VOL 4, NO.1 (2024) 65-85
https://jagst.utm.my

JAGST
Journal of Advanced Geospatial Science & Technology

The Spatial Distribution of Sexual Harassment Sentiment in Twitter

Nur Shafiqah Mohd Shariff[1], Suzanna Noor Azmy[1,2*], Zakri Tarmidi[1], Noradila Rusli[1,3]
[1]Faculty of Built Environment and Surveying, Universiti Teknologi Malaysia
[2]Geoscience and Digital Earth Centre (INSTEG), Universiti Teknologi Malaysia
[3]Centre for Innovative Planning and Development (CIPD), Universiti Teknologi Malaysia

*Corresponding author: suzanna.noorazmy@utm.my

Abstract - In a cultural landscape where discussions about sexual harassment are often considered taboo, this study leverages social media as a platform for the anonymous sharing of experiences related to sexual harassment in Asia. By focusing on Kuala Lumpur, we applied grid-based data mining to analyze the spatial distribution of tweets related to sexual harassment, revealing significant distribution patterns. Moreover, this study utilized two advanced sentiment analysis methods: the Natural Language Toolkit (NLTK) and Azure Machine Learning (AML), to evaluate the polarity of sentiments expressed in these tweets. The findings demonstrate a notable variance in sentiment analysis results between NLTK and AML, with NLTK classifying a majority of tweets as neutral (63.7%), while AML identified a predominant positive sentiment (70%). This discrepancy highlights the complexities of sentiment analysis and the importance of selecting appropriate tools for specific research contexts. In examining the spatial distribution, it becomes apparent that tweets collected through the identified keywords are dispersed within the boundaries of Kuala Lumpur. However, there is a notable concentration in specific areas, particularly evident in the hotspot encompassing Taman Desa and Pantai Dalam. Understanding the sentiment's location enables us to delve deeper into the pronounced cluster of tweets, examining potential correlations with surrounding social and moral issues. This insight empowers us to address the issue through various means, including the implementation of Crime Prevention through Environmental Design and fostering community-based initiatives.

Keywords *– sexual harassment; data mining; Twitter; geospatial; sentiment analysis; spatial pattern; GIS; social media*

Article History: Received 12 January 2024, Accepted 25 February 2024, Published 31 March 2024

How to cite: Shariff, N. S. M., Azmy, S.N., Tarmidi, Z. and Rusli, N. (2024). The Spatial Distribution of Sexual Harassment Sentiment in Twitter. Journal of Advanced Geospatial Science and Technology. 4(1), 65-85.

1.0 Introduction

Sexual harassment is a global phenomenon that can occur anywhere. Generally, sexual harassment can take place in three main settings: the workplace, public spaces, and educational institutions. Typically, sexual harassment happens to women, but it is not denied that men can also be victims of sexual harassment. According to Rahman and Pauzi (2019), victims describe the experience of being harassed as a tremendously distressing and emotionally oppressive experience. Victims face emotional problems such as depression, hatred, anger, fear, anxiety, loss of judgment, and not only on a mental level but also physically. Sexual harassment is a serious issue in Malaysia. A survey conducted by the National Human Rights Commission of Malaysia in 2019 found that 85% of women and 44% of men have experienced some form of sexual harassment. The causes of sexual harassment in Malaysia are complex and varied. However, some key factors include gender inequality. Malaysia is a patriarchal society where men are traditionally seen as superior to women. This can lead to a culture of male entitlement, manifested in the form of sexual harassment and a lack of awareness. Many people in Malaysia are not familiar with what constitutes as sexual harassment. This can make it difficult for the victims to recognize and report harassment. There are still many negative attitudes towards women in Malaysia. According to several reports, official data or reports regarding sexual harassment in Malaysia do not accurately reflect the actual situation (Women's Aid Organisation, 2017; Faruqi, 2020 and Hadzam, 2021). A recent study by the Women's Aid Organisation (WAO) found that 8 out of 10 women have experienced some form of sexual harassment in their lifetime (Baker and McKenzie, 2023). They are also more likely to have low self-esteem and difficulties in forming relationships. Instead of making official reports, victims tend to feel more comfortable expressing their experiences of sexual harassment through social media platforms (Malik, 2018; Budiman et al., 2020; Zakaria et al., 2019). Therefore, this study aims to examine the intensity and spatial distribution of sexual harassment sentiments discussed on Twitter within the locality of Kuala Lumpur. Apart from measuring the intensity of sexual harassment sentiment as medium of sexual harassment expose, this study also emphasized the identification of the keywords describing sexual harassment in tweets, the location, the polarity of tweet sentiment and the pattern distribution.

Meanwhile, social media data mining refers to the process of extracting and analyzing information or patterns from large volumes of data generated on social media platforms. It involves collecting, processing, and interpreting data from various social media sources to derive valuable insights, trends, or patterns (Rusli et al., 2023). The increasing reliance on social networks calls for data mining techniques that is likely to facilitate reforming the unstructured data and place them within a systematic pattern (Injadat et al., 2016). Social data mining for sexual harassment focuses on using data mining techniques to understand and address sexual harassment tendencies on social media platforms. Studies have shown that sexual harassment on these platforms primarily targets young women and involves unwanted sexual attention.

Kumar et al. (2011) conducted a study utilizing machine learning and natural language processing techniques to analyze trends in online sexual harassment. Their research focused on the detection and classification of different types of online harassment, aiming to automate the identification and removal of such content. This work highlighted the importance of leveraging technology to address the complexities of online harassment and improve detection processes

In a similar vein, researchers have delved into the spatial analysis of tweets concerning sexual harassment, aiming to uncover patterns related to tweet distribution and population density. By utilizing web crawlers for data collection, these studies have been able to pinpoint areas with high concentrations of discussions on sexual victimization. This analytical approach has provided valuable insights into the geographic distribution of tweets related to movements like #MeToo, shedding light on potential implications for resource allocation and support systems (Gautam and Bansal, 2022).

2.0 Materials and Methods

2.1 Study Area

This study aims to examine the spatial distribution of sexual harassment sentiments discussed on Twitter within the locality of Kuala Lumpur (Figure 1) and map the hotspots of sexual harassment sentiments in the study area. Data is collected through a data mining approach on the Twitter social media platform within the selected study area of Kuala Lumpur. Kuala Lumpur was chosen as the study area due to its high record of reported sexual harassment cases, as shown in Figure 2. Additionally, sexual harassment is extensively discussed on social media platforms, as millennial generations are less likely to report to the police for various reasons. Therefore, the data mining method on Twitter has been utilized to gather data pertaining to sexual harassment in the study area. In terms of temporal aspects, the data has been mined within the time period from 1 January 2013 to 31 December 2023.

