International journal of science, engineering, and information technology
Volume 07, Issue 11, July 2023
Journal homepage: https://journal.trunojoyo.ac.id/ijseit

Hand Load Analysis Using Text Mining Based on Letter Frequency of Indonesian Language Theses Documents

Akbar Rizki[1], Barokaturrizkia Ameliani[1], Abdul Aziz Nurussadad[2], Bagus Sartono[1], Itasia Dina Sulvianti[1], Auzi Asfarian[3]

[1]Department of Statistics, IPB University, Bogor, Indonesia
[2]Geospatial Information Agency, Bogor, Indonesia
[3]Department of Computer Science, IPB University, Bogor, Indonesia

ABSTRACT

Scientific documents contain valuable knowledge which can be discovered using text analysis. Meanwhile, the act of typing the text may discover the psychological and health condition. The hand load of the typer is essential information for designing a better keyboard, which is relevant to students' well-being in higher education institutions. This study aims to analyze the hand load of students when typing their theses in the Indonesian language. The documents used in this study were selected using stratified random sampling from undergraduate theses in the Faculty of Mathematics and Natural Science at IPB University between 2019 and 2021, which are available in the institutional repository. Letters from the selected documents were extracted by "pdftools" in R software. We calculate letter frequency in the Indonesian language in theses documents, examine the hand load balance based on letter position on a QWERTY keyboard, identify letter hotspots, and examine hand alternation using circular visualization. The results are that the left hand has the higher load hand, indicated by the most frequent letter appearing in the documents, and the letter W as the hotspot, located on the keyboard's left side. Moreover, hand alternations based on the sequence of Indonesian text identify a significant high alteration of letters from the left to the left side when typing Indonesian documents using the QWERTY keyboard. This result confirmed that the left hand has more load and less time to take a break than the right hand.

Keywords: Hand Load Analysis, Letter frequency, Typing, Scientific Documents

Article History
Received 18 February 23
Received in revised form 02 June 23
Accepted 03 July 23

1. Introduction

Scientific documents contain valuable knowledge which can be discovered using text analysis. Scientific documents, including student theses, are essential to analyze to identify trends, organize contents, or suggest new scientific hypotheses [1]. In order to assist such analysis tasks, the well-understood characteristics of language, which include letter frequency analysis, are essential. Previous studies [2] analyze the English corpus's single letter, and bigram counts, [3] analyze the English and Spanish corpus to get a proper function to predict universal letter frequency in all languages, and [4] analyze Filipino, Indonesian, Malay, and English letters, bigram, and trigram frequency in digital newspapers and found significant frequency differences between languages.

Meanwhile, the act of typing the text may discover the psychological [5] and health [6]–[11] conditions of the typer. By analyzing the text, we can discover the hand load [12] of the typer, which is an essential consideration in designing a better keyboard [13]. This issue is essential in university as it is relevant with students' well-being in higher education institutions [14]. Although the analysis is already commonly done in the English language, there is a lack of analysis for documents written in Indonesian.

This study aims to analyze the hand load of students when typing their theses in the Indonesian language. We calculate letter frequency in the Indonesian language theses documents, examine the hand load balance based on letter position on a QWERTY keyboard, identify letter hotspots, and examine hand alternation using circular visualization.

2. Research Method

The documents used in this study were selected randomly from undergraduate theses in the Faculty of Mathematics and Natural Science at IPB University between 2019 and 2021, which are available in the institutional repository. Using Neyman allocation methods, stratified random sampling was used to select from those available documents. This faculty has eight distinct departments, which were used as a stratum in this study. Under Neyman allocation, the total sample size (n) becomes [15]:

* *Corresponding author.* Phone :+62 811-1144-470
E-mail address:akbar.ritzki@apps.ipb.ac.id

E-ISSN 2548-4214

$$n = \frac{(\sum_{i=1}^{L} N_i\sigma_i)^2}{N^2 D + \sum_{i=1}^{L} N_i {\sigma_i}^2} \quad (1)$$

where: $D = \frac{B^2}{4}$; $B = 2\sqrt{V(\underline{y}_{st})}$; and $V\left(\underline{y}_{st}\right)$ denotes the estimated variance for the population. Further, N denotes the population size, N_i denotes the size of the ith stratum, and σ_i denotes the population variance for the ith stratum.

