Paper No: SC-10 **Smart Computing**

LYZGen: A mechanism to generate leads from Generation Y and Z by analysing web and social media data

Janaka Senanayake*
Department of Industrial Management
University of Kelaniya, Sri Lanka
janakas@kln.ac.lk

Nadeeka Pathirana
Department of Information Technology
University of Sri Jayawardenapura, Sri Lanka
pathirana@sjp.ac.lk

***Abstract* - Identifying an appropriate target audience is essential to market a product or a service. A proper mechanism should be followed to generate these potential leads and target audiences. The majority of people who were born between 1981 and 2012 hold top positions in companies. These people are regular social media and website users, since they represent generations Y and Z. They usually keep digital footprints. Therefore, if an accurate method is followed, it is possible to identify potential contact points by analysing publicly available data. In this research, a novel lead generation mechanism based on analysing social media and web data has been proposed and named LYZGen (*L*eads of *Y* and *Z Gen*erations). The input to the LYZGen model was an optimised search query based on the user requirement. The model used web crawling, named entity recognition (NER), and pattern identification. The model found and analysed freely available data from social media and other websites. Initially, person name identification was performed. An extensive search was carried out to retrieve peoples' contact points such as email addresses, contact numbers, designations, based on the identified names. Cross verification of the analysed details was conducted as the next step. The results generator provided the final output, which contained the leads and details. Generated details were verified with responses captured via a survey and identified that the model could detect lead details with 87.3% average accuracy. The model used only the open data posted on the internet by the people. Therefore, it did not violate extensive privacy or security concerns. The generated results can be used, in several ways, including communicating promotional details to the potential target audience.**

Keywords - lead generation, named entity recognition, web crawling, web data analysing

I. INTRODUCTION

There is a high number of instances of communicating about promotional details related to products and services. However, in most cases, these communications are conducted without identifying the potential audience. Resources and time of the advertisers or promotional campaign organisers might be wasted because of this. Therefore, identifying the potential leads should be the initial task of this whole process.

These potential leads can be generated by thoroughly analysing the web data [1]. Many young people who belong to Generation Y and Generation Z tend to keep digital footprints knowingly or unknowingly when they browse the internet and social media. That is the nature of Generation Y [2] and Generation Z [3].

Web crawling and web data analysis techniques can be applied to analyse the content of a web page [4], which is also known as web scraping. By using a spider, the analysis can be performed in web scraping, and by using NER, person names can be identified [5] after analysing a textual input.

The use of web crawlers and web data analysis is not a novel area since various approaches were already proposed by academia. Their strengths and limitations are also discussed [6]. However, combining web crawlers to generate leads after identifying generation Y and Z behaviour in the digital space is not considered. The usage of websites and social media has increased rapidly, especially among the generations that were focused in this study. This increase is due to the travel restrictions imposed with the ongoing Covid-19 pandemic situation. In this paper, a model to detect leads and contact details of persons, using web crawling, web data analysis and named entity recognition, has been proposed. The generated data were validated again using web data analysis to determine the accuracy. In the model, all the steps in data collection and analysis were conducted on publicly available data on the web. Since the details were not extracted using any illegal approaches, there are no significant concerns of privacy violations [7].

Following research questions were answered in this research while building the LYZGen model.

- RQ1: What are the optimising strategies of web search queries?
- RQ2: How to apply web crawling and web data analysis to generate leads?
- RQ3: How to perform valid pattern recognition processes to identify lead-related attributes?
- RQ4: How to validate the accuracy of the contact details of the potential leads?

The generated details were re-evaluated for their accuracies by comparing them with survey results. This survey was conducted to record the name, details of designation, email address, and contact number from volunteers from academic, medical, financial and information technology fields. The survey results contain 179 records.

The rest of the paper is organized as follows: Section II contains related work. Section III gives an overview and the methodology of the LYZGen system to generate potential leads. Section IV presents the results and discussions related to the research. Finally, the conclusions and future work directions are discussed in Section V

II. RELATED WORK

There are various research studies conducted in the research areas of identifying leads, mechanisms of web crawling and web data analysing, NER methods, and user generations. However, to the best of our knowledge, there is no comprehensive research conducted after combining each of these individual areas to build a proper lead generation mechanism. In this section, related research studies in those mentioned areas are discussed.