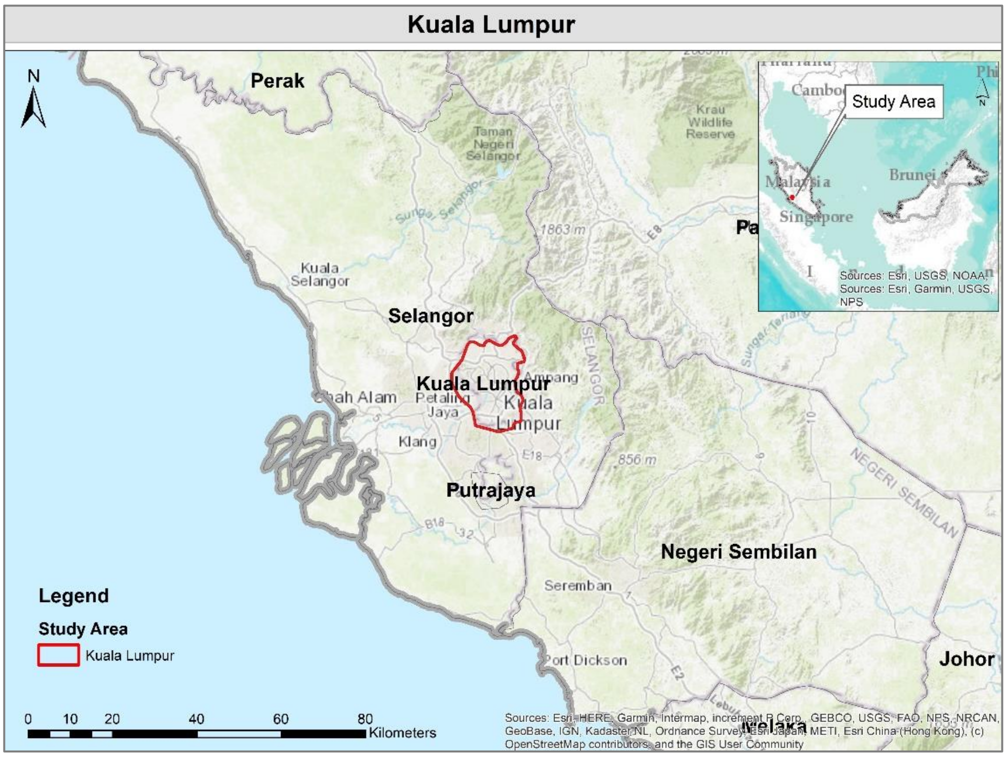


Figure 1: Study area - Kuala Lumpur City

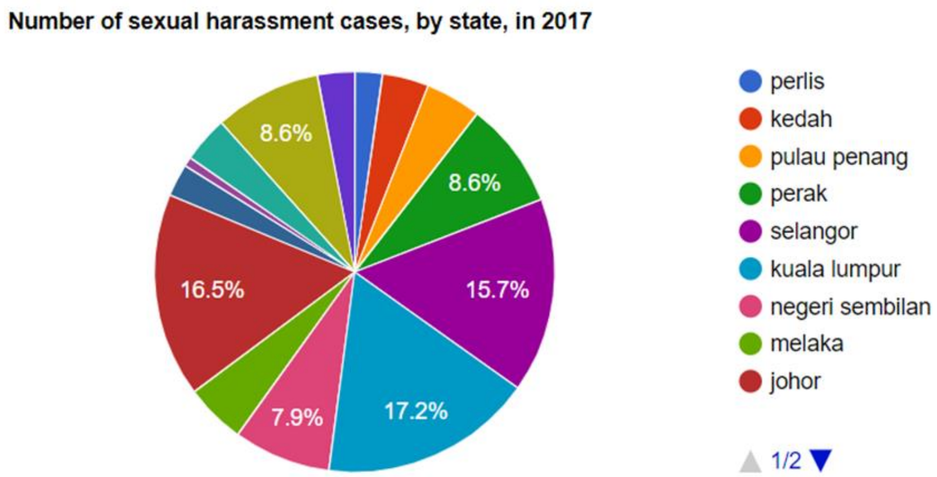


Figure 2: Number of sexual harassments cases

Source: Kamarudin (2022) dan Women's Aid Organisation (WAO), (2017)

2.2 Methodology

The methodology of this study consists of four phases: identify the issue, data collection, data processing, data analysis, and results and discussion (see Figure 3). The first phase defines the problem related to the study. The second phase, data collection. It discusses the keywords commonly used on the Twitter social media platform. The third phase is data processing, which involves data filtering. The final phase is the generation of data analysis and visualization of results.

