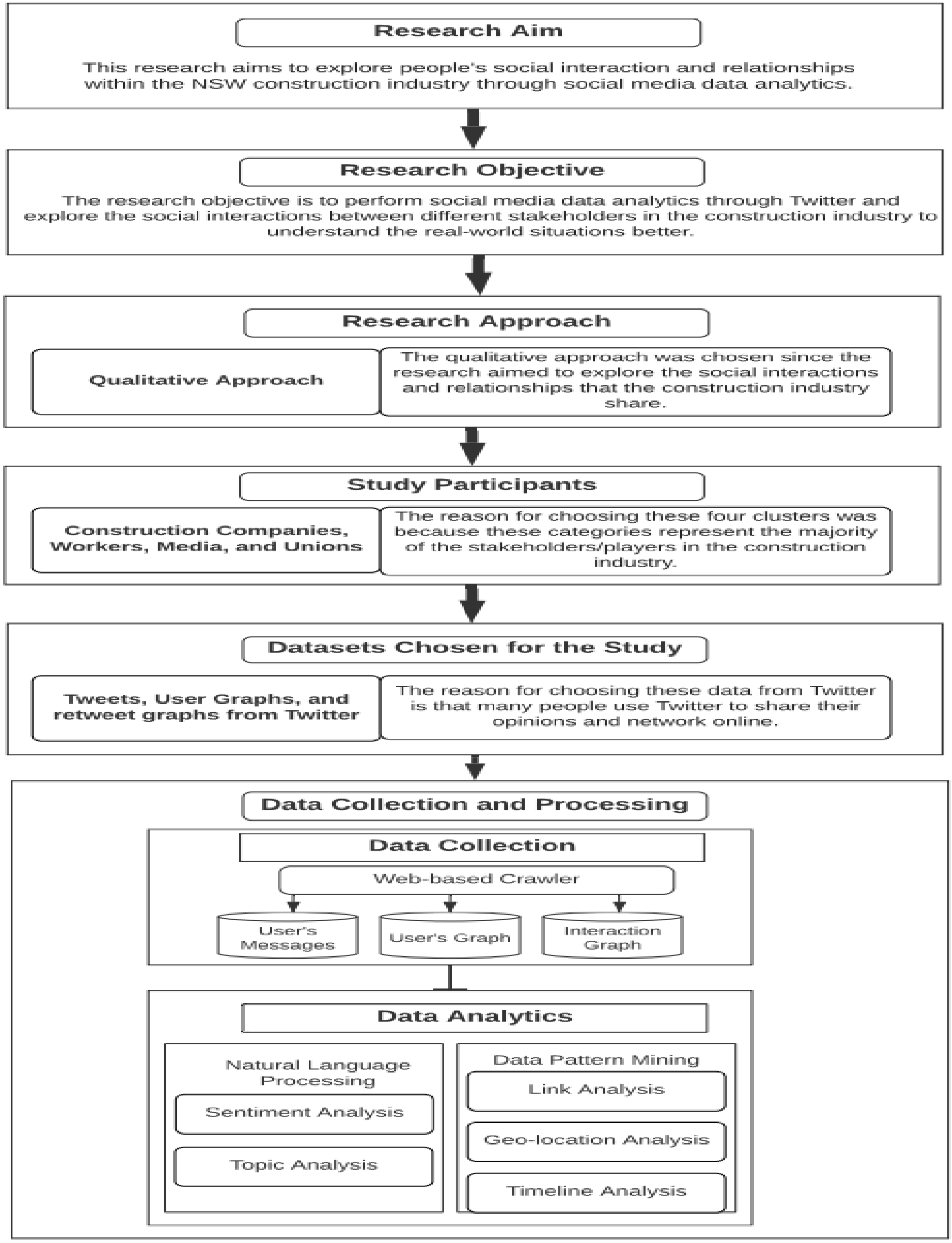
3.5.3 Web-based Data Crawler

In addition to the Twitter API, there was a web-based data crawler that was also used to collect additional data from the Twitter platform. The reason why the web-based data crawler was added to the mix was because the Twitter API was somewhat limited to the amount of data that it could extract from Twitter. As Tang et al. (2017) state, the Twitter API is limited by the

OAuth authentication protocol to obtain only up to 1% of the public tweets from Twitter, within a given time. This means that the API could only retrieve a limited number of tweets, which, in turn, also limits the scope of data to work with in this study. Therefore, to avoid this limitation, the researchers included a second data collection tool, precisely the web-based data crawler. Unlike the API, the crawler was not limited to Twitter's policies, and could, therefore, parse Twitters' html web pages and extract more data. Thus, while the API could only retrieve 401 tweets from the platform, the web-based crawler was able to retrieve many more tweets.

3.5.4 Python Jupyter Notebook

Python is a general-based programming language that is available in numerous versions, one of them being within the Jupyter notebook platform. The Jupyter notebook is a web-based interactive development environment that is used to write, edit, and share code in different programming languages, e.g., Python. Jupyter notebook is freely available and accessible from the Anaconda distribution edition. Therefore, the Jupyter notebook was used in this study to develop and run the code for the Tweepy package. Figure 2 below gives a concise outline of the research flowchart for this chapter.



This research aims to explore people's social interaction and relationships within the NSW construction industry through social media data analytics.

The research objective is to perform social media data analytics through Twitter and explore the social interactions between different stakeholders in the construction industry to understand the real-world situations better.

The qualitative approach was chosen since the research aimed to explore the social interactions and relationships that the construction industry share.

The reason for choosing these four clusters was because these categories represent the majority of the stakeholders/players in the construction industry.

Datasets Chosen for the Study
Tweets, User Graphs, and retweet graphs from Twitter

The reason for choosing these data from Twitter is that many people use Twitter to share their opinions and network online.

Figure 2: Research Methodology Flowchart.

Chapter 4. Data Collection and Analysis

4.1 Data Collection

This section seeks to elaborate the step-by-step procedure that was used in collecting and analyzing the primary, raw data set of tweets, user graphs, and retweet graphs. The discussion of this section will address the extraction of the data set using the tools and instrumentation as well as the participants involved. The data collection process is outlined below in 5 steps:

4.1.1 Installing Tweepy Package

The first step was to install the Tweepy package in Python Jupyter notebook environment, using the code syntax shown in Table 1 below.

Table 1: Installing Tweepy Package in Python

```
$ mkdir tweepy-bots

$ cd tweepy-bots

$ python3 -m venv venv

$ source ./venv/bin/activate

$ pip install tweepy
```

4.1.2 Creating a Python Dependency

The second step was to create a requirements.txt file that would hold the names of all the dependencies that would be required to ensure that the new Twitter project would proceed smoothly, without encountering any hiccups or drawdowns whatsoever. This was achieved using the code syntax shown in Table 2 below:

Table 2: Creating a Python Dependency

```
$ pip freeze > requirements.txt
```

4.1.3 Configuring the Authentication Credentials for the Twitter API

The third step was to configure the authentication credentials that would be required by the Twitter API to make an application to the OAuth protocol, so as to have the invoke request from Tweepy approved. Usually, the basic authentication credentials that are required by the protocol, include: the consumer's key, the consumer's secret, the access token, and the access secret. In order to get these details, one needs to first to apply for a Twitter developer account, then, later, create an application, as Twitter only grants the authentication credentials to a Twitter app and never a personal account (Garcia, 2020). Once the Twitter application is approved, then one can go ahead and create the authentication details within the app.

4.1.4 Accessing the Twitter API from Tweepy

The next step was developing a code syntax that would enable Tweepy to invoke the Twitter API and make a request to it to retrieve the data. This was made easy by invoking Tweepy's OAuthHandler class that would incorporate all the authentication credentials that were created in the previous step (see Table 3 below).

Table 3: Invoking the Twitter API Using Tweepy

```
import tweepy


# Authenticate to Twitter

auth = tweepy.OAuthHandler("CONSUMER_KEY", "CONSUMER_SECRET")

auth.set_access_token("ACCESS_TOKEN", "ACCESS_TOKEN_SECRET")


# Create API object

api = tweepy.API(auth, wait_on_rate_limit=True,
    wait_on_rate_limit_notify=True)
```

4.1.5 Extraction of Data from Twitter

After the above steps, the final step was now to extract the actual data. See Table 4 below for a demonstration of the code snippet that was used to achieve this step.

Table 4: Data Extraction from Twitter and Search for Tweets

```
# Define the search term and the dates as variables
search_words = "construction", “project”, "@construction", “@project”, "#construction",
“#project”
date_since = "2008-05-16"
date_to = "2016-03-10"


# Collect tweets
tweets = tw.Cursor(api.search,
              q=search_words,
              lang="en",
              since=date_since,
              to=date_to).items(401)
# Iterate and print tweets
for tweet in tweets:
    print(tweet.text)


# Extract the types of users posting the above keywords and their locations
```

```
users_locs = [[tweet.user.screen_name, tweet.user.location] for tweet in tweets]

users_locs
```

The code snippet in Table 4 is explained in detail in the subsections below:

# Define the search term and the dates as variables: This section of code was written to search Twitter for tweets, i.e., normal messages, retweets, replies, and hashtags. This process was done using a heuristic-based method that would enable the researchers to obtain the seed users from Twitter. The heuristic rules required that the researchers use two core keywords, i.e., “construction” and “project”, to fish out for all the tweets that contained these terms. As seen from Table 4 above, 6 terms were included in the “search_words” variable in order to specify the different types of tweets that the API should search for. The “construction” and “project” search words would extract any normal tweets that contained these two terms, while the “@construction” and “@project” search words would extract any replies or retweets that contained the @ symbol. The last two “#construction” and “#project” search words would extract all the hashtags that contained these two terms and were also accompanied by the # symbol. In addition to the specification of the search words, the code snippet in Table 4 above also specified the time period within which the tweets were published on Twitter. Therefore, the API only collected the tweets that had been published within the time period between May 16th, 2008, and March 10th, 2016, as specified by the variables, “date_since” and “date_to”.

# Collect tweets: This second section of code was written in order to begin the collection of all the tweets that contain either of the 6 search words and which were published within the specified time period.

# Iterate and print tweets: This third section of code was written in order to request the program to continuously loop through the second section of code, collecting the appropriate tweets and printing the results.

# Extract the types of users posting the above keywords and their locations: This last section of code was written to have the program indicate the types of users who were posting these tweets and their geo-locations. The "tweet.user.screen_name" method indicates the user's twitter handle, while the "tweet.user.location" indicates their geo-locations.

Table 5 below lists the number of collected tweets per each participant cluster.

Table 5: Number of Collected Tweets per Each Participant Cluster

Cluster	**Total Tweets Collected**
Companies	401
Workers	401
Media	401
Unions	401
TOTAL	**1604 Tweets**

4.2 Data Analysis Processes

Once all the data was collected, the next step was data analysis. As earlier stated, the data analysis was divided into five phases, i.e., sentiment analysis, topic modeling, link analysis, geo-location analysis, and timeline analysis. Both the sentiment analysis and topic modelling are generally grouped under natural language processing (NLP), while the remaining three techniques, i.e., link analysis, geo-location analysis, and timeline analysis, are classified under data pattern mining. As Tang et al. (2017) state, before either of these analyses is done, the collected data has to, first, undergo a pre-processing stage.

4.2.1 Data Pre-Processing Stage

The data pre-processing stage is the phase whereby the raw data is cleaned in order to make it ready for machine learning processes, e.g., natural language processing and also data pattern mining (Canchen, 2019). Initially, before any pre-processing takes place, the data is often polluted and unsuitable for machine learning models. This means that the data is noisy as it contains a lot of filler words or phrases that are meaningless and only make machine learning processes take longer to complete (Ramirez-Gallego et al., 2016). For example, in the sentence, “Oh gosh, I have been feeling so, so, so, very tired and I would like to just, uh, relax at home the entire day today.” There are a number of filler words in this sentence, which only make the sentence very long and tiring to read. The sentence would read better and easier if it was just phrased like this: “I have been feeling very tired and I would like to relax at home today.” The second sentence is shorter than the first as it has got rid of several punctuation marks, conjunctions, and hesitation sounds, such as “uh”, “oh”, etc. According to Luengo, Garc´ıa, and Herrera (2015), filler words tend to lower the quality of a speech as they are meaningless. Therefore, data has to first under pre-processing in order to clean it and remove noises, such as conjunctions, verbatim sounds, punctuation marks, etc., so as to increase the accuracy of machine learning performance and also decrease the time taken (Canchen, 2020). In this study, several pre-processing steps were performed on the data collected, including tokenization and data clean-up (e.g., removing stop words, punctuation marks, numbers, white spaces, upper case formatting, etc.). See Table 6 below for a demonstration of the code snippet that was used for the pre-processing phase.