People who live all around the world can be categorized based on different dimensions. Among all these dimensions, the "generation" has become one of the important societal categories introduced [8]. In the human context, a generation is defined as a group of people who were born and nurtured at a specific time. They have common characteristics and viewpoints which are affected by their growing time. It implies that there are characteristic discrepancies among generations.

In the current society, four to five generations are working side by side [9]. Among them, generation Z and generation Y are the latest generations who work in society nowadays. They deal with technology frequently. Generation Y is the first generation of people who came into the world of technology [2] when they were born. Generation Z is the first generation born with the technology; known as digital natives [3]. The new generation always tries to perform their tasks efficiently with the help of technology [10]. The research conducted in [11] identifies the fact that the leaders of using technology are the people from generation Y and Z. Compared to the other generations, they spend a significant amount of time surfing and browsing the internet for different purposes. Due to this behaviour, they tend to keep their digital footprints in cyberspace more than the other generations do.

One of the most important facts to consider when dealing with technology is security. Most people use their mobile devices to engage with the digital space actively. Several techniques used to detect and prevent attacks on the users, such as data theft, social engineering, and malware have been identified [12]. However, generation Y and Z people should consider their digital footprint to keep themselves safe since there are some ways for obtaining these footprints, which includes recording of footprints with or without the consent or acknowledgement of the users.

These digital footprints of generations Y and Z can be collected and analysed to generate leads. Lead generation is one of the most common marketing approaches used to identify potential customers. This method helps identify the target audience for a particular domain. Through identifying contact points, it is easy to reach the right people [13]. Various lead generation methods are being practised on several occasions [14].

To find the leads and related details, one way that can be used is to analyse the web page contents. An automated mechanism should be integrated to achieve that. Web crawling and web data analysis are the methods, which can be applied to this [15]. Web crawling is also known as web scraping. In web scraping, the feature known as spider visits websites and scrapes all the data after performing an analysis. In [4] and [16], many methods are proposed to conduct web scraping. Out of those methods, the spider-based web scraping method was identified as the efficient method, and this is used in many web search engines.

The scrapped content should properly be analysed though an extensive web crawling method. If the content can be formed into a textual string, it is possible to apply several text mining methods to identify patterns [17]. NER is one of the commonly used methods to identify names with the help of text mining and natural language processing. NER can be categorised into three main categories as Hand-made Rule-based NER, Machine-Learning based NER and Hybrid NER. Mining names using human-made rules set is known as hand-made rule-based NER. Machine Learning-based NER can identify problems and classify problems, and then the System identifies patterns and relationships. After that, it makes a model using available statistical models and machine learning algorithms. Hybrid NER is the combination of rule-based and Machine Learning based NER approaches [18]. Based on the requirements, the types that need to be recognized could vary. Recognition can be done for a person, contact details, location, or other information related to a specific task.

Privacy and security of web data are also important to be considered. With technological enhancement, people tend to use online resources to do their day-to-day activities efficiently. When people use different web applications and mobile applications, they create social networks through digital platforms. Due to this, people make their details available in public, knowingly or unknowingly. These details may contain their experiences, opinions and knowledge. There can be private data such as name, contact information, gender, etc. [19] among those details. Sharing this type of information could have both. a negative and a positive effect. If a person shares sensitive information, a negative impact for that user can also be generated. For example, insurance companies can collect that information to identify users as risky clients [20].

III. METHODOLOGY

To overcome the identified problem by addressing the formulated research questions, the LYZGen model is proposed, as described in this section.

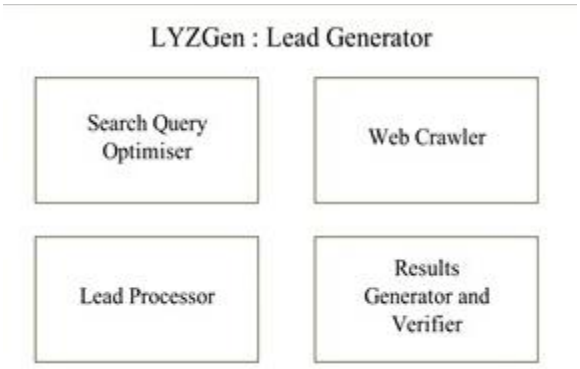

Fig 1. LYZGen architecture

The methodology of this work was distributed among four sub-systems. These subsystems were named Search Query Optimiser, Web Crawler, Lead Processor, and Results Generator and Verifier. The overview of the system is illustrated in Figure 1. Each of these subsystems were connected to generate verified results on potential leads.