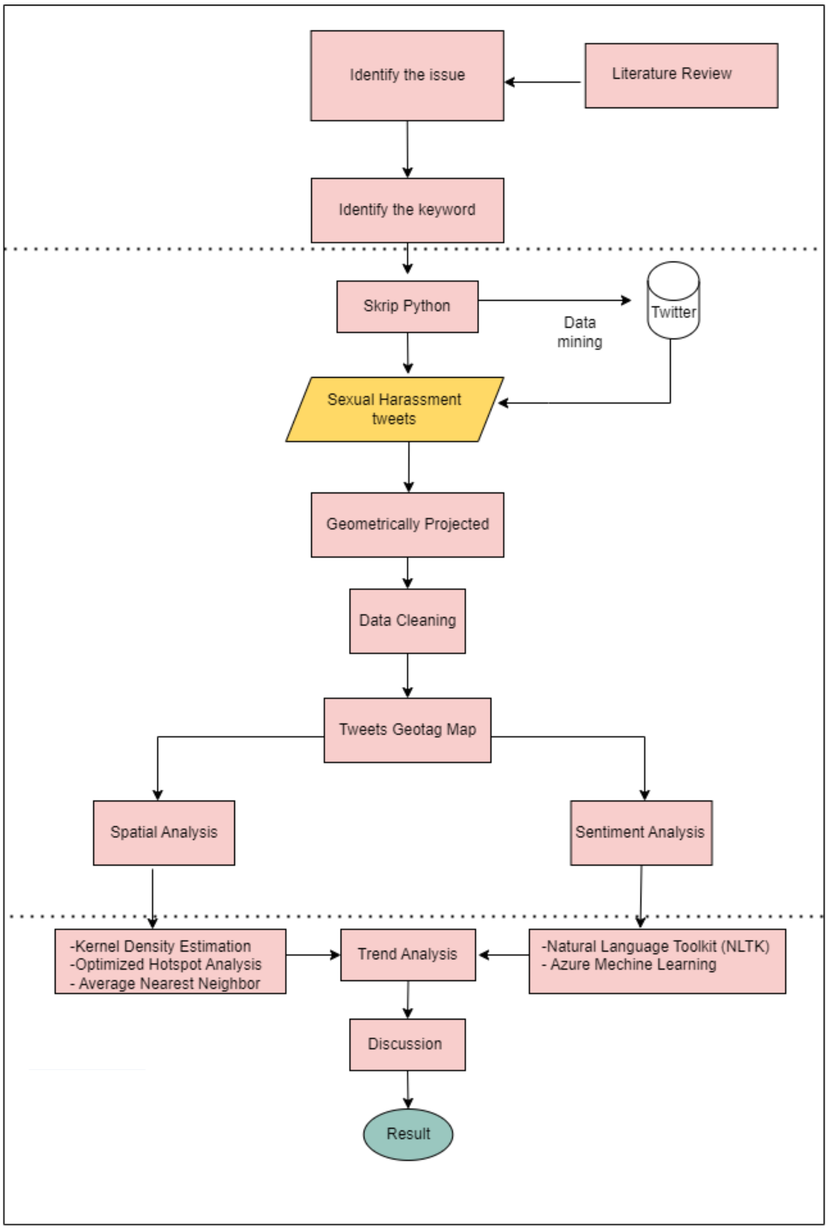


Figure 3: The flow of research methodology

2.3 Data Collection

In this phase, the data collection process is carried out using the data scraping method by developing phyton script incorporating snscrape module. Twitter platform was selected due to Twitter's architecture facilitates the rapid dissemination and discussion of information, making it an ideal platform for capturing real-time public sentiment on issues like sexual harassment. Unlike some other platforms where data might be more private or harder to access due to privacy settings, Twitter offers a vast amount of publicly available data, enabling researchers to gather and analyze discussions on sexual harassment extensively. The inherent character limit on Twitter encourages users to express their opinions succinctly, often making it easier to analyze sentiments compared to longer, more nuanced posts that might be found on other platforms. In addition, Twitter has a wide-reaching global user base, including significant usage in Kuala Lumpur, the study area. This diversity offers a broad perspective on the subject matter.

There are twelve keywords namely "Mengorat", "Seksual," "Memegang," "Liwat," "Dicium," "Rogol," "Dirty jokes," "Raba," "Catcalling," "Bercumbu", "Cabul" and "Gangguan seksual" are used to scrape data from the Twitter. The selection of keywords for this study was strategically made to align closely with the prevalent discourse on sexual harassment, particularly as it is discussed within social media contexts. This selection process was guided by the dual criteria of relevance to sexual harassment scenarios and their frequency of use in online conversations. This methodological choice guaranteed the relevance of the collected data to the study's objectives. Notably, as illustrated in Figure 2, Kuala Lumpur registers the highest incidence of sexual harassment cases, which underscores the importance of these specific terms. By choosing keywords that are commonly associated with real-world reports and incidents of sexual harassment in Kuala Lumpur, this study ensures that the language used in the data collection mirrors the local expressions and experiences, thus accurately reflecting the regional specificity of such cases. The data is collected from 1 January 2013 to 31 December 2023, as shown in Table 1.

Table 1: Data from Twitter

Data	Criteria	Number of Data
Extracted	Sexual harassment keyword	7614 tweets
Removed	Non-related to sexual harassment, spam, non-English and Malay tweets, non-coordinate and geotagged tweets, incorrect coordinates	4956 tweets
Data used for processing	Coordinate in Kuala Lumpur, correct keyword for sexual harassment	2658 tweets

2.4 Data Processing

The Jupyter (Anaconda) platform (Figure 4) is used to extract data from social media platform Twitter using data mining techniques with keyword-based searches. Jupyter Notebook is also utilized to perform sentiment analysis in order to classify the obtained tweet texts as positive, negative, or neutral. Jupyter Notebook is a web application that enables the creation and sharing of documents containing live code, equations, visualizations, and narrative text.

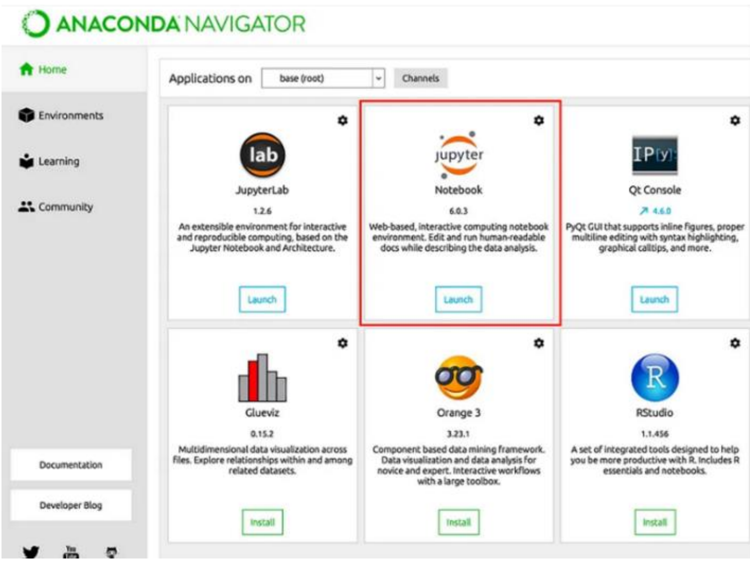


Figure 4: Jupyter Platform

To provide a visual representation of the sentiment analysis location, we utilized the ArcGIS software. We discovered a way to depict the data mined from Twitter. Using GIS tools, we mapped the latitude and longitude information of each tweet and displayed them as points on a map. ArcGIS is a popular GIS software used for creating and displaying maps, as well as analyzing and managing spatial data. Visualizing the data using ArcGIS aids in understanding trends and patterns within the data (Rusli et al., 2023).

During the data processing stage, it involves the process of data filtering as not all data can be considered usable due to not meeting the established criteria to ensure the quality of the research outcomes. There are several reasons why data filtering may be necessary in the study of sexual harassment on Twitter. For instance, tweets that are not related to sexual harassment need to be excluded, and spam tweets or tweets in languages other than English and Malay should be discarded. Data filtering is also performed to remove non-coordinate tweets, non-geotagged tweets, and incorrect coordinates resulting from phone settings or the location sharing policies of the tweet owners.

2.5 Spatial Distribution Analysis

The research aims to analyze the spatial distribution of sentiment, investigating how sentiment varies across different geographical locations. By integrating sentiment analysis techniques with spatial data, the study seeks to uncover patterns and trends in sentiment within specific regions or areas. Additionally, an analysis applied to the data is Exploratory Spatial Data Analysis (ESDA), which comprises a set of statistical techniques used to explore the spatial distribution of data. ESDA techniques can be employed to identify patterns and trends in spatial data, test hypotheses regarding spatial relationships, and visualize spatial data. Some specific analyses conducted include optimized hotspot analysis, kernel density estimation, and nearest neighbour analysis.

2.6 Sentiment Analysis

Sentiment analysis seeks to determine the contextual polarity of text based on the author's declared views on a subject or message. The fundamentals of sentiment categorization and evaluation are thoroughly documented by Pang and Lee (2008), Sharma and Jain (2020). Natural language processing methods are used to analyse tweets to find and extract pertinent information on sexual harassment in this study's sentiment analysis of sexual harassment using Twitter data mining. This entails spotting potentially disruptive speech and actions as well as patterns and trends in the data that might point to more serious problems.

The tweets were auto-classified into positive, negative, and neutral sentiment using the "Natural Language Toolkit (NLTK)" and Azure Machine Learning. The Natural Language Toolkit (NLTK), which provides a variety of text processing libraries and corpora for tasks including text categorization, information extraction, machine translation, and more, is a useful tool for researchers, students, and developers. Contrarily, Azure Machine Learning has strong capabilities for sentiment analysis, enabling us to ascertain the sentiment or emotion portrayed in text data like customer reviews, social media posts, or survey replies (Bandi & Fellah n.d.; Harfoushi, Hasan & Obiedat 2018).