Subsequently, the number of elements in each stratum (n_i) calculated by using the formula:

$$n_i = n\left(\frac{N_i\sigma_i}{\sum_{i=1}^{L} N_i\sigma_i)}\right) \quad (2)$$

The Population and sample size in each stratum shows in Table 1.

Table 1. The population and sample size in each stratum

Stratum (Department)	Population (N)	Sample (n)
Statistics	150	2
Geophysics and Meteorology	143	4
Biology	236	6
Chemistry	233	8
Mathematics	216	9
Computer Science	252	18
Physics	145	2
Biochemistry	185	5
Total	1.560	54

The study is based on the quantitative approach, covering all of the letters from the cover until the end of the page in each document selected. The data analysis procedure in this study applied the following stages:

1. Extract character from the selected documents by using the package "pdftools"[16] in R software. This tools using "libproppler" which has F1 accuracy around 80% [17].
2. Calculating the total frequency of letters in the selected documents.
3. Examining the balance activity of the right and left hand based on the letters' position on the QWERTY keyboard. The t-student test for paired data which compares the mean of two matched group [18] was used to ensure the results were obtained. The test applied the right and left-hand balance is different in typing the documents as the alternative hypothesis. The letters locations were adapted from [19] on which the left side are A, C, D, E, F, Q, R, S, V, W, X, and Z, while the letters located on the right side are I, J, K, L, M, N, O, and P.
4. Identifying hotspots on the QWERTY keyboard layout features based on the documents' letter frequency using the Getis-Ord Gi* method. This method requires a weight matrix. Queen Contiguity, which used the queen criterion where the eight neighbors of each button in all directions are given the value 1, and all others 0, was used as a weight matrix [20]. Getis-Ord Gi* states the spatial association of high and low letter frequency in the keyboard layout features based on the sample documents [21], [22]. It finds spatial associations among neighboring letters and identifies spatial groups with high and low letter frequency values as hotspots and coldspots, respectively. a Z-score is provided by Getis-Ord Gi* output which is used to identify the specific location [23]. A high Z-score denotes the hotspot, while a low Z-score denotes the coldspot. Suppose $G_i = \frac{\sum_{j=1}^{n} w_{i,j}a_j}{\sum_{j=1}^{n} a_j}$; $E(G_i) = \frac{W_i}{n-1}$; $Var(G_i) = \frac{W_i(n-1-W_i)U_{i2}}{(n-1)^2(n-2)U_{i1}^2}$; $U_{i1} = \frac{\sum_j a_j}{n-1}$; $U_{i2} = \frac{\sum_j a_j^2}{n-1} - U_{i1}^2$ for: $j \neq i$; and $W_i = \sum_j w_{i,j}$. Here a_j denotes the value of the region-j, $w_{i,j}$ denotes the value of spatial weight matrix in *i-th* row and *j-th* column, $j \neq i; j = 1, 2, \dots, n; i = 1, 2, \dots, n$, and n are the number of observed areas. The formula from Z-score for each letter in the keyboard layout (Z_i) can be seen as follow:

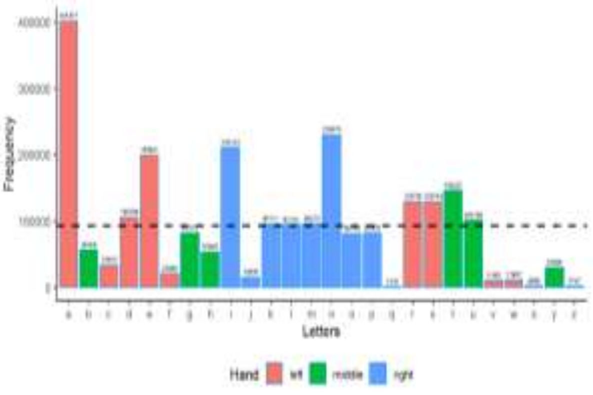
   $$Z_i = \frac{G_i - E(G_i)}{\sqrt{Var(G_i)}} \quad (3)$$