An interface of the prototype system which performed those four subsystem processes is illustrated in Figure 2. The prototype was developed using the Java programming language.

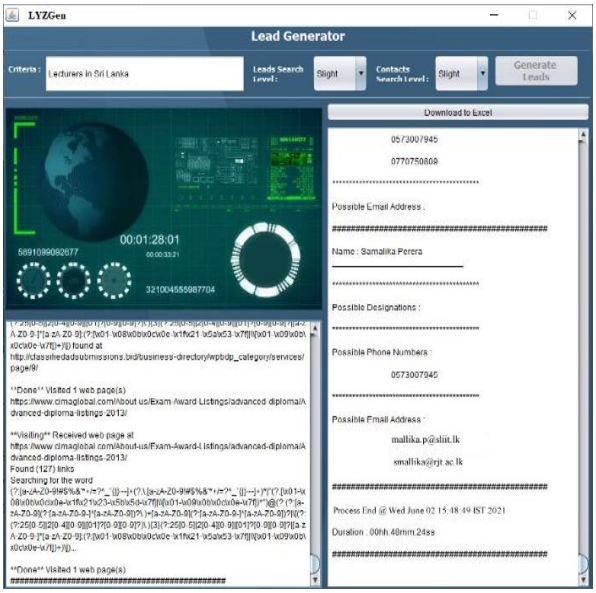

Fig 2. LYZGen prototype

A. Search query optimiser

The initial input to the model is the search criteria. There is a difference between web search queries entered by a person having good computer literacy and a regular person. But this model can be used by anyone. Therefore, search query optimisation should be performed as the first task to retrieve accurate search results [21]. The search criteria entered by the user was split into words, and a string array was created. Then, using "OR" and "AND" operators, the search query was optimised. The pre-stored article words were not taken into consideration initially when preparing the search string. Some examples for optimised search queries are listed in Table I.

TABLE I. SEARCH QUERY OPTIMISATION

User Input	Optimised Search Query
Lecturers in Sri Lanka	("Lecturers") AND ("Sri Lanka") OR ("Sri Lanka") AND ("Lecturers") OR ("Lecturers Sri Lanka") OR ("Sri Lanka Lecturers") OR ("Lecturers in Sri Lanka")
Cricket Players in Sri Lanka	("Cricket") AND ("Players") AND ("Sri Lanka") OR ("Cricket") AND ("Sri Lanka") AND ("Players") OR ("Sri Lanka") AND ("Cricket") AND ("Players") OR ("Sri Lanka") AND ("Players") AND ("Crickets") OR ("Players") AND ("Cricket") AND ("Sri Lanka") OR ("Players") AND ("Sri Lanka") AND ("Cricket") OR ("Cricket") AND ("Players") OR ("Cricket") AND ("Sri Lanka") OR ("Players") AND ("Cricket") OR ("Players") AND ("Sri Lanka") OR ("Sri Lanka") AND ("Cricket") OR ("Sri Lanka") AND ("Players") OR ("Cricket Players in Sri Lanka") OR ("Cricket Players Sri Lanka")

B. Web crawler

Web crawler performs the searching and crawling process of the model. Once the search query was optimised, the web crawler was activated. The crawler can be customised with search depth (known as the Leads Search Level) as "Slight", "Low", "Moderate", "Strong", and "Extreme". The number of outputs depends on the depth level. The time it takes to complete the search depends on the number of words the user input and the search depth. Then the search depth was converted into a numeric value. Values from 1 to 5 were assigned from Slight to Extreme. For example, if the user selects Strong (value is 3) as the search depth, the web crawler visits 30 (3×10) links and their sub-links in search engine results. The reason for multiplying by 10 is that one page of a search engine results contains ten results (links). The links visited by the web crawler were stored in a Java Collection to further process in the Lead Processor subsystem. Once the Lead Processor requests the crawl process to identify names and contact details, the Web Crawler subsystem crawled web pages while considering "Contact Search Level" as one parameter which defines the depth of the data analysis process of a given web link. The Contact Search Level also has five levels: "Slight", "Low", "Moderate", "Strong" and "Extreme". Similar to the Leads Search Level parameter, this also represents values from 1 to 5, and the value was multiplied by 10. "Mozilla/5.0 (compatible; Googlebot/2.1", "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/535.1 (KHTML, like Gecko) Chrome/13.0.782.112 Safari/535.1", and "Mozilla/5.0 (Windows; U; Windows NT 6.1; en-US)" were used as the user agents when crawling the web pages [22]. The web pages can be either regular websites or social media sites.