As referred to Figure 5, The process commences with the initialization phase, where essential libraries are imported to facilitate data manipulation and analysis. These include the 'CSV' library for handling CSV file operations, 'matplotlib.pyplot' for plotting graphs, 'nltk.sentiment' for sentiment analysis, and 'langid' for language identification.
Subsequently, the operational parameters are defined. This entails specifying the input CSV file containing the Twitter data and establishing an output CSV file to store the results of the sentiment analysis. The next phase involves the initialization of empty lists, namely 'tweet_texts', 'themes', 'longitudes', 'latitudes', and 'sentiment_labels'. These lists are intended to store processed data such as the text of tweets, thematic categorizations, geographic coordinates, and sentiment classifications, respectively.

Following this setup, the algorithm enters a loop to iterate over each row in the CSV file, which represents individual tweets. During each iteration, the sentiment of the tweet is analyzed and the result is recorded. Once all rows have been processed, the collected data on sentiment is utilized to print the results, providing immediate insights into the analyzed sentiments. In the final step, a pie chart is generated to visually represent the distribution of sentiment scores, thus concluding the data analysis process. This visual representation allows for an at-a-glance understanding of sentiment proportions within the dataset.

The procedure effectively terminates after the pie chart is produced, signaling the end of the sentiment analysis workflow. This structured approach ensures systematic data processing, from raw input through to the final visual output, thereby facilitating a comprehensive analysis of sentiment data derived from Twitter.

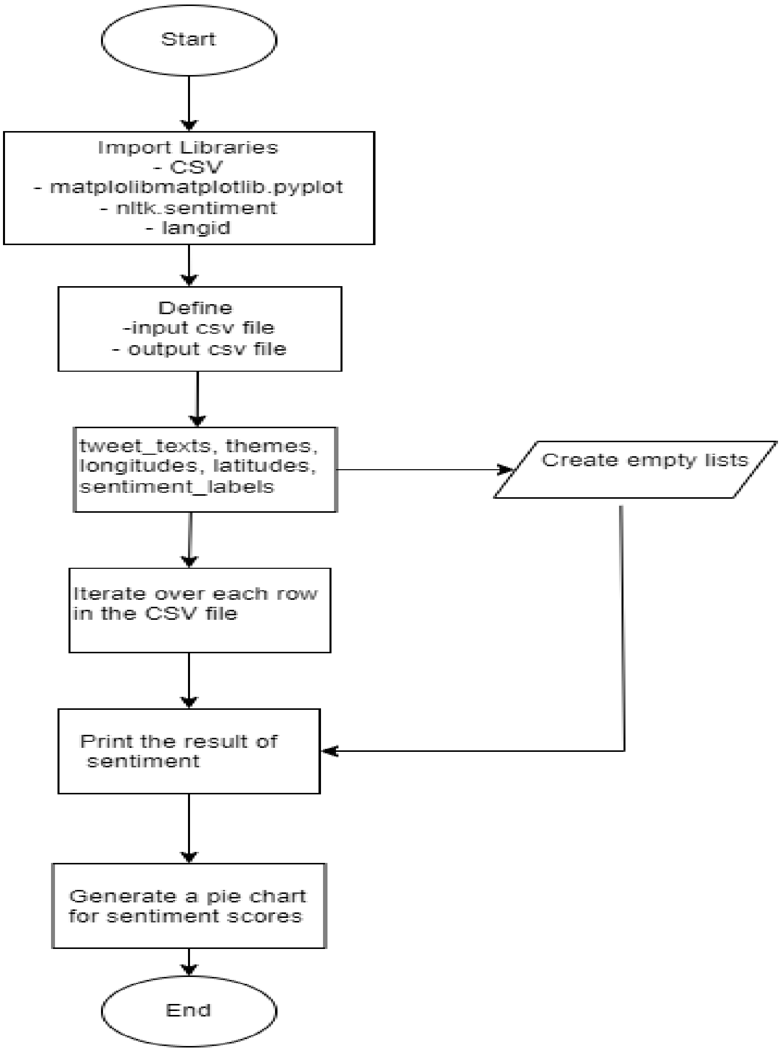


Figure 5: Pseudocode flow diagram

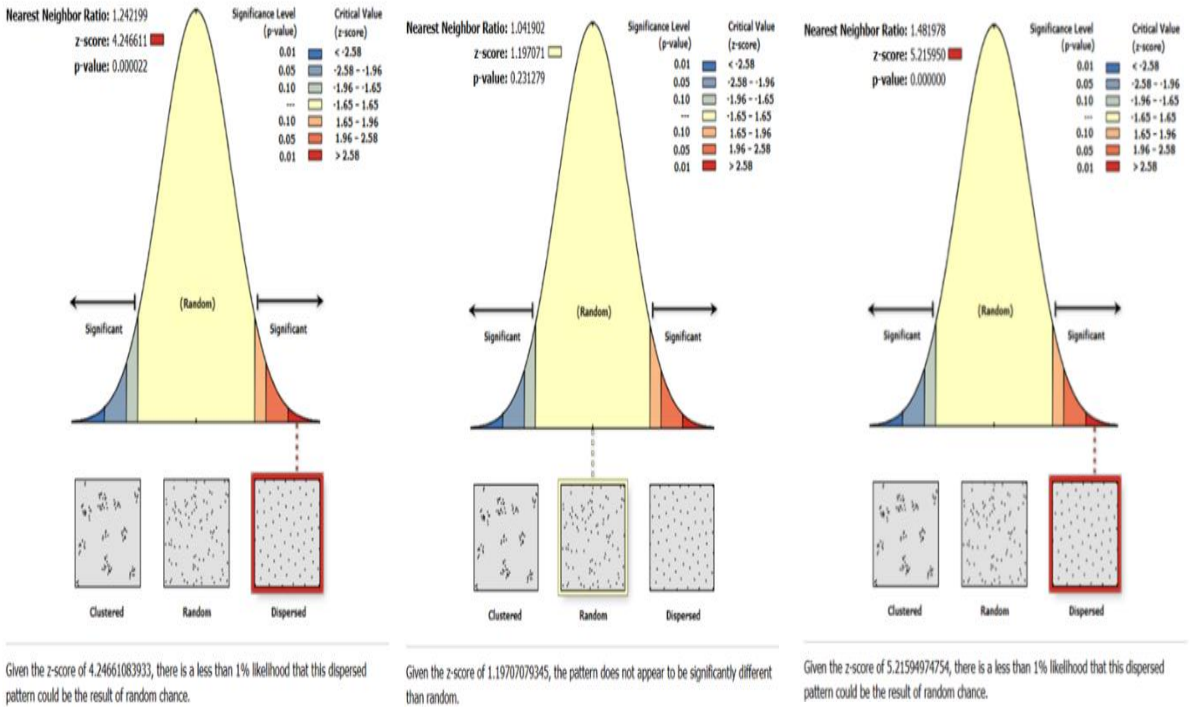
3.0 Results and Discussion

3.1 Spatial Patterns

To fulfil the third objective of analyzing the spatial distribution patterns in Kuala Lumpur, the average nearest neighbour method was employed to determine whether the patterns were clustered, random or dispersed. In Figure 6 (a) shows the pattern observed for positive sentiment was classified as dispersed with a z-score of 4.24661083933. This z-score indicates that there is less than a 1% chance that the observed dispersed pattern could be attributed to random chance alone. Furthermore, the nearest neighbour ratio of 1.242199 further supports the dispersed distribution pattern.