Table 6: Code Snippet for the Data Pre-Processing Step

```
#Step 1: Installing the necessary libraries into the R workspace
install.packages('tm') #Install package for natural language processing
library(tm)
install.packages("readr") #Install package for reading external files into R
library(readr)
install.packages('stringr') #Install package for tokenization of tweets into words
library(stringr)
install.packages("wordcloud")
library(wordcloud)
install.packages("topicmodels") #Installing package for topic modeling
library(topicmodels)
install.packages(reshape2) #installing package for reshaping dtm
library(reshape2)
```
~~~
#Step 2: Load dataset into the workspace
```{r}
data<-read_csv("/cloud/project/NSW/Company.csv")
~~~
~~~
#Step 3: Select the column with the Twitter tweets/ text
```{r}
~~~

```
company_tweets<-data$text
```
#Step 4: Clean and preprocess the company_tweets by doing the following:
```{r}
data1<-gsub(pattern="\\W",replace=" ",company_tweets) #Remove punctuation from file
data2<-gsub(pattern="\\d", replace=" ", data1) #Remove numbers from the file
data3<-removeWords(data2, stopwords()) #Remove stop words from the file
data4<-gsub(pattern="\\b[A-z]\\b{1}", replace=" ", data3)
data5<-stripWhitespace(data4) #Strip white spaces
data6<-tolower(data5) #Transform the text from upper case to lower case
```
#Step 5: Perform Tokenization and Unlist Object into a Character
```{r}
tokenized_words<-str_split(data6, pattern="\\s+")
tokenized_words<-unlist(tokenized_words)
tokenized_words
```

#Step 1: Installing the necessary libraries into the R workspace: The first section of the code was to first install all the necessary packages and libraries in the R programming workspace. The packages that were considered for this exercise were tm, readr, stringr, wordcloud, topicmodels, and reshape2. Each of these packages had a specific function, as defined below:

- Tm – Contains classes and methods for natural language processing (for sentiment analysis and topic modelling)

- Readr - Contains classes and methods for reading any external files into R language workspace. For this study, the data that was collected from Twitter were compiled together in MS Excel csv files.
- Stringr- Contains classes and methods for tokenization of the tweets into single, separate words
- Topicmodels - Contains classes and methods for analyzing topics (topic modelling)
- Reshape2 - Contains classes and methods for reshaping document term matrices (dtm)

#Step 2: Load dataset into the workspace: The second section of the code was to enable the program to load, or read, the external csv files into the current working directory.

#Step 3: Select the column with the Twitter tweets/ text: The third section of the code was to enable the program to pick out only one column, from the entire csv file, which contained the tweets from Twitter.

#Step 4: Clean and preprocess the company tweets by the following steps: The fourth section of the code was to perform the data clean-up, so as to remove the noises in the tweets. Various types of noises were removed, including punctuation marks, numbers, stop words, and white spaces. Also, the last line of code in this section was run to change all the upper-case formatting into lower case, since the R program treats upper case letters and lower case letters as two different entities, even if the words or phrases are grammatically the same.

#Step 5: Perform Tokenization and Unlist Object into a Character: The fifth section of the code was written to perform tokenization of the tweets into words. Luengo et al. (2015) define tokenization as the act of splitting down whole sentences, paragraphs, or entities into smaller units, which can either be in form of words, phrases, etc. Tokenization is often considered as a basic step in pre-processing since it returns smaller units, which the other machine learning processes, e.g., natural language processing, can easily and quickly analyze.

4.2.2 Sentiment Analysis

After the pre-processing phase was over, the next phase of data analysis was the sentiment analysis. As earlier stated, sentiment analysis was performed to primarily determine the polarity of the opinions, emotions, or sentiments within a given dataset. The purpose of the sentiment analysis was to see if the users' tweets were positive, negative, or simply neutral; and, if yes, what was the extent of these polarities. As previously highlighted, this study undertook the corpus-based approach when performing the sentiment analysis. The reason for adopting this approach was because the corpus-based approach seemed a more effective technique than the dictionary-based approach. While the dictionary-based approach is comprehensive and contains numerous opinion words, Khoo and Johnkhan (2017) state that it is rather limited to detect opinion within single words, and not within an entire sentence. This means that this approach is incapable of detecting the opinion or sentiment of a sentence and, therefore, cannot decipher the contextual meaning of a document.

The corpus-based approach, on the contrary, is not limited to single words (Medhat et al., 2-14). Rather, the approach is designed to detect an opinion, sentiment, or emotion within whole phrases, sentences, and documents. Therefore, the approach is capable of deciphering sentiments within context-specific orientations (Sadia et al., 2018). According to Aung and Myo (2017), the corpus-based approach uses syntactic patterns to determine sentiment in phrases that frequently occur with a seed list of opinion terms. Therefore, the corpus-based approach can be considered an upgrade of the dictionary-based approach, as the former technique adds peripheral terms to the basic seed lists that are developed by the dictionary-based approach. Therefore, this thesis seeks to use the corpus-based approach to analyze the sentiment scores of the Twitter tweets of the four clusters in the construction industry. In order to achieve this end, the following code snippet was executed (see Table 7 below).

Table 7: Sentiment Analysis in R

~~~
#-------------------------SENTIMENT ANALYSIS--------------------------------#

#Step 1: Load the positive lexicon into the workspace

```{r}

positive<-read_tsv(file.choose())

```

#Step 2: Load the negative lexicon into the workspace

```{r}

negative<-read_tsv(file.choose())

```

#Step 3: Match the tokenized words with positive lexicon

```{r}

positive_results<-sum(!is.na(match(tokenized_words, positive)))

positive_results

```

#Step 4: Match the tokenized words with negative lexicon

```{r}

negative_results<-sum(!is.na(match(tokenized_words, negative)))

negative_results

```
~~~

~~~~
#Step 5: Find the sentiment analysis score

```{r}

final_score<-positive_results - negative_results

final_score

```
~~~~

Elaboration of the code

The corpus-based approach that was used in this study encompassed two lexicons, namely a positive and negative lexicon. The positive lexicon contained purely positive terms, e.g., beautiful, pretty, pleasant, welcoming, refreshing, comforting, wonderful, etc., while the negative lexicon contained only negative terms, e.g., bad, ugly, disrespectful, annoying, uncomfortable, unpleasant, etc. Therefore, first, the two lexicons were loaded into the R workspace, after which, they were compared against the dataset (the tokenized tweets), to assess the number of words or phrases that the tweets and the lexicons had in common. A score was, thereafter, given to indicate the sentiment polarity of the dataset.

#Step 1: Load the positive lexicon into the workspace: This first line of code was to load the positive lexicon into the R workspace.

#Step 2: Load the negative lexicon into the workspace: This second line of code was to load the negative lexicon into the R workspace.

#Step 3: Match the tokenized words with positive lexicon: The third line of code was compared to the tweets with the positive lexicon to see how many words they had in common

#Step 4: Match the tokenized words with negative lexicon: Just like the third line of code, the fourth line of code was to compare the tweets with the negative lexicon in order to see how many words they had in common

#Step 5: Find the sentiment analysis score: The final line of code was to calculate the sentiment analysis score. If the tweets were found to have a higher number of negative terms than positive terms, then they were, generally, considered as negative, and vice versa. If negative, the sentiment score will be given as -1, while if it is positive, the sentiment score will be given as 1 (Jalayer Academy, 2017). However, if it is neutral (i.e., does not contain either positive or negative terms), then the sentiment score will be given as 0.

4.2.3 Topic Modelling

The next step was the topic modelling analysis, whose purpose was to identify or pick out the unique topics that are present within the Twitter tweets posted online by the construction workers, companies, unions, and the media. To achieve this end, the following code snippet was executed in R (see Table 8 below).

Table 8: Topic Analysis Source Code

~~~
#------------------------------TOPIC ANALYSIS--------------------------------#

#Step 1: Convert text into corpus and document term matrix

```{r}

corp<-Corpus(VectorSource(tokenized_words)) #Conversion tokenized_words into corpus

dtm<-DocumentTermMatrix(corp1)

dtm1<-as.matrix(dtm)

```

#Step 2: Get frequency of the top 25 keywords
~~~

~~~
```{r}

freq<-colSums(dtm1)

order<-order(freq, decreasing=TRUE)

freq1<-freq[head(order)]

freq1

```
~~~

Elaboration of the code

#Step 1: Convert text into corpus and document term matrix: This first section of the code was to transform the tweet tokens into a document term matrix, since the 'topicmodels' package (demonstrated in Table 4 above) only works with document term matrices.

#Step 2: Get frequency of the top 25 keywords: This second section of the code was to enable the program to show the top 25 most frequent topics in the document term matrix.

4.2.4 Link Analysis

The third step was the link analysis, which was performed in order to establish the connections and associations between the user clusters present within the construction industry. The end goal of the link analysis was to learn whether there are any forms of associations or connections between construction workers, companies, unions, and the media. To achieve this end, the following code snippet was used to carry out the link analysis exercise (See Table 9 below).

Table 9: Link Analysis in R

```
# Step 1: Install packages

install.packages(c("igraph", "tidyr", "tidygraph", "RColorBrewer"))
```

```
#Step 2: Load the network dataset

library (readr)

csvfile<-read_csv("/cloud/project/R/LinkAnalysis.csv")


#Step 2: Manage dataset

B<-as.data.frame(table(csvfile)) # Create an edge weight column named "Freq"

B1<-subset(B,Freq>0) # Delete all the edges having weight equal to 0


#Step 3. Create an igraph object from the dataframes

library(igraph)

ljnks<-graph_from_data_frame(d=B1, directed = TRUE)


#Step 4: Measuring degree centrality

links_deg<-degree(links,mode=c("out"))

V(links)$degree<-links_deg

V(links)$degree

which.max(links_deg)
```

Elaboration of the code

# Step 1: Install packages: This first section of the code was to install all the packages necessary for performing the link analysis. Each package had a specific purpose, as elaborated further below:

- Igraph – Contains classes and methods for creating an igraph object (similar to that in Figure 2 above) that shows the links, or arrows, between various nodes in a given network.
- Tidyr & tidygraph – Contains the necessary tools for making a given dataset tidy and manipulating the shapes of node and edge dataframes
- RColorBrewer – Contains color palletes that may be used to give igraph objects different color shades

#Step 2: Load the network dataset: This second section of the code was to load the network dataset into the R workspace.

#Step 2: Manipulate edges data: This third section of the code was to create an edge column with real integers, greater than zero. Edge data is important as it is used, alongside nodes data, to create an igraph object. The second line of code [B1<-subset (B,Freq>0) # Delete all the edges having weight equal to 0] was also used to delete any edges equal to 0.

#Step 3. Create an igraph object from the dataframes: This fourth section of code was to create an igraph object, showing the links between the user clusters in the construction industry.

#Step 4: Measuring degree centrality: This fifth section of code was run so as to calculate the number of links (or arrows) incident upon a node. The higher the number of links incident upon a node(s), the higher the degree centrality of that node. For example, if the media had 100 number of links while the workers had 80 number of links directed towards them, then the media node would be considered to have a higher degree centrality.

4.2.5 Geo-location Analysis

A geo-location analysis was performed next, by simply reading the geo-location information shared by the users below their tweet posts. As previously mentioned, most, if not all, users' geo-location data can be shared through social media networks. However, users have to activate these functions by allowing these apps to access their GPS location features. As Tang et al. (2017) state, not all users activate these features, which means that there are some tweets that indicate the users' locations and then there are others that do not. This thesis only collected the tweets that were indicative of the users' geo-locations, by executing the code snippet shown in Table 10 below.

Table 10: Geo-location Code Snippet

```
# Extract the types of users posting the above keywords and their locations

users_locs = [[tweet.user.screen_name, tweet.user.location] for tweet in tweets]

users_locs
```

The geo-location data will help in identifying the locations of the four user clusters in the construction industry and understanding their sentiments. For example, we will be able to reveal which states have users with the most negative, neutral, and positive sentiments. This way, we shall know the extent of construction-related issues by location, thus aiding in the implementation of feasible solutions by location as well.