C. Lead processor

The Lead Processor is the subsystem where most of the important steps happen in the overall model. Initially, this subsystem takes the input as the Java Collection (List) generated by the web crawler. As the first step of the lead processor, the names of the leads were identified using pattern recognition and NER. The lead processor sent a request to the Web Crawler subsystems with a list of links, and then the crawler visited each link. The pattern of the name was determined using "[A-Z]([a-z]+) [A-Z]([a-z]+)" regular expression. Identified possible names were stored in a Java Hash Set to avoid duplicates. The set was iterated through several NER classifiers to identify the person names using a similar process which was followed in [23]. This model used *english.nowiki.3 class*, *english.conll.4 class*, *english.all.3 class* and *english.muc.7 class* NER classifiers [24]. Java libraries developed using those classifiers were used in the LYZGen with the category information of "PERSON" [25]. Once the names are properly identified, the Lead Processor calls the Web Crawler subsystem to determine their contact numbers, email addresses, and designations. When calling the web crawler, the search queries were modified to receive accurate results based on the type of information (i.e. contact number, email, designation). For the email address pattern recognition, an advanced regular expression *(?:[a-zA-Z0-9!#$%&'*+/=?^_`{|}~-]+(?:\\.[a-zA-Z0-9!#$%&'*+/=?^_`{|}~-]+)*|\"(?:[\\x01-\\x08\\x0b\\x0c\\x0e-\\x1f\\x21\\x23-\\x5b\\x5d-*

\\x7f]|\\\\[\\x01-\\x09\\x0b\\x0c\\x0e-\\x7f])\")@(?:(?:[a-zA-Z0-9](?:[a-zA-Z0-9-]*[a-zA-Z0-9])?\\.)+[a-zA-Z0-9](?:[a-zA-Z0-9-]*[a-zA-Z0-9])?|\\[(?:(?:25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\\.){3}(?:25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?|[a-zA-Z0-9-]*[a-zA-Z0-9]:(?:[\\x01-\\x08\\x0b\\x0c\\x0e-\\x1f\\x21-\\x5a\\x53-\\x7f]|\\\\[\\x01-\\x09\\x0b\\x0c\\x0e-\\x7f])+)\\]*), was used. Contact numbers were generated using "(0|94)[1-9]\\d{8}" regular expression. Finally, the results generated in the Lead Processor were stored in a Java Map.

D. Results generator and verifier

The generated results from the Lead Processor were used in the Results Generator and Verifier subsystem to provide the results after the verification process. A two-step verification process was conducted. This subsystem sent requests with four parameters which were 1) generated name, 2) designation, 3) contact number, and 4) email address, to the Web Crawler subsystem to perform the initial verification of the details available in the map generated in the Lead Processor. Requests were formatted, and the map was iterated to retrieve every detail. Once the Web Crawler subsystem sent more accurate responses, the revised version of the leads map was generated. As the second step of verification, the Truecaller [26] Application Programming Interface was used to generate further accurate data on the contact number and email addresses.

As the final step of the LYZGen, it generates a comma-separated version (CSV) file with all the details. The overall model is illustrated in Figure 3.