Figure 6 (b) the spatial pattern distribution of neutral sentiment, which appears to be "Random" with a Z-score of 1.19707079345. This indicates that the observed pattern does not deviate significantly from what would be expected in a random distribution. Additionally, the p-value of 0.231279 is relatively high. Higher p-values suggest that the observed pattern is not statistically significant, indicating that the neutral sentiment pattern is like random chance. Figure 6 (c) depicts the pattern analysis for negative sentiment, which is "dispersed" with a z-score of 5.21594974754. This indicates that the likelihood of this dispersed pattern occurring randomly is less than 1%. Locations with negative sentiment exhibit a dispersed pattern, with a low probability of occurring randomly. The combination of the z-score, nearest neighbour ratio, and p-value provides strong evidence supporting this spatial pattern.

In summary, according the analysis the different spatial patterns for each sentiment category. Both positive and negative sentiments show a dispersed pattern with high statistical significance, indicating a non-random distribution. On the other hand, neutral sentiment shows a random pattern. Positive and negative sentiments are more likely to be expressed in a strong and emotional way, which can lead to them being clustered in certain areas. Neutral sentiment, on the other hand, is more likely to be expressed in a more neutral and objective way, which can lead to it being more evenly distributed.



a) Positive Sentiment b) Neutral Sentiment c) Negative Sentiment

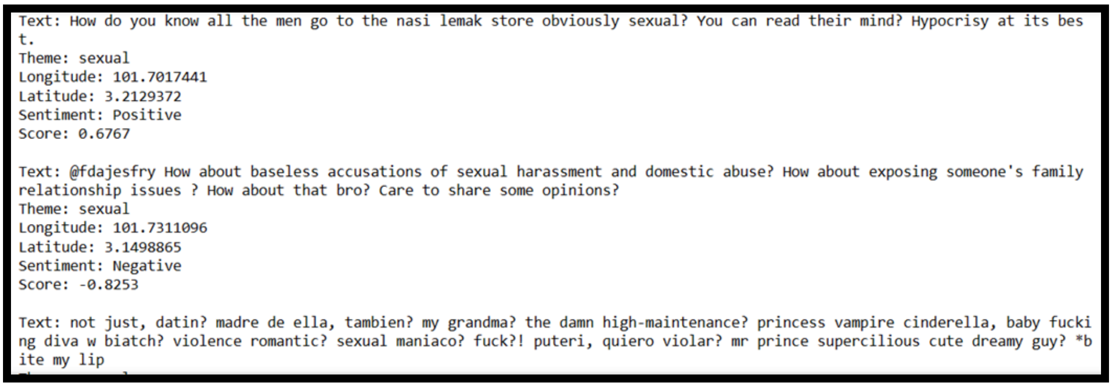
Figure 6: (a) Spatial Pattern of Positive Sentiment (b) Neutral Sentiment (c) Negative Sentiment

Meanwhile, Figure 7 presents a word cloud depicting sexual harassment in the study area. Words that appear larger in the word cloud are those that are more frequently used in tweets about sexual harassment. The words in the word cloud are arranged in a way that reflects their relative frequency.

Figure 7: word cloud on Twitter after analysis

3.2 Analysing Sentiment in Tweets Using the Natural Language Toolkit (NLTK) and Azure Machine Learning Method

Figure 8 presents the sentiment analysis conducted using the Natural Language Toolkit (NLTK) and Figure 10 is the sentiment analysis data using Azure Machine Learning. On the other hand, Figure 9 illustrates the percentage distribution of sentiment analysis results obtained through NLTK. According to the pie chart, it can be observed that the majority of analyzed tweets were classified as having a neutral sentiment, accounting for approximately 63.7% of the total. Following this, positive tweets make up around 22.2%. Finally, approximately 14.1% of the analyzed tweets have been categorized as having a negative sentiment.



```
Text: How do you know all the men go to the nasi lemak store obviously sexual? You can read their mind? Hypocrisy at its bes
t.
Theme: sexual
Longitude: 101.7017441
Latitude: 3.2129372
Sentiment: Positive
Score: 0.6767

Text: @fdajesfry How about baseless accusations of sexual harassment and domestic abuse? How about exposing someone's family
relationship issues ? How about that bro? Care to share some opinions?
Theme: sexual
Longitude: 101.7311096
Latitude: 3.1498865
Sentiment: Negative
Score: -0.8253

Text: not just, datin? madre de ella, tambien? my grandma? the damn high-maintenance? princess vampire cinderella, baby fucki
ng diva w biatch? violence romantic? sexual maniaco? fuck?! puteri, quiero violar? mr prince supercilious cute dreamy guy? *b
ite my lip
```

Figure 8: Results of sentiment analysis using NLTK.