4.2.6 Timeline Analysis

The timeline analysis was also performed in a similar manner like the geo-location analysis. The following code snippet, in Table 11 below, was used for to achieve the timeline analysis exercise.

Table 11: Timeline Analysis Code Snippet

```
# Extract the types of users posting the above keywords and their locations

users_time = [[tweet.user.screen_name, tweet.user.timeline] for tweet in tweets]
```

```
users_time
```

The purpose of conducting the timeline analysis was to find out the level of social media engagement that the four user clusters have. The results, hence, helped to answer several questions, such as the time of the day, and the day of the week, when the users tend to post most frequently on Twitter. Knowing this type of information may be of great use as it will enable the appropriate parties to look out for those periods when these users are actively engaged on Twitter and respond appropriately to their posts, tweets, retweets, and replies. This will increase engagement and social interactions between the four clusters of construction stakeholders online and ease the flow of communication amongst everyone. A timeline analysis will also help evaluate the number of times, or frequency, each cluster tends to post content on Twitter. If it is observed that the clusters post too frequently and are persistent about a particular topic, or have a constant negative sentiment, then this means that that particular user has a very pressing issue that needs to be addressed immediately. At other times, it could mean that that particular user may be experiencing a mental health problem that requires immediate treatment or therapy.

Chapter 5. Findings and Results Chapter

5.1 Introduction to the Chapter

This is the fourth chapter of the thesis, and it seeks to present all the findings and results that were obtained, following the five data analysis processes that were conducted in the previous chapter. The findings will be presented in different formats, some in form of graphs, others in histograms, and the rest in tables.

5.2 Sentiment Analysis Results

This section presents the results obtained from the sentiment analysis process, in form of four figures (see Figures 3-6 below).

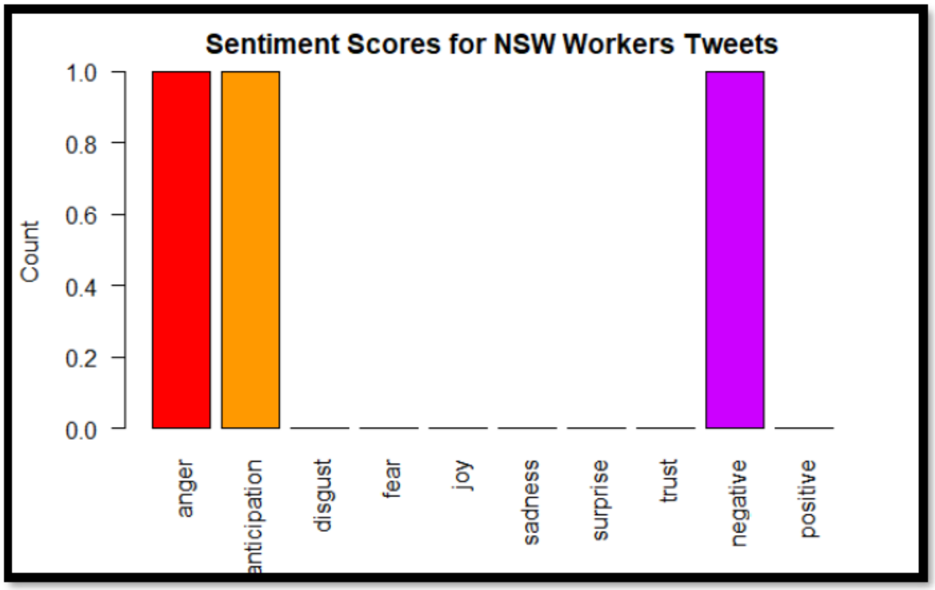


Figure 3: Sentiment Scores for NSW Workers Tweets.

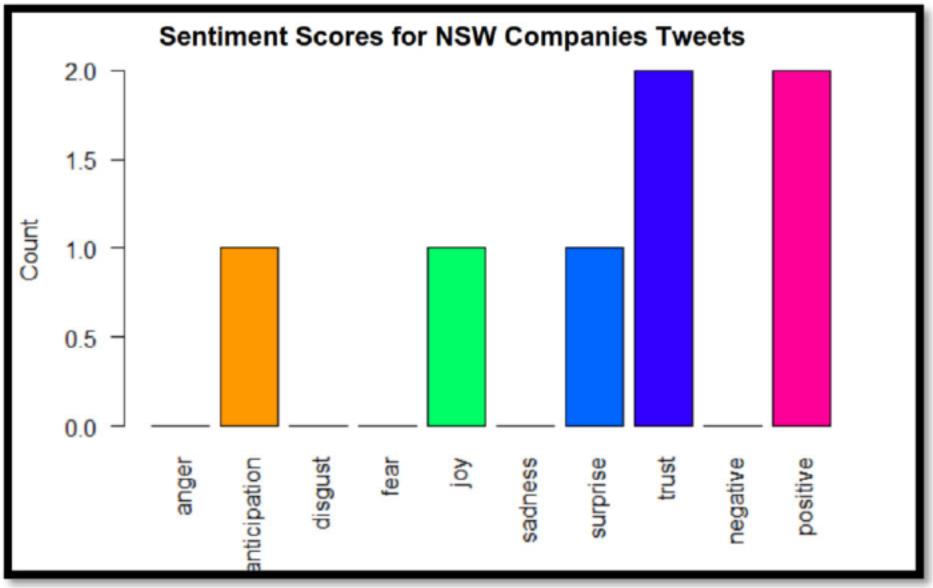


Figure 4: Sentiment Scores for NSW Companies Tweets

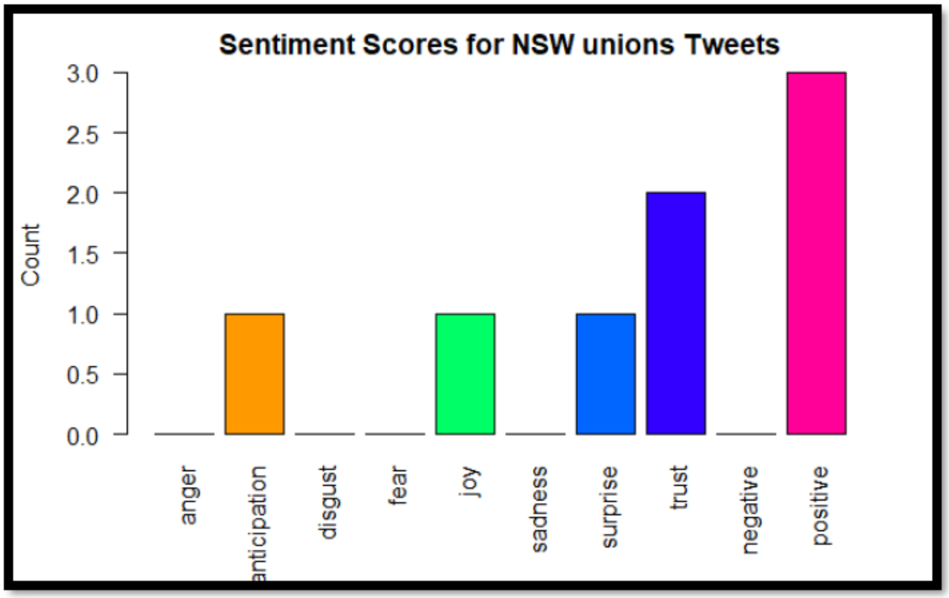


Figure 5: Sentiment Scores for NSW Unions Tweets

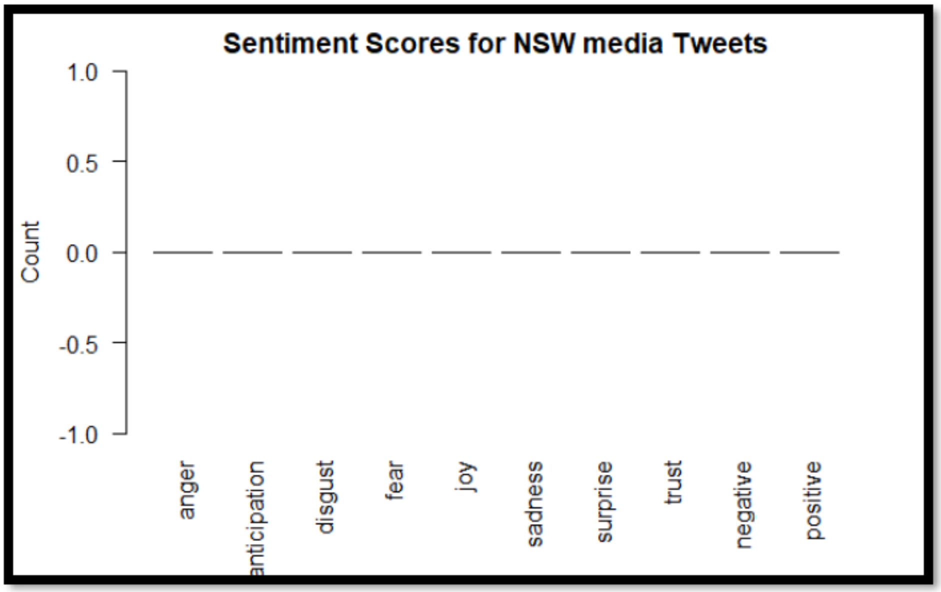


Figure 6: Sentiment Scores for NSW Media Tweets.

Figures 3-6 above demonstrate the sentiment scores of the workers, companies, unions, and media, in that order. As can be seen, the construction workers seem to have the highest percentage of negative sentiments, defined mostly by anger and anticipation. The reason why workers' tweets may have more negative sentiments compared to the other three user clusters may be because workers tend to post more personal content regarding their experiences at their workplaces (Tang et al., 2017). This personal information may be disparaging as it may contain derogatory statements about employers and company managerial teams. For example, workers may be taking out their grievances against their employers and complaining about how they treat them, poor salaries, toxic working conditions, lack of job security, minimal work benefits or perks, or even reporting fraudulent activities within the companies (Adzovie et al., 2017).

As for positive sentiments, the companies have the highest number of positive sentiments than all the other user clusters. The reason for this is because companies, unions, and the media, on the other hand, are less likely to post personal information on social media platforms. Rather, they tend to post business-related matters, which are, for the most part, neutral. Companies, specifically, will most likely use social media platforms to discuss new product launches and projects, digital marketing content and advertising, work policies, etc. (Siricharoen, 2012). This is done in a bid to market their products and services, build their brand image, brand awareness, and corporate identity. Therefore, Twitter, for the, is seen as a venue to further their advertising efforts, conduct market research, as well as maintain healthy relations with their clients and other stakeholders (e.g., investors, suppliers, and the local or international community). Unions, as well, will use social media platforms as a platform to make known to their members about their work policies, labor rights, employer-employee relations, new infrastructure, and any other general issues concerning the workers (Carneiro & Costa, 2020). It is also clear that the media has the highest percentage of neutral sentiments, compared to all the other user clusters in the construction industry; reason being that the media post neutral content on their pages, e.g., news about companies' development, workers' safety, press releases about new products or upcoming projects, financial news, etc. (Hanley, 2014).

5.3 Topic Modelling Analysis Results

This section presents the results obtained from the topic modelling analysis process, in form of bar graphs. The bar graphs illustrate the top 25 keyword topics that most frequently appeared within the companies, workers, unions, and media tweets (See Figures 7-10 below).

5.3.1 Workers

See Figure 7 below for the top 25 keywords and topics extracted from the workers' tweets.

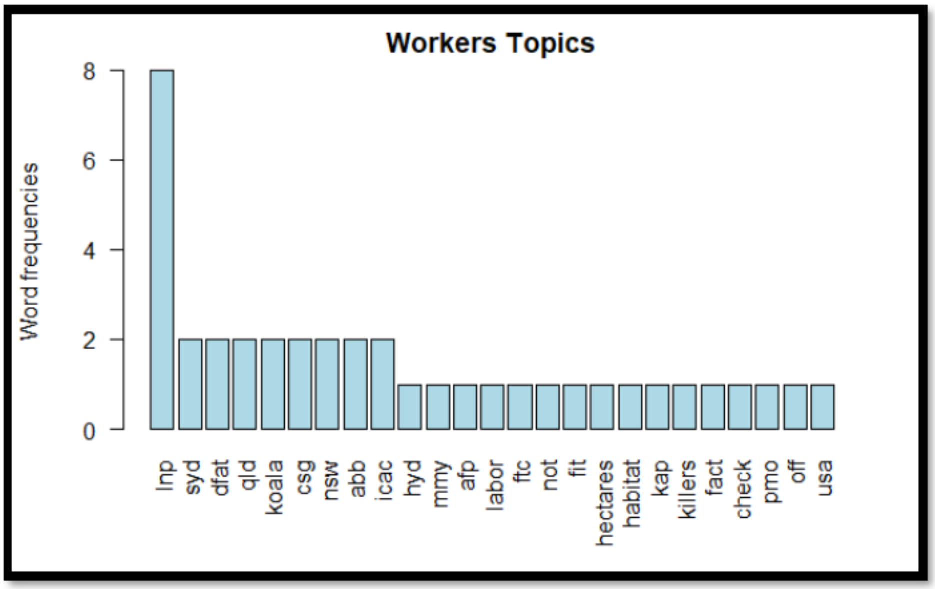


Figure 7: The Top 25 Most Frequent Topics Extracted from Workers.