   The probability of Getis-Ord statistics for each letter (G_i) was calculated, compared to the Z-value to determine the hotspot area. A letter is considered as a hotspot area if $G_i \geq Z_{\alpha/2}$ or $(P(Z \geq G_i) \leq \alpha/2)$.
5. Examining hand-alternation by using circular visualization. A circular layout is composed of sectors and tracks [24]. This visualization will works well if there are not to many sectors to visualize[25]. Each letter is featured in the bottom and top circular layout, so there are 26 sectors in the bottom and top layout. Hand alternation when typing the documents determined by the tracks from the bottom to the top. Hand load balance was determined by the movement of the hands while typing a document. Therefore, what was expected was that there were significant movements to the different side (right-left or left-right) compared to the typing in the left or right side only (left-left or right-right).

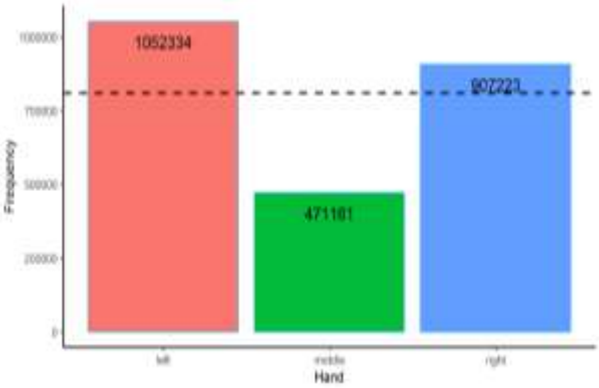
3. Result and Discussion

3.1. Letter Frequency of the documents

Hand load was measured by calculating the letter frequency value of the alphabetic characters based on their location on the QWERTY keyboard layout, on which the right, center, and left. The letters on the right side are accessed by the right hand, and conversely, the letters on the left side are accessed by the left hand. Furthermore, the letters in the middle are assumed to be accessible by the right or left hands. Based on the frequency of the letters shown in Figure 1(a), the most frequent letter in the documents is the letter A, followed by the letters N, I, and E. It is reasonable due to Indonesian and Malay languages were dominated by A and N [26], especially in the suffix. Meanwhile, letter frequency based on their location presented in Figure 1 (b), the frequency of letters appearing is dominated by the letters on the left side. It was followed by the letters on the right, and the letters in the middle have the least frequency in the documents. In other words, the left hand has a higher load than the right hand.



(a)



(b)

Figure 1. (a) Frequency of the letters (b) Letters frequency based on the location in the QWERTY keyboard layout

The paired sample t-test was used to support the results. The data applied in this analysis was the frequency of occurrence of letters located on the right and left only. The alternative hypothesis was there is a different load between the right and left hands. The p-value of the test was 5.04×10^{-17}, which means there is a significant difference between the load of the right and left hands at the 5% significance level.

3.2. Hotspots Analysis

The Z-score of each letter based on the letter frequency in the documents and the location in the keyboard layout is shown in Table 1. The positive Z-score of the Getis-Ord Gi* local spatial autocorrelation could be a hotspot. In contrast, the negative Z-score of the Getis-Ord Gi* local spatial autocorrelation could be a cold spot. It can be seen that the Z-score range from -1.474 to 2.076. Thus, the potential alphabet candidacy as a hotspot can be W, Q, and Z as the alphabetic, which has a high Z-score.