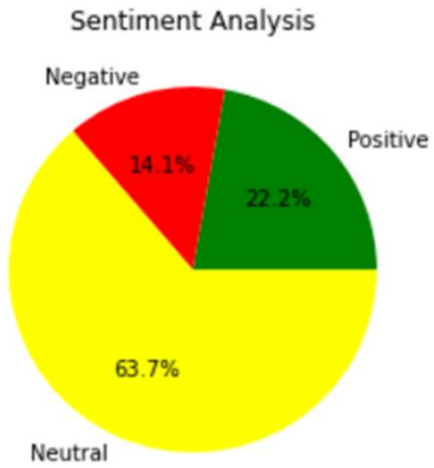


Figure 9: The percentage of sentiment analysis using NLTK

	A	B	C	D	E
1	tweet_text	Longitude	Latitude	Sentiment	Score
2	How do you+A2:A81 know all the men go to the nasi lemak store obviously sexua	101.702	3.21294	positive	90%
3	@fdajesfry How about baseless accusations of sexual harassment and domestic a	101.731	3.14989	positive	94%
4	not just, datin? madre de ella, tambien? my grandma? the damn high-maintenan	101.681	3.17575	negative	4%
5	@taehyuvv Jantan yang acah2 begini kena takot sikit sebab depa senyap2 ada ciri	101.744	3.17547	positive	90%
6	@hafizhudri Bro ni student mana? Sexual harassment ni	101.702	3.21294	positive	62%
7	@1Obefiend Yes.ive been single a long time since divorced. I took medical check-u	101.702	3.21294	negative	1%
8	billboards, iklan-iklan produk "kuat", tak pernah disuarakan pun. Dahlah iklan-	101.702	3.21294	positive	61%
9	My close friends stories can be sexual, can be something about cat and can be any	101.731	3.14989	negative	36%
10	Aku bukan anti-vax tapi benda2 macam ni aku consider personal boundaries and	101.731	3.14989	positive	64%
11	This is pure sexual harassment.	101.702	3.21294	neutral	54%
12	I am struggling when I say spread love not hate because the darkest moment in my life was from sexual harassment by a male who supposed to be family. Trust issues was developed when I was 7 years old.	101.731	3.14989	negative	0%
13	Carnal knowledge - The term derives from carnal, meaning "of the flesh", and the	101.731	3.14989	positive	79%
14	@swingingb2b I'm not sure. Heard that sessions were conducted in zoom, and ea	101.702	3.21294	negative	9%
15	I've heard that a club's virtual booth was sabotaged with explicit sexual material during freshmen orientation week (TAWE).	101.702	3.21294	negative	23%
16	@hanafy_mj Hurmmmmm entah le nak komen.....unless ada unsur sexual, lucah (	101.731	3.14989	positive	63%
17	reading a lot of sexual harassment tweets regarding Polis officers at Jalan Duta Tc	101.731	3.14989	neutral	58%
18	harassment !	101.702	3.21294	neutral	48%
19	Season 17 is definitely giving a nod to heavy sexual themes in the earlier seasons.	101.731	3.14989	positive	80%
20	believe the sexual assault victim first, THEN investigate. don't call someone a liar j	101.731	3.14989	negative	17%
21	@Harishna6 the sexual energy between me and burning my physics book	101.731	3.14989	neutral	54%
22	So blackpink broke the 24 record then bts released the following week and broke	101.731	3.14989	negative	15%
23	Some Indonesians finna make this sexual	101.702	3.21294	positive	71%
24	Belum lagi lawyers who have longgggggggg history of sexual harassment cases being DEFENDED and protected because they are superstars who bring in	101.702	3.21294	negative	38%
25	These issues of rape culture, sexual violence and predators using socmed to groor	101.702	3.21294	neutral	52%
26	@Hidayat_Manaf being LGBTQ+ is a sexual orientation. I'm just saying, using YOL	101.731	3.14989	negative	40%

Sentiment_AML

Figure 10: The sentiment analysis of data using Azure Machine Learning.

Figure 11 presents a pie chart depicting the sentiment distribution of tweets related to sexual harassment. The chart reveals that most of these tweets, accounting for 70%, exhibit positive sentiment. The next prominent category is neutral tweets, making up 20% of the total. Lastly, 10% of the tweets analyzed demonstrate negative sentiment.

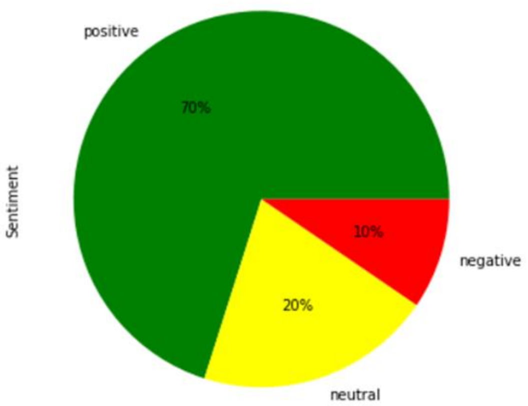


Figure 11: The percentage of sentiment analysis using AML

Here, we can see how the sentiment analysis findings produced by the Azure Machine Learning platform and the Natural Language Toolkit (NLTK) are significantly dissimilar. The diverse sentiment analysis outcomes between these two platforms demonstrate how various training sets, algorithms, and approaches can result in various sentiment categorization findings. Considering these factors, it is important to understand that sentiment analysis is a difficult endeavor, and that different tools or platforms may produce varying results due to differences in their underlying methodologies and settings. As a result, it is crucial to carefully choose the appropriate tool based on the investigation's particular needs and environment.

3.3 Heat Map to Optimized Hotspot Analysis

The heat map (Figure 12) use count incidents within hexagon polygon to analyze the area of heat map. Pantai Dalam is a sizable region that includes the city's core and other neighboring suburbs. Since neutral sentiment has 10 output features that are statistically significant based on an FDF adjustment for multiple testing and spatial dependence, a comparison of the heat map data shows that Pantai Dalam has a higher concentration of neutral sentiment than positive or negative sentiment. This indicates that residents of Pantai Dalam are more likely to express neutral opinions regarding sexual harassment than they are to express opinions that are favourable or unfavourable. According to an FDF adjustment for multiple testing and spatial dependence, the positive sentiment heat map has 6 statistically significant output features, while the negative sentiment heat map only has one such feature.

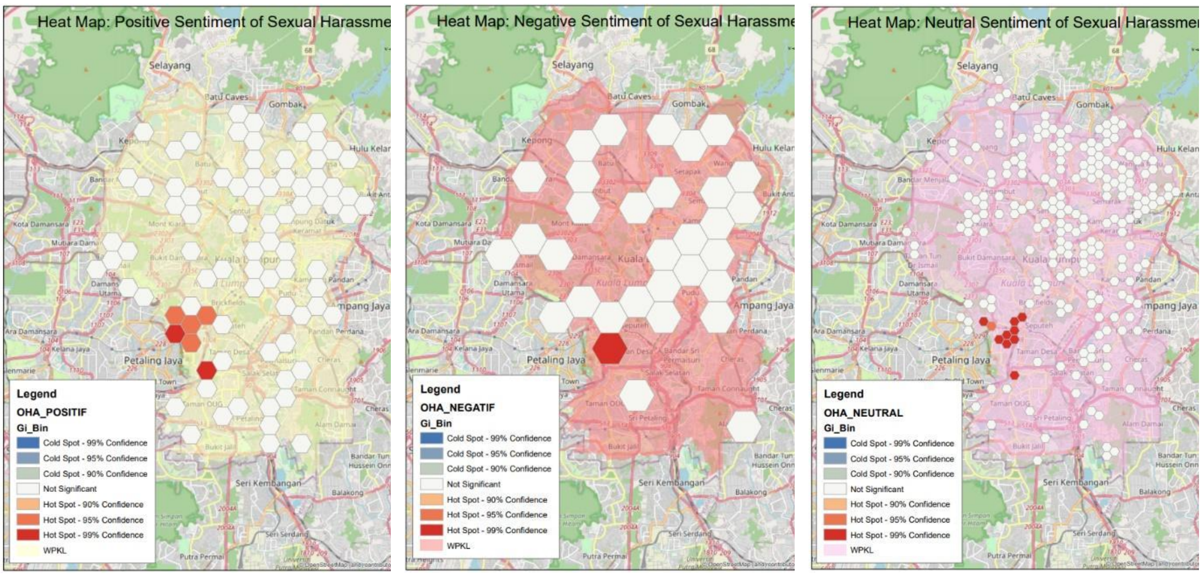


Figure 12: Heat map

3.4 Hot Spot Map using Kernel Density Analysis

The Kernel Density Estimation (KDE) analysis, as depicted in Figure 13, serves as an appropriate tool for pinpointing areas with a high concentration of each sentiment type. Kalinic and Krisp (2018) describe KDE as a statistical method that approximates the probability density function of a random variable. Within the realm of hotspot analysis, this technique is adept at identifying clusters of data points, revealing geographic locations where expressions of specific sentiments are notably prevalent. This allows for a nuanced understanding of spatial patterns in sentiment distribution.