Figure 7 above illustrates the top 25 most frequent topics that were extracted from the workers' tweets. The keyword topic that leads the list is lnp, followed by syd, dfat, qld, koala, csg, nsw, abb, icac, hyd, mmy, afp, labot, ftc, and hectares, just to mention a few. All these keywords are abbreviations of different construction companies and contractors around the globe, but mostly within Australia. The workers' conversations on Twitter can be said to be mostly focused on construction companies, whereby they compare different organizational cultures across different corporations. For instance, they widely discuss various companies, such as the LNP Construction, DFAT, Koala Construction Limited, CSG Built Pty Ltd., and ABB Construction Ltd, to compare their unique organizational cultures, worker welfares, treatments, and benefits, in order to share notes amongst each other. The end goal is to see which organization has the most favorable worker treatment and benefits within and outside the News South Wales (NSW) region. Some of the worker welfare aspects that the workers often discuss include labor, habitats, killer environments, etc.

5.3.2 Companies

See Figure 8 below for the top 25 keywords and topics extracted from the companies' tweets.

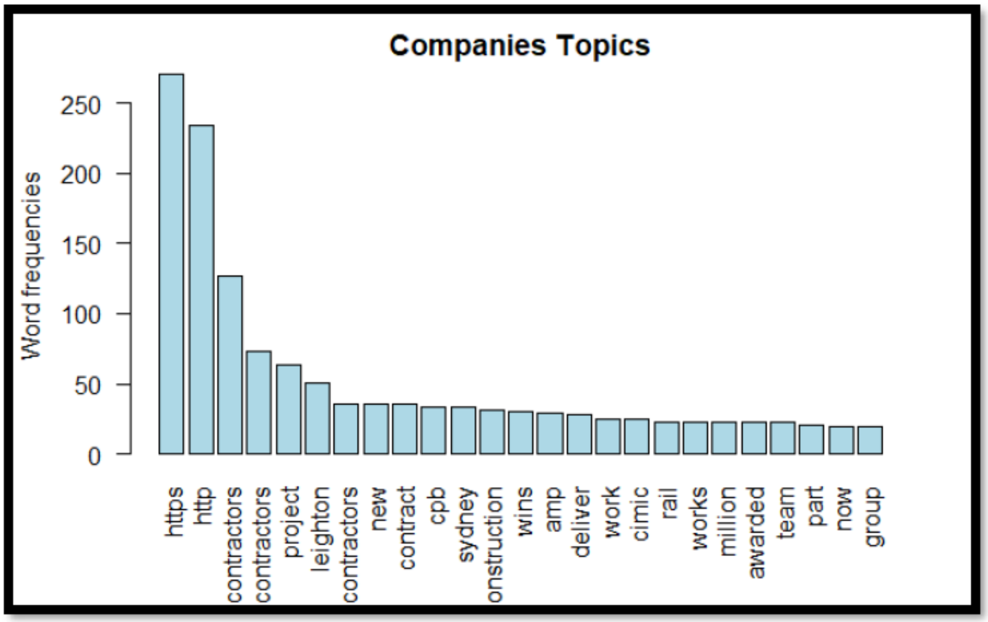


Figure 8: The Top 25 Most Frequent Topics Extracted from Companies.

Figure 8 above shows the top 25 most frequent topics that were extracted from the companies' tweets. They include https, http, contractors, project, Leighton, new, cpb, Sydney, construction, wins, and amp, just to mention a few. As can be seen, the construction companies tend to focus more on business-related matters, as earlier highlighted. Their posts are mostly about advertisements for their products and services as there is a lot of mention of 'https' and 'http' links. The https and http links are redirects to the companies' official websites, where they list their organizational products and services, company agenda, about information, company events, etc. They also frequently talk about their contractors, probably as an effort to market their services and pitch for new tenders from potential clients. The topic of 'cpb' is also frequently mentioned. CPB Contractors is a well-known construction company of the CIMIC Group, based in Sydney, Australia. The firm was formerly known as Leighton Contractors and Theiss and now focuses in areas such as design and construct, construct only, alliances and joint ventures, as well as construction management, just to mention a few.

5.3.3 Unions

See Figure 9 below for the top 25 keywords and topics extracted from the unions' tweets.

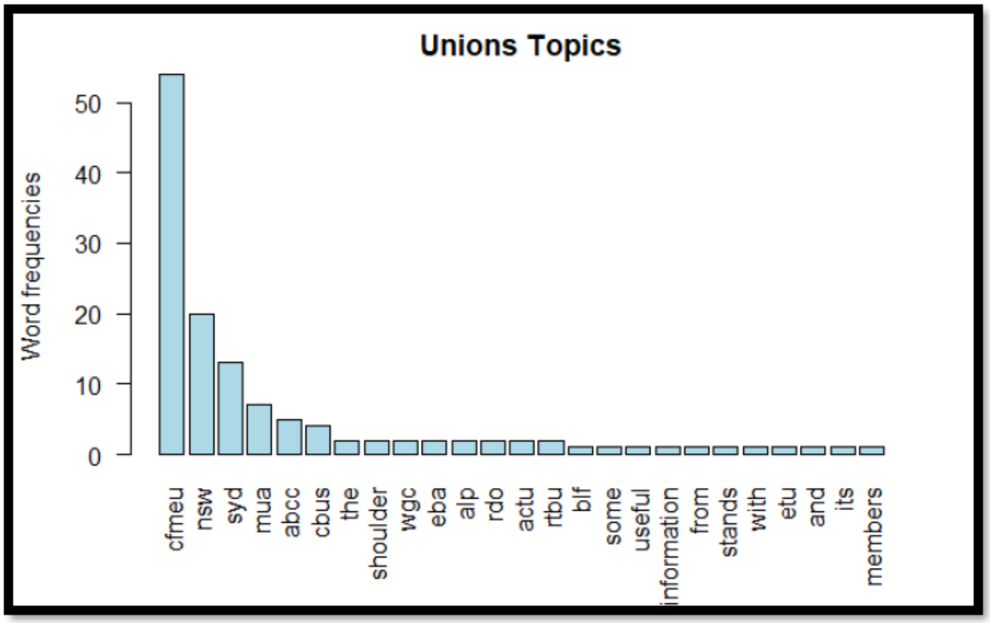


Figure 9: The Top 25 Most Frequent Topics Extracted from the Unions

Figure 9 above illustrates the top 25 most frequent topic keywords that were extracted from the unions' tweets. The most common topics were cfmeu, nsw, syd, mua, abcc, cbus, wgc, eba, alp, rdo, actu, rbtu, blf, and etu, amongst many others. All these keywords are simply abbreviations of the construction unions based in Australia. Therefore, the graph above indicates that majority of the tweets from the unions cluster are often regarding matters to do with construction union organizations. The most talked about union organization that is talked about is the CFMEU (Construction, Forestry, Maritime, Mining and Energy Union), MUA (Maritime Union of Australia), ABCC (Australian Building and Construction Commission), CBUS, ALP (Australian Labor Party), the RBTU (Rail, Tram, and Bus Union), EBU (Electrical Trades Union), and the BLF (Builders Labourers Federation), amongst others. The discussions surrounding the organizations mostly consist of information about the member benefits offered therein, and other issues such as worker advocacy issues, such as employment policies, fair employment for employees, protest news and events (e.g., city council involvement, release, August protest events, etc.), working loads, reforms, etc.

5.3.4 The Media

See Figure 10 below for the top 25 keywords and topics extracted from the media tweets.

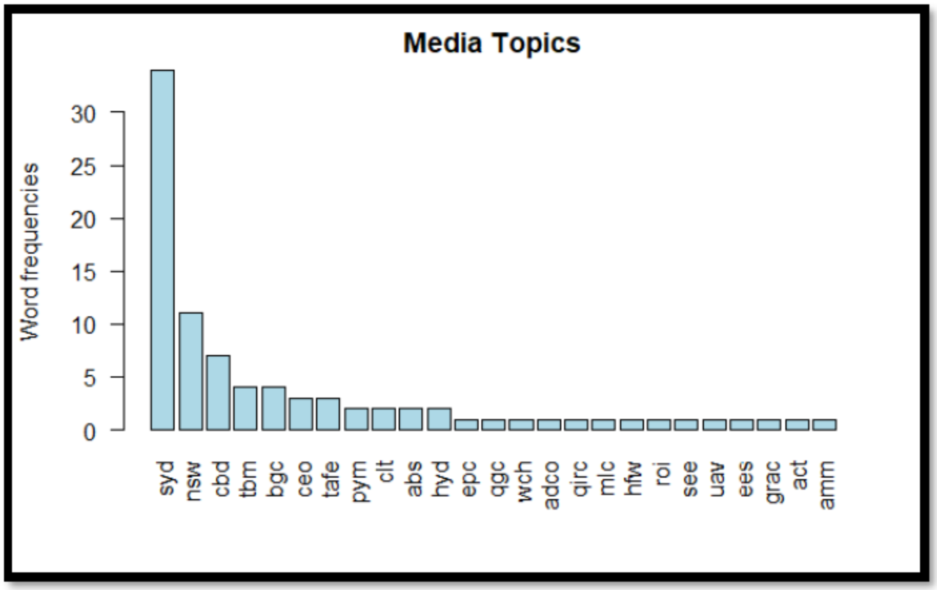


Figure 10: The Top 25 Most Frequent Topics Extracted from the Media.

Figure 10 above shows the top 25 most frequent topic keywords that were extracted from the media tweets. As can be seen, the most common topics are syd, nsw, cbd, tbm, bgc, ceo, tafe, pym, and clt, just to mention a few. It is evident that the construction media cluster have a wide range of topics and are not focused on just one item, unlike the other three clusters, i.e., companies, workers, and unions. From the figure above, it is clear that the media have diversified discussions about various things, such as construction companies (e.g., bgc – BGC Housing Group, ADCO International Pty Ltd.), Australian cities (e.g., syd – Sydney, nsw – New

South Wales), construction company roles (e.g., ceo – Chief Excecutive Officer), and even media groups (e.g., epc – EPC Media Group, mlc – MLC Media Centre).

5.4 Link Analysis Results

The results of the link analysis method are presented in this section. These results are shown in three tables, namely the follower-following ratio and types of tweets amongst the four clusters (see Tables 12 and 13 below).

5.4.1 Follower-Following Ratio

The results of the FF ratio are arranged from the highest to the lowest in Table 12 below. As can be seen, the companies have the highest number of followers and following, which translates to having the highest FF ratio than all the other clusters. The likely reason for this is because the companies always strive to stay professional throughout their social media content and engagement, posting information about the organizations and corporate events. The cluster with the second highest FF ratio is the media, followed by the unions, and lastly by the workers.

Table 12: Follower-Following (FF) Ratio among the Four Clusters

Cluster	**Number of followers**	**Numbers of following**	**FF Ratio**
Companies	380	21	18.09523
Media	269	132	2.037878
Unions	212	189	1.121693
Workers	163	238	0.684874

5.4.2 Types of Tweets

Table 13 below shows the different types of tweets that Twitter contains, e.g., retweets, replies, retweeted, favorites, hashtags, and solo tweets. As can be seen, the media had the highest number of retweets and then followed by workers, thanks to their more personal engagement with each other and closer interactions. 28.7% of the workers are likely to reply to tweets which is statistically significant when compared to media and union with a lot of normal tweets. Though, less tweets from workers were from solo tweets (8.5%) which do not have no interaction with other three clusters. The media posts were likely to be retweeted by users due to educational and news nature of the posts useful to the followers (Tang et al., 2017).