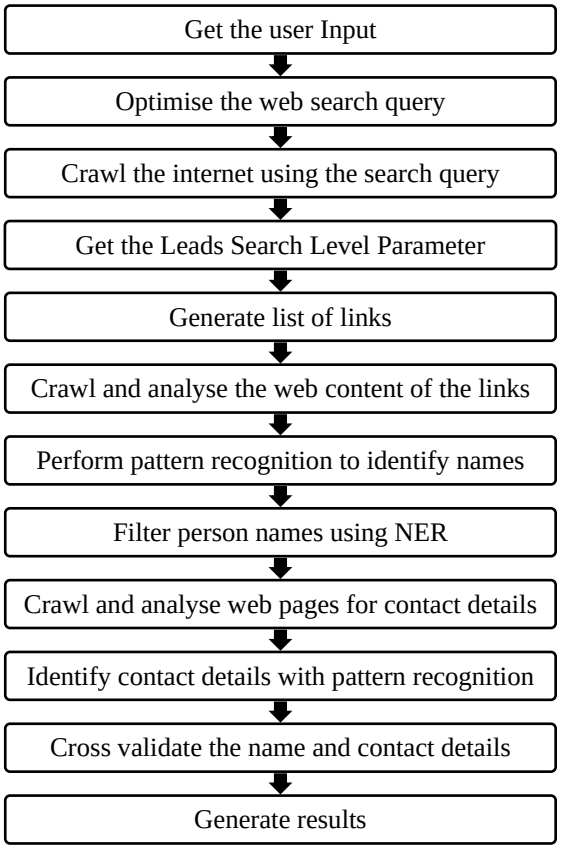

Fig 3. LYZGen model

IV. Results and Discussion

Though the LYZGen provides high accuracy results, a survey-based method was also used to validate the model accuracy. The survey was conducted to capture the details (name, designation, email address, and contact number) of individuals. The data was obtained for one month, from 1st March to 31st March 2021. The questionnaire was distributed electronically to selected categories such as medical officers, lecturers, banking officers and software engineers representing generations Y and Z. There were 179 records available in the survey results. LYZGen model was also executed for the same criteria (i.e. Medical Officers in Sri Lanka, Lecturers in Sri Lanka, Banking Officers in Sri Lanka and Software Engineers in Sri Lanka). The LYZGen model was executed with the parameter of Moderate for both Leads Search Level and Contact Search Level.

Afterwards, a comparison was conducted between the results generated from the LYZGen model and the survey results. The results count is illustrated in Figure 4. In this comparison, it was identified that Generation Y and Z people keep digital footprints compared to other generations. Many of them work in industries such as IT and financial companies, where the possibility of keeping digital footprints is high since these industries are closely associated with digital space. Due to the selection of a limited sample for the survey, it was not possible to differentiate all the results from the LYZGen model since the model contains some data from people from other generations who were also good at using technologies and were closely associated with cyberspace. But it is possible to say that those results can be treated as outliers.

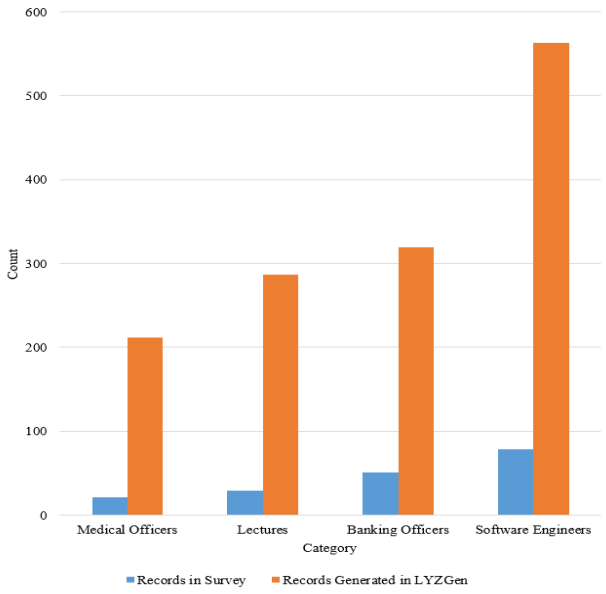

Fig 4. Results of survey and LYZGen model

A series of analyses were conducted on the names, designations, email addresses and contact numbers. Each record of the survey results was compared with the results of the LYZGen model and the matching records were identified. Table II shows the comparison of matching records in the two results sets, and Table III compares the accuracies of the LYZGen generated results in terms of

name, designation, email address, and contact number. According to analysis in Table III, it is possible to say that the people from generations Y and Z are closely associated with cyberspace as the number of records in the survey were closely matched the LYZGen results.

TABLE II. COMPARISON OF MATCHING RECORDS

Category	# Records in the Survey	# Matching Records with LYZGen			
		Names	Designations	Email Add.	Contact Nos
Medical Officers	21	19	18	19	17
Lectures	29	27	26	26	24
Banking Officers	51	46	44	45	43
Software Engineers	78	71	68	67	63

TABLE III. COMPARISON OF ACCURACIES OF LYZGEN RESULTS

Category	Accuracy of LYZGen Results (%)			
	Names	Designations	Email Add.	Contact Nos
Medical Officers	90.48	85.71	90.48	80.95
Lectures	93.10	89.66	89.66	82.76
Banking Officers	90.20	86.27	88.24	84.31
Software Engineers	91.03	87.18	85.90	80.77

Fig. 5 shows the average accuracies of the LYZGen model when identifying attributes of leads.