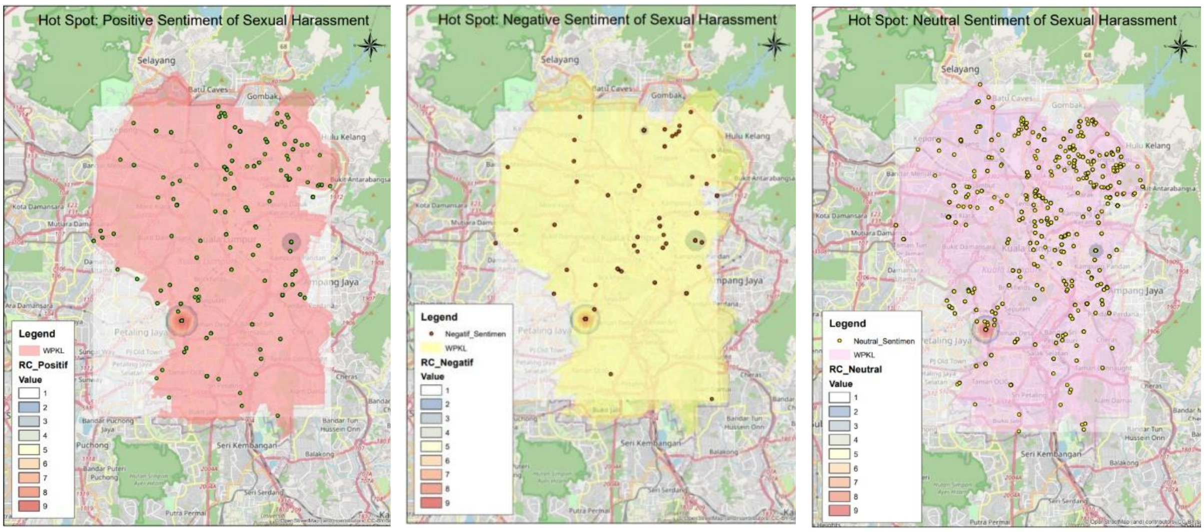


Figure 13: Hotspot Map

We found that positive and neutral sentiment have the same hotspots and coldspot locations when comparing hotspots and cold spots of positive, negative, and neutral sentiment. Positive and neutral feeling, however, share the same hotspot sites, while negative sentiment has separate cold spot locations. This shows that while positive and neutral feeling are both present in high concentrations, negative sentiment is not as prevalent. Similarity in the expression of sentiment is indicated by the presence of the same hotspots for positive, neutral, and negative sentiment. Different cold spot locations, however, suggest that negative sentiment occurs less frequently than positive and neutral sentiment. These results highlight the geographical dynamics of sentiment dispersion and can help us comprehend the sentiment environment in a certain area. The underlying causes of these patterns can be ascertained with the aid of additional analysis and consideration of contextual factors, which can then guide focused tactics or interventions.

3.5 Discussion

The methodology articulated in this research offers an insightful exploration into the spatial manifestation of sentiments related to sexual harassment in Kuala Lumpur, derived from social media data mining. By concentrating on keywords prevalent in Twitter discourse on sexual harassment, the study has illuminated the distribution of spatial patterns tied to varied sentiments.

Our investigation has uncovered distinctive spatial patterns linked to types of sentiment, revealing a significant dispersion of positive and negative sentiments. This suggests that expressions of strong emotions are not random but are concentrated in certain locales, reflecting potential hotspots of intense discourse on sexual harassment. Conversely, neutral sentiments are more diffusely distributed, indicating a more measured engagement with the topic across the urban fabric.

The analytic diversity presented by Azure Machine Learning and the Natural Language Toolkit (NLTK) underscores the complex nature of sentiment analysis. The variance in findings between these platforms underscores the critical need for careful selection of analytic tools, which may yield divergent outcomes based on their specific algorithms, training sets, and configurations. Such variation necessitates a cautious approach in interpreting sentiment analysis results, as different tools offer unique lenses through which to view the data.

Beyond the analysis of sentiment dispersion, it is vital to contextualize these patterns within the broader sociocultural and gender dynamics of Kuala Lumpur. The isolated study of sentiment distribution provides only a snapshot, lacking in the depth needed to grasp the root causes and dynamics of sexual harassment. To this end, our spatial analysis results should be synthesized with findings from prior research that delve into the contributory factors of sexual harassment, such as cultural norms, gender dynamics, socio-economic factors, and organizational structures.

Moreover, reflecting on the accuracy and reliability of our sentiment analysis, we acknowledge the methodological confines and potential biases that might influence data interpretation. To enhance confidence in our findings, we propose verifying the sentiment classification through cross-referencing reported cases or supplemental studies involving direct feedback from affected individuals.

While our study contributes valuable insights into the discourse around sexual harassment on Twitter in Kuala Lumpur, it also paves the way for future research directions. These include, but are not limited to, multi-platform analyses, longitudinal studies to track sentiment evolution, and incorporation of these findings into policies and interventions aimed at mitigating sexual harassment. It is through such layered, multifaceted inquiry that we can aspire to construct a more complete and actionable understanding of sexual harassment within urban environments.

4.0 Conclusion

In conclusion, the present study methodically investigated the spatial distribution of sentiments pertaining to sexual harassment as expressed on Twitter within the Kuala Lumpur area. Recognizing the rise of social media as a critical medium through which the millennial generation communicates and processes their experiences, this research leveraged data mining techniques to discern the geographic nuances of these discussions in urban public spaces. The employment of specific, locally-relevant keywords allowed for the extraction and subsequent analysis of tweets, leading to the identification of distinct spatial sentiment patterns.

This exploration has yielded a map of sentiment hotspots, offering unprecedented insight into where conversations about sexual harassment are most concentrated within Kuala Lumpur. Such spatial intelligence is invaluable, not only for understanding the pervasiveness of the issue but also for informing potential interventions by policymakers and social workers. The analytical power of ArcGIS, combined with the versatility of the Jupyter Notebook (Anaconda) environment, proved instrumental in visualizing and interpreting the complex web of data, thereby enhancing our comprehension of the relationship between place and sentiment in the context of sexual harassment.

Ultimately, this study not only delineates the landscape of sentiment associated with sexual harassment on a popular digital platform but also signifies a step forward in harnessing the potential of social media analytics to inform and potentially transform the societal response to such critical social issues.

Acknowledgements

Authors would like to express our gratitude to the Ministry of Higher Education (MOHE) Malaysia for funding this project under the Fundamental Research Grant Scheme (FRGS/1/2023/WAB11/UTM/02/1).