Table 14 shows the number of accounts at each of the cluster. It can be depicted that the construction workers were rarely stated by the other three clusters. The companies are influenced by media and unions, and sometimes the followers of the union and companies may decide not to respond to workers posts. Besides, unions are influenced by all the three clusters, and they occasionally reply or retweet workers' posts. Finally, both companies and unions have an influence on the media, which may be due to the fact that they collect and publish fascinating posts for their followers every day (Tang et al., 2017).

Table 13: Types of Tweets Posted by the Four Clusters

Type of message	**Company**		**Worker**		**Union**		**Media**	
	Total	%	**Total**	%	**Total**	%	**Total**	%
All	401	100	401	100	401	100	401	100
Retweet	45	11.2	48	12.0	14	3.5	60	15.0
Reply	130	32.4	115	28.7	87	21.7	98	24.4
Normal tweets	226	56.4	238	59.4	300	74.8	243	60.6
Retweeted	51	12.7	112	27.9	114	28.4	112	27.9
Favorite	62	15.5	70	17.5	82	20.4	72	18.0
Hashtags	53	13.2	22	5.5	55	13.7	14	3.5
Solo tweets	60	15.0	34	8.5	49	12.2	45	11.2

Table 14: Number of @ at among each other

Data set	Number of @workers	Number of @companies	Number of @unions	Number of @media
Worker	320	0	11	0
Company	0	347	36	21
Union	6	13	147	35
Media	0	54	92	265

5.5 Geo-location Analysis Results

With continued increase in social media analytics, there is need to perform the analysis in a way to consider a geospatial aspect of the tweets or likes or even the sentiments made by users in various locations. Doing this would enable the gain of insights to the links in information from social media analysis and the zones from which the users come from (Ho et al., 2020). To begin with, the analysis of number of favorite and number of retweets is analyzed by location. The aim is to gain insights on the locations which found the tweets more likeable and the locations that chose to retweet the tweets more considered to others. The social media interactions have reach to about 37 different locations according to the dataset.

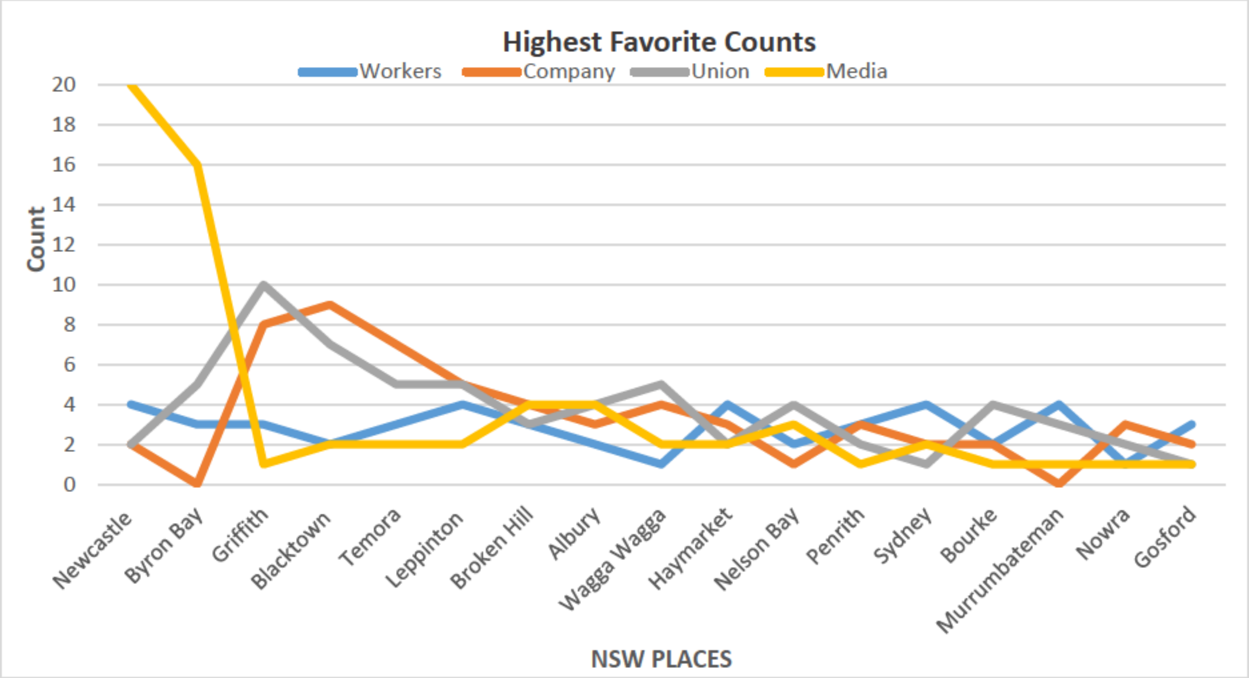


Figure 11: Top Locations with highest number of Favorite Counts

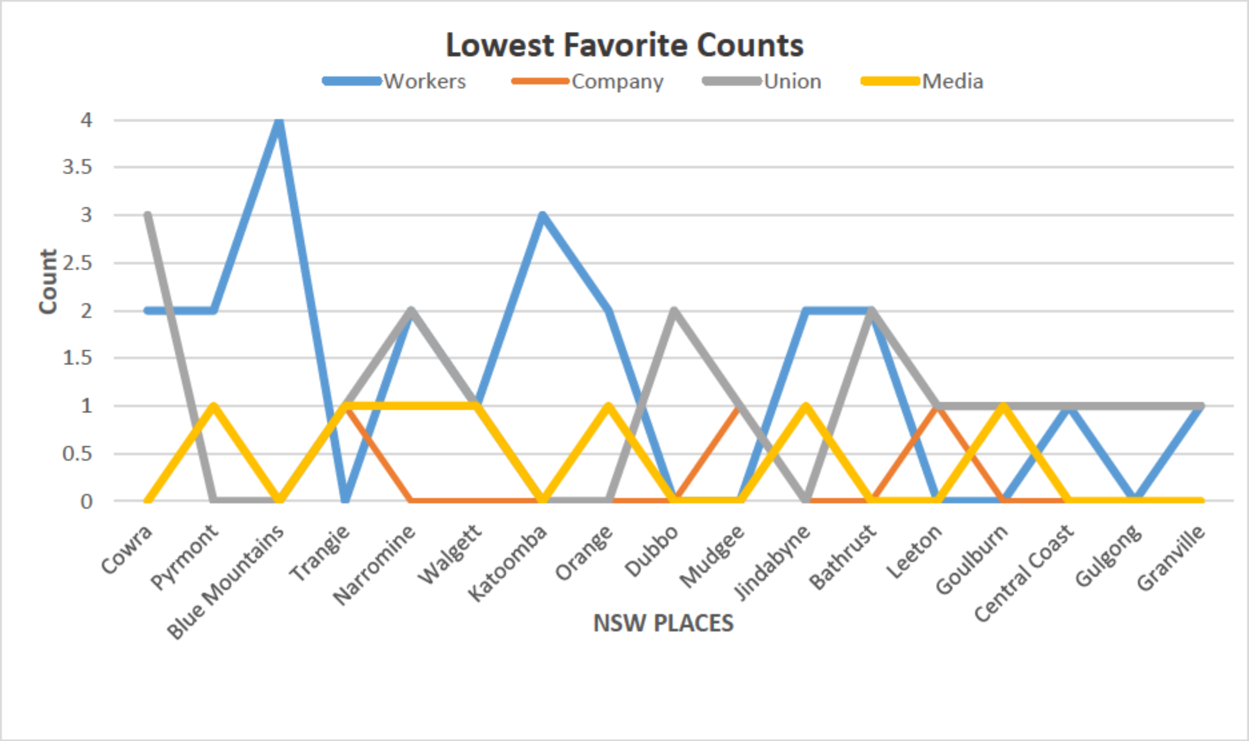


Figure 12: Locations in NSW with low counts of favorites.

Figure 11 and 12 above show the locations with highest number of favorite counts and lowest number of counts respectively. Most of the tweets were like by users from Newcastle as shown by Figure 11. Users from Graville and Gulgong had only one tweet each likable as shown by Figure 12 above.

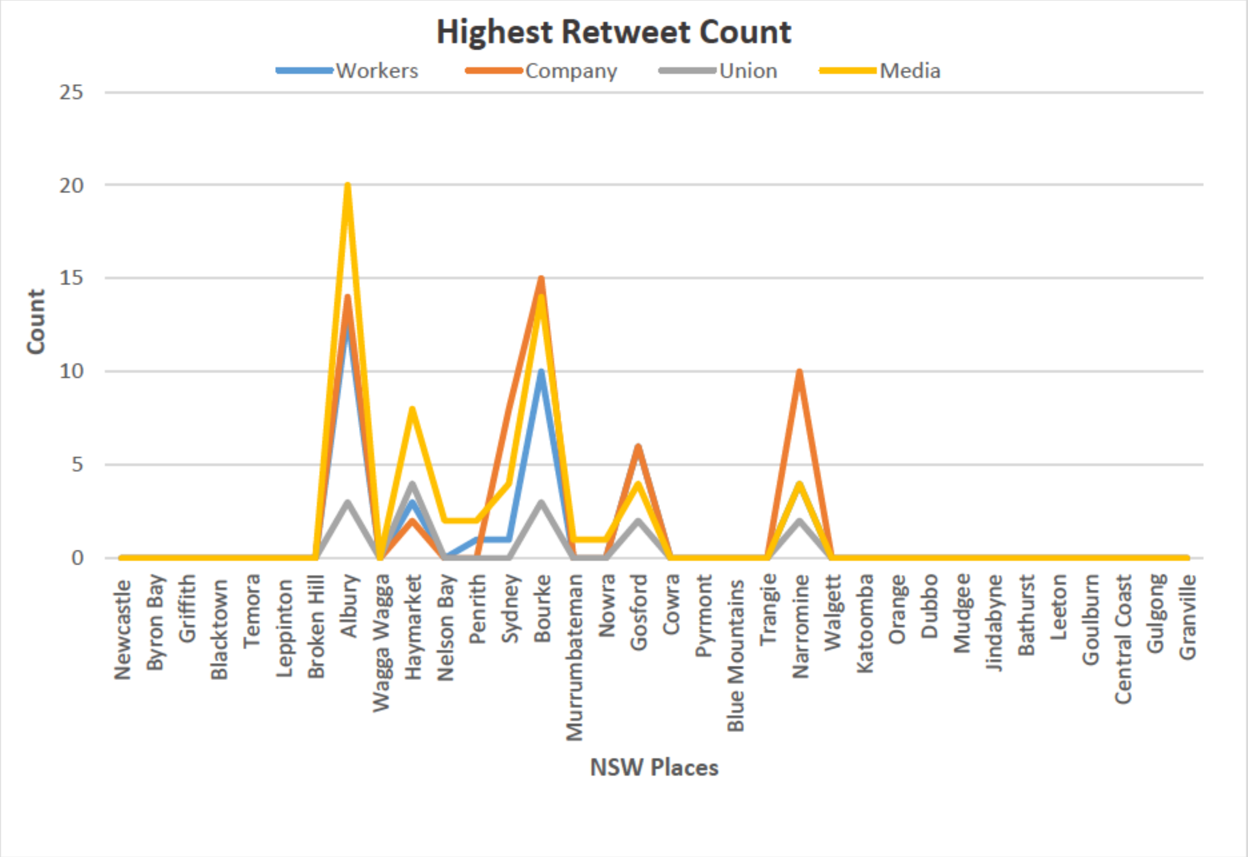


Figure 13: Top locations with highest number of retweets for the tweets

Figure 13 above shows the top 17 locations with the highest and lowest number of retweets from the four clusters. The highest number of retweets was from Albury, followed by Bourke, then Narooma, Gosford, Hay market and Sydney. Other places registered zero retweets such as Newcastle, Byron Bay, Griffith, Blacktown, Temora, Leppington, Broken hill, Wagga

Wagga, Cowra, Pyrmont, Blue Mountains, Trangie, Walgett, Katoomba, Orange, Dubbo, Mudgee, Jindabyne, Bathrust, Leeton, Goulburn, Central Coast, Gulgong and Granville.

5.6 Timeline Analysis Results

Figures 14 and 15 shows the timeline analyses graphs for the four construction clusters i.e., company, worker, union, and media.