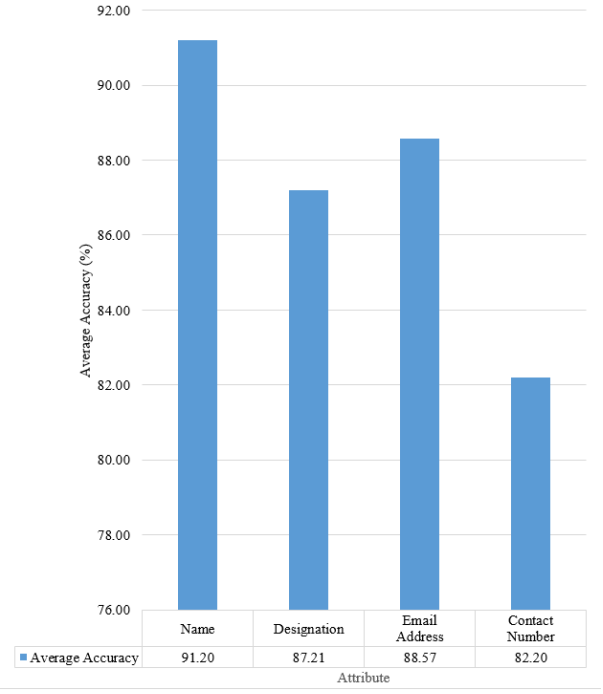

Fig 5. Average accuracies of lead details

Therefore, by analysing the above results, it is identified that the LYZGen model has a high accuracy of detecting names. The reason for that might be the attribute to be easily found when performing a web crawling process is the name. The email address is the second-highest attribute, and designation is the third. The reason for that would be, the email addresses and employee designations are relatively easy to find in the publicly available web data. However, generating contact numbers is a difficult task. The reason for that is, they are hard to find in public sources. The accuracy of identifying contact numbers is somewhat low compared to the other attributes. The reason for that might be due to some pattern recognition issues. Overall, it is identified that the LYZGen model can identify leads and attributes with 87.3% average accuracy.

V. CONCLUSION AND FUTURE WORK

Having a proper lead generation mechanism is valuable in communicating promotional activities to the appropriate audience. Since generation Y and Z use technology and the internet more, it is possible to find digital footprints. In this paper, a novel lead generation mechanism was proposed, named LYZGen, to identify leads' details such as name, designation, email addresses, and contact numbers by analysing digital footprints and freely available data in websites and social media sites. There were four subsystems in the proposed model to perform lead generation with cross-validations. A survey was also conducted to validate the model. It was identified that the model can generate data with an average accuracy of 87.3%. The LYZGen model can be used by anyone who wants to generate leads from publicly available data without violating major privacy concerns. LYZGen can be used to generate leads to improve the strategies of marketing campaigns by identifying the most suitable target audience.

Though the generated results were conducted only in the Sri Lankan context, this model can generate results without limiting them to the context. The accuracy of generating results can be increased by improving some of the areas in the LYZGen model. We identified that the model sometimes detects incorrect person names not from the specific country due to the limitation of the NER classifier. That can be omitted if a context-based NER classifier is introduced. Currently, if the search level is selected as "Extreme", it will take a lot of time to generate the results since the crawler has to visit many web pages. The efficiency of the model can be further improved. Furthermore, the dataset generated from the current LYZGen model can be used in future research areas related to leads and contact details. Once a high number of data are collected, it will be possible to apply machine learning to improve accuracy.

REFERENCES

[1] J. M. D. Senanayake and W. P. N. H. Pathirana, "Developing a Lead Generation Mechanism to Identify People's Contact Points Using Web Data Analytics," in Uva Wellassa University of Sri Lanka, Badulla, Sri Lanka, 2019.

[2] S. Prasad, A. Garg and S. Prasad, "Purchase decision of generation Y in an online environment," Marketing Intelligence & Planning, vol. 37, no. 4, pp. 372-385, 2019.

[3] W. P. N. H. Pathirana and D. N. Wickramaarachchi, "Software usability improvements for Generation Z oriented software application," in 2019 International research conference on smart computing and systems engineering (SCSE), Colombo, Sri Lanka, 2019.