Table 2. Z-score of each alphabet

Alphabet	Z-Score	Alphabet	Z-Score
A	-1.474	N	-0.806
B	-0.011	O	0.640
C	-1.438	P	-0.095
D	-0.204	Q	1.786
E	0.125	R	0.628
F	-0.347	S	0.877
G	-1.237	T	-0.602
H	-0.272	U	-0.370
I	-0.372	V	-1.155
J	1.074	W	2.076
K	0.171	X	-0.701
L	-0.105	Y	0.003
M	0.376	Z	1.652

Figure 2 presents a keyboard map of Getis-Ord Gi* local spatial autocorrelation Z-score. The white part of the keyboard is letter with a negative Z-score. On the contrary, the button of the keyboard with an orange gradation is letter with a positive Z-score. The keyboard part with a dark orange color indicates a high Z-score, and conversely, a light orange color indicates a low Z-score. The Q, W, and Z have a deep red color, with the darkest part on the W. Based on the 5% significance level, the character W is a hotspot. The W character is in the left-hand position. Thus, the keyboard part in the left-hand position has a spatial grouping of alphabet with a high letter frequency.

The hotspot analysis results show a local spatial grouping of the letter with a high frequency in the left-hand position. It impacts the high intensity of typing on the keyboard layout in the left-hand position so that the load on the left hand becomes greater than the load on the right hand when typing. It also indicates the inappropriate use of the QWERTY keyboard layout in preparing Indonesian-language scientific documents, which in this study were theses documents of the students.

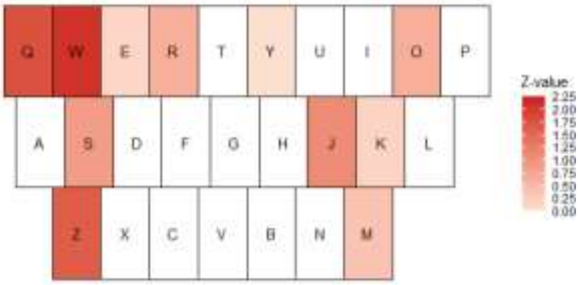
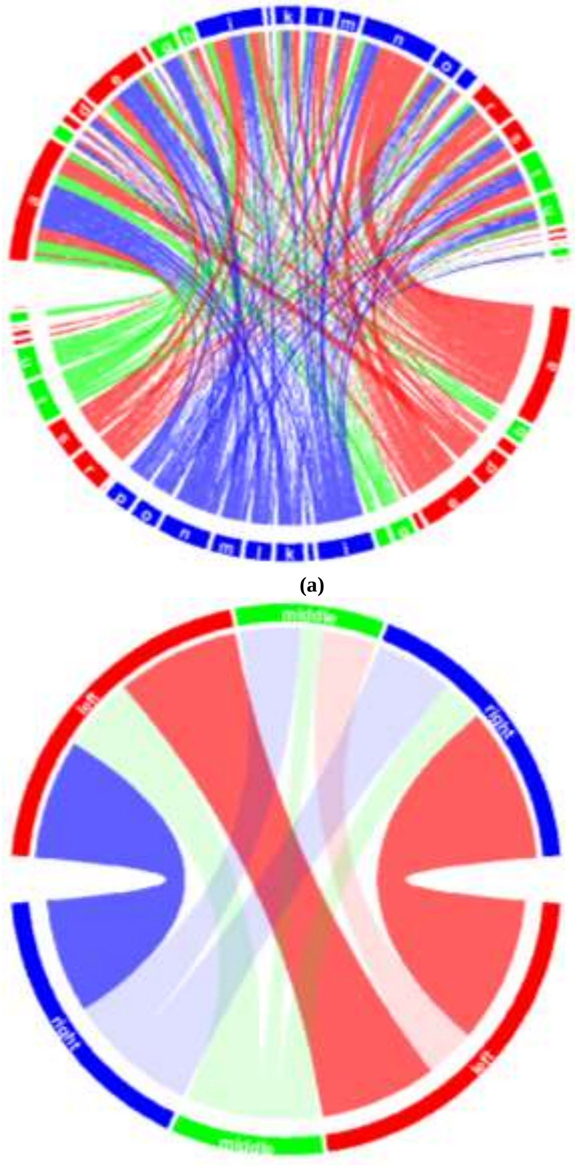