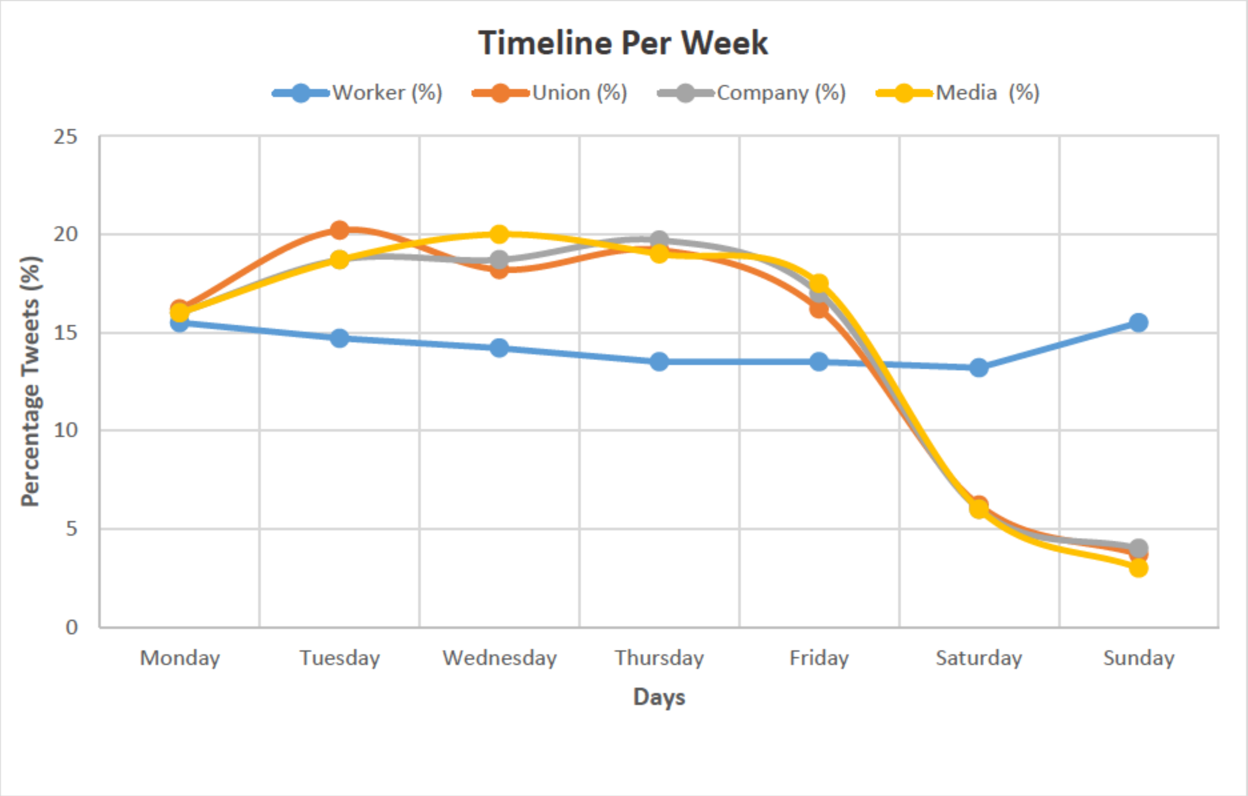


Figure14: Timeline per week for social media users of four clusters

Generally, as can be seen from figure 14, depicts that, workers greatly used social media from Friday to Sunday. As from Monday, their interaction with social media steeply increases and only greatly changed when Friday reached. Companies greatly used social media on Thursday. From Friday onwards it reduced. Unions and Media used social media more during Tuesdays and Wednesday respectively compared to other days. For Saturday and Sunday their interactions with social media were relatively low compared to other days. For the case of workers, they tend to post daily at earlier morning and number of posts reduces during working hours. The media publishes the tweets in working hours and posts in the afternoons and evening hours while the union and companies tweets in the morning hours and posts in the morning due to public relations staff needed to post earlier for creating the attentions.

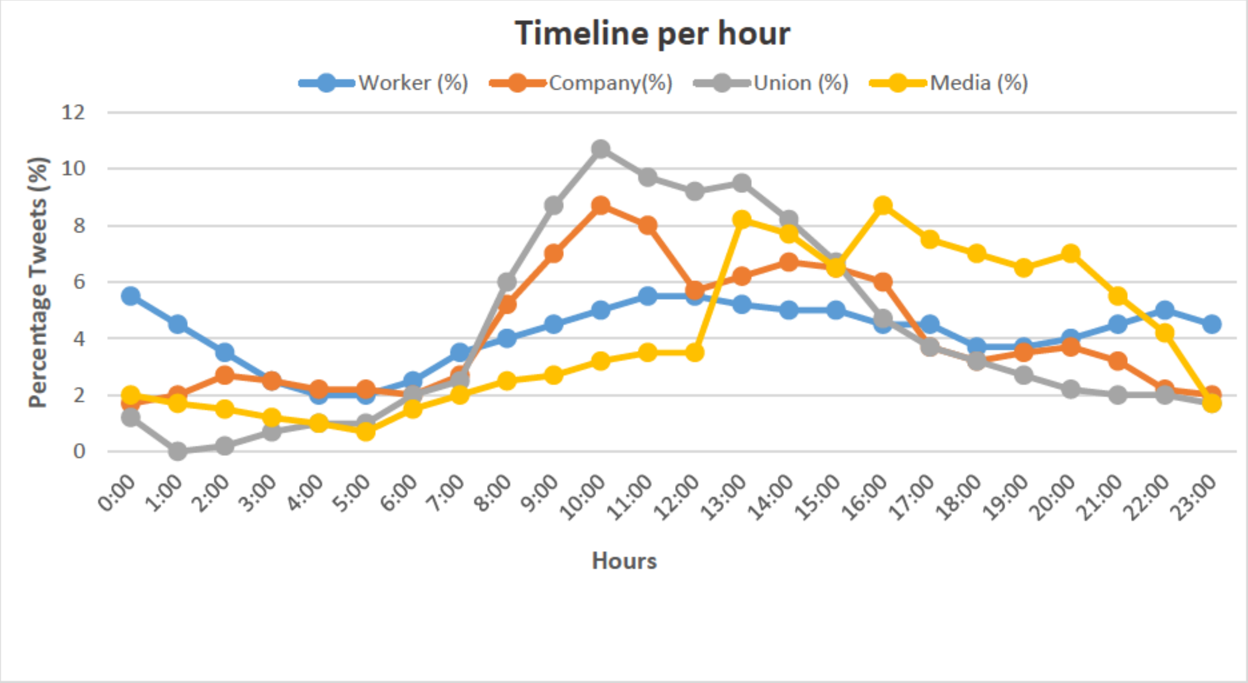


Figure 15: Timeline for hourly interactions on social media by the four cluster groups in the construction industry

The timeline analysis results also proved that those workers tend to more often during the waking morning hours and towards the late evening hours and less often during the working hours, which are typically from 9:00 am – 6:00 pm. This trend was also the same for the company group. The reverse is true for the other three clusters, who are seen to post more often during the working hours and less often during non-working hours. The companies and trade unions are more drawn to posting in the morning hours, probably since they have public relations personnel, who are tasked to post content in the morning hours, so as to attract the public's attention (Tang et al., 2017). The media and the Union generally posted all through the hours there were no great or notable peak posting hours for them in the social media as compared to the workers and the companies who showed some hours having peak interactions in the social media.

Chapter 6. Discussion and Concluding Remarks

6.1 Introduction to the Chapter

This chapter aims to discuss the results, from the previous chapter, and break down their meaning and implications on the social dynamics of stakeholders in the construction industry. The purpose of discussing these topics is to re-evaluate the findings of the five data analyses and see whether they fulfill the aim and objectives of the thesis. To achieve this, the aim and purpose of the thesis will be restated and then re-evaluated to determine whether the research efforts of this study have fulfilled this aim or not. In addition to this, the chapter will also give several concluding remarks to the thesis, including the limitations encountered during the study as well as feasible recommendations for future researchers.

6.2 Restating the Aim and Purpose of the Thesis

As earlier stated, the aim of this research was to explore and investigate the social interactions and links between the different stakeholders that are present in the construction industry. Investigating these interactions will help reveal a multitude of other related social aspects about the stakeholders, e.g., their genuine attitudes about the construction industry and how they feel being involved in this field of work. Learning the sociocultural dimensions of a company is also fundamental to understand the social dynamics between the stakeholders of a company, or organization, and model an organizational culture that best aligns itself to these dynamics (Di Fabio, 2017). Indeed, structuring the organizational culture to fit the stakeholders' sociocultural values is fast becoming a top strategy in creating great workplaces and healthy relationships between different departments within a company (Britt, 2020). De Waal (2018) emphasizes that establishing and maintaining healthy social interactions in an organization is key in ensuring that there is effective teamwork amongst the employees, trust and commitment, credibility amongst the members, and respect for each other. Strong social interactions also reinforce people’s loyalty, communication, shared values, and goals, as well as a sense of belonging and togetherness (Blustein et al., 2019). When people trust and respect each other, they are more likely to work together in unity and, thus, elevate the organization’s performance, productivity, profitability, and prosperity. In general, the facilitation of this research study and thesis will contribute to a more defined understanding of the qualitative attributes of the construction industry, including the attitudes and sentiments of the stakeholders regarding the construction industry, the relationships between the different stakeholders, and even their geographical locations.

6.3 The Social Relationship in the Construction Industry

Findings from the study revealed that the construction workers seem to have a poor relationship and social dynamics with the rest of the other clusters, i.e., companies, unions, and the media. The reason for this claim is because the findings of the study show that workers had generally negative sentiments, compared to the companies, unions, and the media. The other three clusters seemed to have a neutral stance against construction-related matters. When a topic modelling analysis was performed, the workers were, once again, found to be discussing rather personal topics that were related to work and labor. The other three clusters, on the other hand, focused on business-related topics. For example, majority of the posts published by the companies were about advertisements for their products and services, their ongoing or new construction projects, technologies, company events, social events, the construction market, earnings, and the general construction industry. The media too based most of their topics on issues such as employment policies in the construction industry, market trends and technologies, political news, construction news, and making advertisements on behalf of the construction companies, etc. Similarly, the unions focused on topics like worker advocacy issues, such as employment policies, fair employment for employees, protest news and events (e.g., city council involvement, release, August protest events, etc.), and working loads and reforms.

When it came to link analysis results, the companies had the highest number of followers and following, which translates to having the highest FF ratio than all the other clusters. The likely reason for this is because the companies always strive to stay professional throughout their social media content and engagement, posting information about the organizations and corporate events. The cluster with the second highest FF ratio is the media, followed by the unions, and lastly by the workers. As for types of tweets, the workers had the highest number of retweets and favorites, thanks to their more personal engagement with each other and closer interactions. Apparently, workers conduct themselves in a rather casual, informal, and easygoing manner, while the media, companies, and unions have to portray a professional and corporate image to the public. It is for this reason that the media and unions had lesser retweets, due to the professional and business-based nature of their tweets and the need to interrelate with each on a corporate basing.

6.4 The Key Improvement Factors

As has been discussed above, it is quite evident that there are poor social relations between workers and the other three clusters (i.e., companies, unions, and the media). There are varied reasons for these observations. The first and foremost reason may be because workers operate in a rather informal and casual manner. The companies, unions, and the media define their behavior in a much more formal and corporate attitude, hence they tend to relate to one another more than they do with workers. For example, when posting tweets, employees are more likely to post personal information on social media platforms and other online sites (Tang et al., 2017). This personal information may be disparaging as it may contain derogatory statements about employers and company managerial teams. For example, workers may be taking out their grievances against their employers and complaining about how they treat them, poor salaries, toxic working conditions, lack of job security, minimal work benefits or perks, or even reporting fraudulent activities within the companies. The companies, unions, and the media, on the contrary, only concentrate on business-related topics and never their personal issues.

There are several ways to achieve this, for example promote a healthy work-life balance, reward great work from the employees and provide constructive feedback, provide employee perks and benefits, as well as encourage job rotation and teamwork between various departments (Haar et al., 2019). As Britt (2020) states, employees also appreciate being kept in the loop and involved in various decision-making tasks. One of the ways to keep the employees involved in organizational affairs is to promote horizontal communication channels and using business proprietary platforms, such as Slack, Twine, Microsoft Teams, to communicate freely between people of different departments and hierarchies (De Waal, 2018). For example, a worker using Microsoft Teams can communicate directly to upper management and give certain updates, without having to draft long emails and going through all the office formalities. Getting rid of communication barriers between workers and the management will help improve the interaction flow between these two entities, hence enhance their relations. This way, workers will feel more connected with the companies they work for and even motivated to perform better within their job roles.