References

Alawneh, Eiad, Muhannad Al-Fawa’reh, Mohammad T. Jafar, and Mohannad al Fayoumi. 2021. "Sentiment Analysis-Based Sexual Harassment Detection Using Machine Learning Techniques." In 2021 International Symposium on Electronics and Smart Devices (ISESD), 1–6. IEEE.

Awang, F., R.K.R.M. Khalid, and A.M. Zam. 2020. "Pemahaman dan Pengalaman Gangguan Seksual Dalam Kalangan Masyarakat Malaysia." International Journal of Social Policy and Society 16.

Aziz, A.R.A., et al. 2021. "Persepsi Pekerja Terhadap Sensitiviti Gangguan Seksual dan Hubungannya Dengan Tingkah Laku Devian." e-BANGI 18, no. 5: 212–229.

Azman, F., et al. 2018. "My Guardian: A Personal Safety Mobile Application." In 2018 IEEE Conference on Open Systems (ICOS), 37–41. IEEE.

Baker and McKenzie. 2023. "Malaysia: Anti-Sexual Harassment Act 2022." Employment Notes. Accessed January 19, 2023. https://hsfnotes.com/employment/2023/12/07/malaysia-anti-sexual-harassment-act-2022/.

Basir, M.K.B.A. "Bentuk-Bentuk Keganasan Terhadap Wanita di Tempat Awam: Kajian di Kuala Lumpur."

El-Gibaly, O., and E.-T.R. 2017. "Public Transportation: A Hot Spot for Sexual Harassment of Women." Accessed January 19, 2023.

Faruqi, I. 2020. "Jenayah Seksual Perlu Kesedaran Tuntas." Accessed January 19, 2023. https://www.bernama.com/bm/tintaminda/news.php?id=187215

Garfias Royo, M., P. Parikh, and J. Belur. 2020. "Using Heat Maps to Identify Areas Prone to Violence Against Women in the Public Sphere." Crime Science 9, no. 1: 1–15.

Gautam, Arvind Kumar, and Abhishek Bansal. 2022. "A Review on Cyberstalking Detection Using Machine Learning Techniques: Current Trends and Future Direction." *International Journal of Engineering Trends and Technology* 70, no. 3: 95-107. https://doi.org/10.14445/22315381/IJETT-V70I3P211.

Hand, D., M.H., and S.P. 2001. Principles of Data Mining. Cambridge, MA: MIT Press.

Hishamuddin, M.S., Lekha L., H.H. Low, and M.S. Maisarah. 2003. "A Study on Sexual Harassment in Small and Medium Enterprises of Malaysia." Universiti Teknologi Malaysia.

IPSOS. 2022. "Gangguan Seksual Merupakan Kebimbangan Utama Rakyat Malaysia." AstroAwaniNetworkSdn.Bhd. Accessed January 19, 2023. https://www.astroawani.com/berita-malaysia/gangguan-seksual-merupakan-kebimbangan-utama-rakyat-malaysia-351028.

Isa, K., and S. Samad. 2012. "Tahap Kesedaran Pekerja Hubungannya dengan Salah Laku Gangguan Seksual di Tempat Kerja." Journal of Techno-Social 4, no. 2.

Ishar, M.I.M., et al. 2021. "Amalan Tidak Berintegriti: Remeh Tapi Salah." Malaysian Journal of Social Sciences and Humanities (MJSSH) 6, no. 8: 215–223.

Kamarudin, L. 2022. "KL, Selangor, Johor Catat Kes Gangguan Seksual Tertinggi." SinarHarian. Accessed January 19, 2023. https://www.sinarharian.com.my/article/206082/BERITA/Nasional/Kes-seksual-tinggi-di-KL-Selangor-Johor.

Karami, A., S. Swan, and M.F. Moraes. 2020. "Space Identification of Sexual Harassment Reports with Text Mining." In Proceedings of the Association for Information Science and Technology, 57, no. 1: e265.

Khalid, R. 2021. "Kes Gangguan Seksual Meningkat." Accessed January 19, 2023. https://www.jpw.gov.my/index.php/ms/media-ms/keratan-akhbar/1480-kes-gangguan-seksual-meningkat.

Kisiel, M.A., et al. 2020. "Medical Students' Self-Reported Gender Discrimination and Sexual Harassment over Time." BMC Medical Education 20, no. 1: 1–9.

Mad Shah, I., et al. "Gangguan Seksual di Tempat Kerja: Kajian Ke Atas Mangsa Dan Pelaku di Sekitar Johor Bahru, Johor."

Min, Y.H., G. Byun, and M. Ha. 2022. "Young Women's Site-Specific Fear of Crime within Urban Public Spaces in Seoul." Journal of Asian Architecture and Building Engineering 21, no. 3: 1149–1160.

Misriyanti, M., and R.Y. Primanita. 2022. "Hubungan Self-Acceptance dan Forgiveness pada Wanita yang Mengalami Pelecehan Seksual di Tempat Kerja." Ranah Research: Journal of Multidisciplinary Research and Development 5, no. 1: 139–144.

Oni, H., and T. Tshitangano. 2019. "Sexual Harassment and Victimization of Students: A Case Study of a Higher Education Institution in South Africa." African Health Sciences 19, no. 1: 1478–1485.

Prathap, B.R., and K. Ramesha. 2018. "Twitter Sentiment for Analysing Different Types of Crimes." In 2018 International Conference on Communication, Computing and Internet of Things (IC3IoT), 483–488. IEEE.

Rahman, S., and S.F.M. Pauzi. 2019. "Gangguan Seksual di Tempat Kerja." Gading Journal for Social Sciences 22: 124–129.

Reel, J.J., and E. Crouch. 2019. "#MeToo: Uncovering Sexual Harassment and Assault in Sport." Journal of Clinical Sport Psychology 13, no. 2: 177–179.

Rusli, N., Nordin, N.Z., Ak Matusin, A.M.R. ,Yusof, J.N., Rosley, M.S.F., Ling, G.H.T., Mohd Hussain, M.H., Abu Bakar, S.Z. 2023. "Geospatial Mapping of Suicide-Related Tweets and Sentiments among Malaysians during the COVID-19 Pandemic" *Big Data and Cognitive Computing* 7, no. 2: 63. https://doi.org/10.3390/bdcc7020063

Sharma, S., and A. Jain. 2020. "Role of Sentiment Analysis in Social Media Security and Analytics." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10, no. 5: e1366.

Srivastava, S., Y. Arif, and J. Abraham. 2018. Critical Themes in Indian Sociology. SAGE Publications India.

SS, S.F., and S.S. SS. 2015. "The Occurrence of Sexual Harassment Among Sports Practitioners in Bukit Jalil, Kuala Lumpur." Procedia-Social and Behavioral Sciences 211: 917–923.

Syahirah, S., et al. 2021. "Sexual Harassment Deterrent Mechanism to Safeguard the Well-Being of Sports Practitioners." Global Academy of Training and Research (GATR) Enterprise.