[4] Hernández, C. R. Rivero and D. Ruiz, "Deep Web crawling: a survey," World Wide Web, vol. 22, no. 4, pp. 1577--1610, 2019.

[5] Goyal, V. Gupta and M. Kumar, "Recent named entity recognition and classification techniques: a systematic review," Computer Science Review, vol. 29, pp. 21-43, 2018.
[6] M. Kumar, R. Bhatia and D. Rattan, "A survey of Web crawlers for information retrieval," WIREs Data Mining and Knowledge Discovery, vol. 7, no. 6, p. e1218, 2017.
[7] S. Ribeiro-Navarrete, J. R. Saura and D. Palacios-Marqués, "Towards a new era of mass data collection: Assessing pandemic surveillance technologies to preserve user privacy," Technological Forecasting and Social Change, vol. 167, p. 120681, 2021.
[8] L. Duxbury and C. Higgins, "An empirical assessment of generational differences in work-related values," Human Resources Management Ressources Humaines, p. 62, 2005.
[9] J. Bejtkovsk`y, "The employees of baby boomers generation, generation X, generation Y and generation Z in selected Czech corporations as conceivers of development and competitiveness in their corporation," Journal of Competitiveness, 2016.
[10] J. M. D. Senanayake and W. M. J. I. Wijayanayake, "Applicability of crowd sourcing to determine the best transportation method by analysing user mobility," International Journal of Data Mining & Knowledge Management Process, vol. 8, no. 4/5, pp. 27-36, September 2018.
[11] T. Issa and P. Isaias, "Internet factors influencing generations Y and Z in Australia and Portugal: A practical study," Information Processing & Management, vol. 52, no. 4, pp. 592-617, 2016.
[12] J. Senanayake, H. Kalutarage and M. O. Al-Kadri, "Android Mobile Malware Detection Using Machine Learning: A Systematic Review," Electronics, vol. 10, no. 13, p. 1606, 2021.
[13] M. Rodriguez and R. M. Peterson, "The role of social CRM and its potential impact on lead generation in business-to-business marketing," International Journal of Internet Marketing and Advertising, vol. 7, no. 2, pp. 180-193, 2012.
[14] Gupta and N. Nimkar, "Role of Content Marketing and it's Potential on Lead Generation," Annals of Tropical Medicine and Public Health, vol. 23, no. 17, 2020.
[15] D. Shestakov, "Current challenges in web crawling," in International Conference on Web Engineering, 2013.
[16] R. Janbandhu, P. Dahiwale and M. Raghuwanshi, "Analysis of web crawling algorithms," International Journal on Recent and Innovation Trends in Computing and Communication, vol. 2, no. 3, pp. 488-492, 2014.
[17] T. Jo, "Text mining," Studies in Big Data, 2019.
[18] N. e. r. approaches, "Mansouri, Alireza; Affendey, Lilly Suriani; Mamat, Ali," International Journal of Computer Science and Network Security, vol. 8, no. 2, pp. 339-344, 2008.
[19] M. Taddicken, "The 'privacy paradox'in the social web: The impact of privacy concerns, individual characteristics, and the perceived social relevance on different forms of self-disclosure," Journal of Computer-Mediated Communication, vol. 19, no. 2, pp. 248-273, 2014.
[20] L. Scism and M. Maremont, "Insurers test data profiles to identify risky clients," The Wall Street Journal, vol. 19, 2010.
[21] D. Sharma, R. Shukla, A. K. Giri and S. Kumar, "A Brief Review on Search Engine Optimization," in 2019 9th International Conference on Cloud Computing, Data Science Engineering (Confluence), 2019.
[22] T. Tanaka, H. Niibori, S. Li, S. Nomura, H. Kawashima and K. Tsuda, "Bot Detection Model using User Agent and User Behavior for Web Log Analysis," Procedia Computer Science, vol. 176, pp. 1621-1625, 2020.
[23] S. Sulaiman and R. A. a. S. S. a. O. N. Wahid, "Using stanford NER and Illinois NER to detect malay named entity recognition," Int. J. Comput. Theory Eng, vol. 9, no. 2, pp. 147-150, 2017.
[24] C. M. Costa, G. Veiga, A. Sousa and S. Nunes, "Evaluation of Stanford NER for extraction of assembly information from instruction manuals," in 2017 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC), 2017.
[25] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit," in Association for Computational Linguistics (ACL) System Demonstrations, 2014.