Better communication efforts and assurance of stable job security and employment are particularly dire in Utopia, where the construction workers reside. Therefore, construction companies and trade unions need to consider the plight of construction workers and engage with them at a large scale. Construction companies, for instance, can work towards achieving higher employee motivation in order to inspire workers and, hopefully, change their trend of posting negative sentiments and topics to post more positive sentiments. This way, the posting behavior of the workers will shift from using profane and obscene terminologies to being more positive and in line with the companies' objectives. Rather than castigate their employers for poor treatment, unstable job security, toxic working conditions, and low engagement, workers will start to talk more about their work achievements, career progress and roles, teamwork, and networking ventures. The construction unions, on the other hand, can try and engage more with their members (i.e., construction workers) in order to determine their grievances at the workplace and, in turn, offer more effective worker advocacy and support.

6.5 Limitations to the Study

The adoption of social media data analytics in construction industry is still infancy and limited by the mindset and technological challenges for now. Besides in this research there are many limitations including bias of some statistics such as number of tweets collected in a given time due to Twitter privacy on data collection. The limitation discovered during data research was that there was little literature on how to implement social media data analytics using the R programming language. This was initially mentioned as a research gap in the introduction chapter. There was clearly a literature research gap since there still is minimal literature that discusses the sociocultural dimensions in the construction industry, i.e., the work practices in construction companies, the relationships between different stakeholders in the companies, the organizational culture, and inclusivity and connectedness of stakeholders (or lack thereof) within this field of work. Instead, majority of the literature published on the construction industry mainly entails topics like safety risk identification and prevention. Studies such as Brandenbury et al. (2006), Sun et al. (2008), Hallowell (2012), and Fang et al. (2015) only cover topics like human resources in construction, safety risk identification and assessment, safety knowledge management, and supervisory worker safety behavior. Thus, for the most part, the research studies are only limited to workers' safety and preventive measures as well as construction business processes, such as project strategies etc. In addition to this, there was also a dearth of literature in social media data analytics in the construction industry, as the use of social media data analytics as a research methodology is still in its infancy stages. The lack of literature thus required extensive time to write and test the code for successfully performing sentiment analysis, topic modelling, link analysis, geo-location analysis, and timeline analysis.

6.6 Future Research Direction

A number of counteractive approaches need to be enforced in future to restore and develop healthy social relations between workers and the other three clusters. To begin with, the company management teams should endeavor to develop stronger interactions with the workers and improve the working conditions, in overall. Some of the long pattern of data is only observed through analysis of database from the time scale and the 4 clusters for Twitter analysis which may not be applicable in areas with different population (Dang-Xuan, L., 2013). However, the validation of the Twitter data for the 4 clusters opens a multiple opportunities for future research which includes; Collection of additional information from other social platforms to validate the methodology applied externally or internally, Identification of the requirement for the decision maker, Conducting the social influence on analysing the construction projects and Identification of construction labor's mental stability through sentimental analysis and implementation of the social media data analytics by use of R programming language (Batrinca, B.2015).

References

Abuhashesh, M. (2014). Integration of social media in business. *International Journal of Business and Social Science, 5*(8), 202-209.

Addo, M. & Eboh, W. (2014). Qualitative and quantitative research approaches. In Ruth Taylor (Ed.), *The Essentials of Healthcare and Nursing Research* (pp. 137–154). London: Sage Publications Ltd

Adzovie, D., Nyieku, I, & Keku, J. (2017). Influence of Facebook usage on employee productivity: A case of university of cape coast staff. *African Journal of Business Management, 11*(6), 110-116.

Agbaimoni, O., & Bullock, L. (2013). *Social media marketing – Why businesses need to use it and how (Includes a study of Facebook).* Warsaw, Poland: Institute of Aviation.

Alghamdi, R., & Alfalqi, K. (2015) A survey of topic modeling in text mining. *International Journal of Advanced Computer Science and Applications (IJACSA), 6*(1), 147-153.

Ali-Hassan, H., Nevo, D., & Wade, M. (2015). Linking dimensions of social media use to job performance: The role of social capital. *Journal of Strategic Information Systems, 24*(2), 65-89.

Alzahrani, H. (2016). Social media analytics using data mining. *Global Journal of Computer Science and Technology: C Software & Data Engineering, 16*(4), 1-4.

Anavizio. (2019). *Can social media analytics support conventional market research?* Retrieved from https://anavizio.com/can-social-media-analytics-support-market-research/

Andryani, R., Negara, E., & Triadi, D. (2019). Social media analytics: Data utilization of social media for research. *Journal of Information Systems and Informatics, 1*(2), 193-205.

Aung, K., & Myo, N. (2017). Sentiment analysis of students' comment using lexicon based approach. *Computer and Information Science (ICIS), IEEE/ACIS 16th International Conference IEEE,* pp. 149-154.

Barasa, A. (2012). *Social media as an effective advertising tool in Kenya.* Retrieved from http://erepository.uonbi.ac.ke/bitstream/handle/11295/76959/Thesis%20-%20Role%20of%20Social%20Media-%20%20Its%20Over.pdf?sequence=4

Batrinca, B., & Treleaven, P. (2015). Social media analytics: A survey of techniques, tools, and platforms. *AI & Society, 30*, 89-116.

BBC. (2013). *The age of big data: BBC Documentary (Low).* Retrieved from https://www.dailymotion.com/video/x16lvg8

Behrens, M., & Jacoby, W. (2004). The rise of experimentalism in German collective bargaining." *British Journal of Industrial Relations, 42*(1), 95–123.

Bizzi, L. (2018). *Employees who use social media for work are more engaged – but also more likely to leave their jobs.* Retrieved from https://hbr.org/2018/05/employees-who-use-social-media-for-work-are-more-engaged-but-also-more-likely-to-leave-their-jobs#:~:text=Social%20media%20can%20be%20a,media%20support%20decision%2Dmaking%20processes.

Blustein, D., Kenny, M., Di Fabio, A., & Guichard, J. (2019). Expanding the impact of the psychology of working: Engaging psychology in the struggle for decent work and human rights. *Journal of Career Assessment, 27*(1), 3–28.

Bo, H., Cook, P., & Baldwin, T. (2012). *Geolocation prediction in social media data by finding location indicative words.* Retrieved from https://www.aclweb.org/anthology/C12-1064.pdf

Bradbury, D. (2013). *Effective social media analytics.* Retrieved from https://www.theguardian.com/technology/2013/jun/10/effective-social-media-analytics

Bradley, A. (2010). Taking a strategic approach to social media. Retrieved from https://cdn.ymaws.com/www.tasscc.org/resource/resmgr/annual_conference/presentations/doc2011bradley_anthony.pdf

Brandenbury, S., Jaas, C., & Byrom, K. (2006). Strategic management of human resources in construction. *Journal of Management in Engineering, 2*(89), 89–96.

Brantner, C., & Pfeffer, J. (2018). Content analysis of Twitter: Big data, big studies. In S. A. Eldridge & B. Franklin (Eds.). *The Routledge handbook to developments in digital journalism studies.* Abingdon: Routledge.

Britt, R. (2020). Beyond the bottom line: *Why putting people first matters.* Retrieved from https://www.uschamber.com/co/start/strategy/putting-employees-before-profits

Brucker, P., Drexl, A., Möhring, R., Neumann, K., & Pesch, E. (1999). Resource-constrained project scheduling: Notation, classification, models, and methods. *European Journal of Operational Research, 112*(1), 3–41.

Bruhn, M., Schoenmueller, V., & Schäfer, D. B. (2012). Are social media replacing traditional media in terms of brand equity creation? *Management Research Review, 35*(9), 770-790.

Buntain, C., McGrath, E., Golbeck, J., & LaFree, G. (2016). *Comparing social media and traditional surveys around the Boston Marathon Bombing.* Retrieved from http://ceur-ws.org/Vol-1691/paper_02.pdf

Cambria, E., & White, B. (2014) Jumping NLP curves: A review of natural language processing research [review article]. *IEEE Computational Intelligence magazine*, *9*(2):48–57.

Canchen, L. (2019). Preprocessing methods and pipelines of data mining: An overview. *Seminar Data Mining*, 1-7.

Carneiro, B., & Costa, H. (2020). Digital unionism as a renewal strategy? Social media use by trade union confederations. *Journal of Industrial Relations*, *0*(0), 1-26.

Castronovo, C., & Huang, L. (2012). Social Media in an Alternative Marketing Communication Model. Journal of Marketing Development & Competitiveness, 6(1), 117-132.

Chung, C., & Austria, K. (2010). *Social Media Gratification and Attitude toward Social Media Marketing Messages: A Study of the Effect of Social Media Marketing Messages on Online Shopping Value. Proceedings of the Northeast Business & Economics Association.* Retrieved from: http://connection.ebscohost.com/c/articles/56100920/social-media-gratification-attitude-toward-socialmedia-marketing-messages-study-effect-social-media-marketing-messages-online-shopping-value

Comcowich, W. (2017). *Advantages of social media listening for market research.* Retrieved from https://www.business2community.com/marketing/advantages-social-media-listening-market-research-01905027

Crimson Hexagon. (n.d.). *The fundamentals of social media analytics.* Retrieved from https://www.upa.it/static/upload/the/the-fundamentals-of-social-media-analytics.pdf

Das, S., & Lall, G. (2016). Traditional marketing VS digital marketing: An analysis. *International Journal of Commerce and Management Research, 2*(8), 5-11.

de Waal, A. (2018). Increasing organizational attractiveness: The role of the HPO and happiness at work frameworks. *Journal of Organizational Effectiveness People and Performance, 5*(1), 124–141.

DeFeo, N. (2018). *How to run your business to build value and increase profitability.* Retrieved from https://www.growthforce.com/blog/how-to-run-your-business-to-build-value-and-increase-profitability

Deng, S., Lin, Y., Liu, Y., Chen, X., & Li, H. (2017). How do personality traits shape information-sharing behaviour in social media? Exploring the mediating effect of generalized trust. *Information Research: An International Electronic Journal, 22*(3), 1-37.

Dhawan, V., & Zanini, N. (2014). Big data and social media analytics. *Research Matters, 18*, 36-41.

Di Fabio, A. (2017). Positive healthy organizations: Promoting well-being, meaningfulness, and sustainability in organizations. *Frontiers in Psychology, 8*(1938), 1-6.

Fan, W., & Gordon, M. (2014). The power of social media analytics. *Communications of the ACM, 57*(6), 74-81.

Fang, D., Wu, C., & Wu, H. (2015). Impact of the supervisor on worker safety behavior in construction projects. *Journal of Management in Engineering, 31*(6), 04015001.

Farhadloo, M., & Rolland, E. (2016). *Fundamentals of sentiment analysis and its applications. In: Sentiment analysis and ontology engineering (pp. 1-24).* Retrieved from https://www.researchgate.net/publication/300965436_Fundamentals_of_Sentiment_Analysis_and_Its_Applications

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communication ACM, 56*, 82-89.

Fouss, F., Saerens, M., & Shimbo, M. (2016). *Algorithms and models for network data and link analysis*. Cambridge: Cambridge University Press.

Fox, J. (1998). Salting the construction industry. *William Mitchell Law Review*, *24*(3), 681-712.

Garcia, M. (2020). *How to make a Twitter bot in Python with Tweepy*. Retrieved from https://realpython.com/twitter-bot-python-tweepy/

Gidofalvi, G. (2012). *Spatio-temporal data mining for location-based services*. Retrieved from http://www.diva-portal.org/smash/get/diva2:501531/fulltext01.pdf

Gousios, G. (2020). *Big and fast data*. Retrieved from https:gousios.orgcoursesbigdatabig-data.html

Groen, B., Wouters, M., & Wilderom, C. (2017). Employee participation, performance metrics, and job performance: a survey study based on self-determination theory. *Management Accounting Research*, *36*(3), 51-66.

Guellil, I, & Boukhalfa, K. (2015) Social big data mining: A survey focused on opinion mining and sentiments analysis. In: *Programming and Systems (ISPS), 2015 12th International Symposium IEEE*.

Haar, J., Schmitz, A., Di Fabio, A., & Daellenbach, U. (2019). The role of relationships at work and happiness: A moderated mediation study of New Zealand managers. *Sustainability*, *11*(3443), 1-16.

Hallowell, M. (2012). Safety-knowledge management in American construction organizations. *Journal of Management in Engineering*, *28*(2), 203-211.

Hameed, H. (2020). *Quantitative and qualitative research methods: Considerations and issues in qualitative research*. Retrieved from https://www.researchgate.net/publication/342491265_Quantitative_and_qualitative_research_methods_Considerations_and_issues_in_qualitative_research

Hanley, K. (2014). *The impact of digital and social media on local television news stations*. Retrieved from https://core.ac.uk/download/pdf/190335592.pdf

Hardjono, T., & Marrewijk, M. (2001). The social dimensions of business excellence. *Corporate Environmental Strategy*, *8*(8), 1-13.

Hawkins, D. (2011). *The importance of relationships: Evaluating the dynamics and challenges of relationships in business*. Retrieved from http://kmhassociates.ca/resources/4/The%20importance%20of%20relationships%20in%20business.pdf

Hevey, D. (2015). Network analysis: A brief overview and tutorial. *Health Psychology and Behavioral Medicine*, *6*(1), 301-328.

Hodder, A., & Houghton, D. (2019). Unions, social media and young workers – Evidence from the UK. *New Technology, Work and Employment*, *35*, 40–59.

Hoit, D. (2013). *Big data, big expectations (Centre for Digital Education Report Q2 2013)*. Retrieved from http://www.centerdigitaled.com/paper/2013-Q2- Special-Report-Big-Data-Big-Expectations.html

Holsapple, C., Hsiao, S., & Pakath, R. (2014). Business social media analytics: Definition, benefits, and challenges. *Twentieth Americas Conference on Information Systems,* 1-12.

Hooder, A., & Houghton, D. (2015). Union use of social media: A study of the university and college union on Twitter. *New Technology Work and Employment, 30*(3), 173-189.

Houston, E. (2020). *The importance of positive relationships in the workplace.* Retrieved from https://positivepsychology.com/positive-relationships-workplace/

IBM. (2013). *Social media analytics: Making customer insights actionable.* Retrieved from http://www-01.ibm.com/software/analytics/solutions/customeranalytics/social-media-analytics

Icha, O., & Agwu, E. (2015). Effectiveness of social media marketing on organizational performance. *Journal of Internet Banking and Commerce, 21*(1), 1-7.

Jaafar, N., Al-Jadaan, M., & Alnutaifi, R. (2015). Framework for social media big data quality analysis. *New Trends in Database and Information Systems II,* 301–314.

Jafar, R., Geng, S., Ahmed, W., & Niu, B. (2019). *Social media usage and employees' job performance: The moderating role of social media rules.* England: Emerald Publishing Limited.

Jalayer Academy. (2017). *Text mining (part 3) – Sentiment analysis and wordcloud in R (single document).* Retrieved from https://www.youtube.com/watch?v=JM_J7ufS-BU

Jiang, H., Lin, P., & Qiang, M. (2016). Public opinion sentiment analysis for large hydro projects. Journal of Construction Engineering and Management, 142(2), 1-12.

Jurgens, P., & Jungherr, A. (2016). *A tutorial for using Twitter data in the social sciences: Data collection, preparation, and analysis.* Retrieved from https://www.researchgate.net/publication/289524980_A_Tutorial_for_Using_Twitter_Data_in_the_Social_Sciences_Data_Collection_Preparation_and_Analysis

Kalil, T. (2012). *Big data is a big deal.* Retrieved from https://obamawhitehouse.archives.gov/blog/2012/03/29/big-data-big-deal

KDnuggets. (2014). *Online education in analytics, big data, data mining, and data science.* Retrieved from http://www.kdnuggets.com/education/online.html

Khan, M., Durrani, M., Ali, A., Inayat, I., Khalid, S., & Khan, K. (2016). Sentiment analysis and the complex natural language. *Complex Adaptive Systems Modeling, 4*(2), 1-19.

Kharde, V., & Sonawane, S. (2016). Sentiment analysis of Twitter data: A survey of techniques. *International Journal of Computer Applications, 139*(11), 5-15.

Kherwa, P., & Bansal, P. (2018). Topic modelling: A comprehensive review. *ICST Transactions on Scalable Information Systems, 7*(24), 159-623.

Khoo, C., & Johnkhan, S. (2017). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science, 44*(6), 1-21.

Kumar, G. (2015). *Guide for building Facebook applications with PHP.* Finland: Oulu University of Applied Sciences.

Lawrence, R., Melville, P., Perlich, C., Sindhwani, V., Meliksetian, S., Hsueh, P, … Liu, Y. (2015). *Social media analytics.* Retrieved from https://www.researchgate.net/publication/283360335_Social_media_analytics

Leung, M., Yu, J., & Liang, Q. (2013). Improving public engagement in construction development projects from a stakeholder's perspective. Journal of Construction Engineering and Management, 139(11), [4013019].

Liu, L., Tang, L., Dong, W., Yao, S., & Zhou, W. (2016) An overview of topic modeling and its current applications in bioinformatics. *Springer Plus*, *5*(1608), 1-22.

Luengo, J., Garc´ıa, L., & Herrera, F. (2015). *Data preprocessing in data mining.* New York: Springer.

Mahmud, J., Nichols, J., & Drews, C. (2012). Where is this tweet from? Inferring home locations of twitter users. *Proceedings of Sixth International AAAI Conference on Weblogs and Social Media*, *6*(1), 511-514.

Maks, I., & Vossen, P. (2012). A lexicon model for deep sentiment analysis and opinion mining applications. *Decision Support Systems*, *53*(4), 680–688.

Mankins, M. (2017). *Stop focusing on profitability and go for growth.* Retrieved from https://hbr.org/2017/05/stop-focusing-on-profitability-and-go-for-growth

Mayer-Schönberger, V. & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think.* Boston: Houghton Mifflin Harcourt.

McAllister, P. (2015). *The importance of trade unions.* Retrieved from https://www.ethicaltrade.org/blog/importance-trade-unions

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1-21.

Moetti-Lysson, J., & Ongori, H. (2011). Effectiveness of trade unions in promoting employee relations in organizations. *Global Journal of Arts & Management*, *1*(4), 57-64.

Nguyen, H. (2014). *Should social media analytics replace traditional marketing research?* Retrieved from https://netbasequid.com/blog/social-media-analysis-replace-traditional-marketing-research/

Olson, D., & Laughoff, G. (2019). *Link analysis: Descriptive data mining (pp. 107-128). Retrieved* from https://www.researchgate.net/publication/332882713_Link_Analysis

Othman, A. (2020). *Social media data mining and analytics.* Retrieved from https://www.academia.edu/37513686/Social_Media_Data_Mining_and_Analytics_pdf

Pasquier, V., & Wood, A. (2018). The power of social media as a labor campaigning tool: Lessons from our Walmart and the fight for 15. *European Economic, Employment and Social Policy*. Retrieved from https://www.researchgate.net/publication/334576073_The_power_of_social_media_as_a_labour_campaigning_tool_lessons_from_OUR_Walmart_and_the_Fight_for_15

Ramirez-Gallego, S., Garcia, S., Luengo, Benitez, J., & Herrera, F. (2016). Big data preprocessing: Methods and prospects. *Big Data Analytics, 1*(9), 1-22.

Roffey, S. (2016). Positive relationships at work. In: *The Wiley Blackwell Handbook of the Psychology of Positivity and Strengths-Based Approaches at Work* (pp.171-190). London, UK: Wiley-Blackwell.

Roller, S., Speriosu, M., Rallapalli, S., Wing, B., & Baldridge, J. (2012). *Supervised text-based geolocation using language models on an adaptive grid.* Retrieved from https://www.researchgate.net/publication/265479236_Supervised_Text-based_Geolocation_Using_Language_Models_on_an_Adaptive_Grid

Rosales, R. (2015). *Energizing social relationships at work: An exploration of relationships that generate employee and organizational thriving.* Retrieved from https://repository.upenn.edu/cgi/viewcontent.cgi?article=1087&context=mapp_capstone

Sadia, A., Khan, F., & Bashir, F. (2018). *An overview of lexicon-based approach for sentiment analysis. International Electrical Engineering Conference.* Karachi, Pakistan: IEP Center.

Sahitreddy, M. (2020). *What is big data and how Facebook is using big data?* Retrieved from https:sahithreddy639.medium.comwhat-is-big-data-and-how-facebook-is-using-big-data-d044fcd52948#:~:text=Big%20Data%20at%20Facebook&text=Every%20day%2C%20we%20feed%20Facebook's,day%20%E2%80%94%20that's%20a%20million%20gigabytes.

Siricharoen, W. (2012). Social media, how does it work for business? *International Journal of Innovation, Management and Technology, 3*(4), 476-479.

Spaiser, V. (2016). *Collecting and analyzing Twitter data – An introduction.* Retrieved from https://quantcrimatleeds.github.io/slides/2016-09-28-V_Spaiser.pdf

Stack Overflow. (2012). *Using the Facebook API compared with the Twitter API.* Retrieved from https://stackoverflow.com/questions/7047195/using-the-facebook-api-compared-with-the-twitter-api

Staff, C., King, H., Roberts, M., Pannell, S., Roberts, D., Wilson, N., … Cooper, A. (2016). *Using social media for social research: An introduction.* Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/524750/GSR_Social_Media_Research_Guidance_-_Using_social_media_for_social_research.pdf

Stieglitz, S., & Linh, D. (2013). Social media analytics and political communication: A social media analytics framework. *Social Network Analysis and Mining, 3*(4), 1277-1291.

Su, Y., Cong, W., & Liang, H. (2019). The impact of supervisor-worker relationship on workers' safety violations: A modified theory of planned behavior. *Journal of Civil Engineering and Management, 25*(7), 631-645.

Sun, Y., Fang, D., Wang, S., Dai, M., & Lv, X. (2008). Safety risk identification and assessment for Beijing Olympic venues construction. *Journal of Management in Engineering, 1*(40), 40–47.

Swanner, N. (2016). *Why Twitter is a better network for real life than Facebook.* Retrieved from https://thenextweb.com/news/twitter-better-than-facebook#:~:text=Facebook%20lets%20you%20connect%20with,is%20still%20more%20feature-rich.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon based methods for sentiment analysis. *Computational linguistics*, *37*(2), 267-307.

Tang, L., Shen, Q., Skitmore, M., & Wang, H. (2015). Procurement-related critical factors for briefing in public-private partnership projects: Case of Hong Kong. Journal of Management in Engineering, 31(6), 1-10.

Tang, L., Zhang, Y., Dai, F., Yoon, Y., Song, Y., & Sharma, R. (2017). Social media data analytics for the U.S. construction industry: Preliminary study on Twitter. *Journal of Management and Engineering*, *33*(6), 1-15.

Tumen, S., & Zeydanli, T. (2015). Social interactions in job satisfaction. *International Journal of Manpower*, *37*(3), 2-31.

Wells, C., & Thorson, K. (2015). Combining big data and survey techniques to model effects of political content flows in Facebook. *Social Science Computer Review, 28,* 24-44.

Wharton, K. (2012). *Why companies can no longer afford to ignore their social responsibilities. Management & Leadership.* Retrieved from https://business.time.com/2012/05/28/why-companies-can-no-longer-afford-to-ignore-their-social-responsibilities/

Woldesenbet, A., Jeong, H. D., & Park, H. (2016). Framework for integrating and assessing highway infrastructure data. *Journal of Management in Engineering, 32*(1), 04015028.

Yin, R. (2015). *Qualitative research from start to finish.* New York, NY: The Guilford Press

Yu, P. (2019). *How to access Twitter's API using Tweepy. Retrieved* from https://towardsdatascience.com/how-to-access-twitters-api-using-tweepy-5a13a206683b

Zhang, M., & Chen, Y. (2018). *Link prediction based on graph neural networks. Retrieved* from https://papers.nips.cc/paper/2018/file/53f0d7c537d99b3824f0f99d62ea2428-Paper.pdf

Ho, C. C., Lim, W. L., & Yee, T. (2020). Sentiment Analysis by Fusing Text and Location Features of Geo-Tagged Tweets. *IEEE Access, 8.* https://doi.org/10.1109/ACCESS.2020.3027845