Figure 2. Hotspot distribution in the QWERTY keyboard layout

3.3. Hand alternation

Hand load balancing involves balancing the alterations of hands when typing documents. The movements from the right to left side or left to the right side when typing letters indicate the load for both hands is more balanced than the sequence pattern of letters, which only involves one side, the left or right side. The hand alternations when typing can be identified based on the order of letters in the documents. Figure 3(a) presents a circular graph illustrating the movement of letters, and Figure 3(b) shows the alterations based on the location of the letters on the keyboard layout.

Based on the circular visualization in figure 3(b), the movements from one side to another (right to left/ left to right) dominated when typing the documents. However, looking into more detail, the alterations from left to left side also take a significant proportion. It can be seen through the thick red line connecting the "left" bottom sector to the "left" in the top sector. It supports the previous results, which indicate that the left-hand load was heavier than the right-hand one. Furthermore, it indicated that the left hand has less time to rest than the right hand.



(a)

(b)

Figure 3. Circular visualization based on (a) letters alternation (b) location of the letters in the QWERTY keyboard layout

4. Conclusion

There is an imbalanced hand load when using a QWERTY keyboard to type Indonesian documents. It identified that the left hand has more hand load than the right hand. It can be seen that the most frequent letters are letters located on the left side of the keyboard layout, which is supported by the result of the t-student paired test with the letter A dominating the writing of Indonesian language documents. This result is also confirmed by hotspot analysis results where the letter W, located on the left side of the keyboard layout, was identified as a hotspot. In other words, the letters which are located around the letter W have a high frequency of occurrence. Additionally, the results of hand movements identified from the sequence of letters in the Indonesian text state that there is still a reasonably high movement of letters from the left to the left side, even though the change of letters from one side to another has dominated. It showed that the left hand had less time to take a break than the right hand.

Acknowledgements

The authors thank to Research Institutions and Community Service of IPB University (LPPM-IPB) for funding and supporting this research.

References

[1] S. Spangler *et al.*, “Automated hypothesis generation based on mining scientific literature,” *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 1877–1886, 2014, doi: 10.1145/2623330.2623667.

[2] M. N. Jones and D. J. K. Mewhort, “Case-sensitive letter and bigram frequency counts from large-scale English corpora,” *Behav. Res. Methods, Instruments, Comput.*, vol. 36, no. 3, pp. 388–396, 2004, doi: 10.3758/BF03195586.

[3] W. Li and P. Miramontes, “Fitting ranked english and spanish letter frequency distribution in US and Mexican presidential speeches,” *J. Quant. Linguist.*, vol. 18, no. 4, pp. 337–358, 2011, doi: 10.1080/09296174.2011.608606.

[4] N. Lin, S. Fu, J. Huang, and S. Jiang, “Exploring Letter’s Differences between Partial Indonesian Branch Language and English,” *Proc. 2019 Int. Conf. Asian Lang. Process. IALP 2019*, pp. 84–89, 2019, doi: 10.1109/IALP48816.2019.9037715.

[5] P. Freihaut and A. S. Göritz, “Does Peoples’ Keyboard Typing Reflect Their Stress Level?: An Exploratory Study,” *Zeitschrift fur Psychol. / J. Psychol.*, vol. 229, no. 4, pp. 245–250, 2021, doi: 10.1027/2151-2604/a000468.

[6] A. Pimenta, D. Carneiro, P. Novais, and J. Neves, “Monitoring mental fatigue through the analysis of keyboard and mouse interaction patterns,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8073 LNAI, pp. 222–231, 2013, doi: 10.1007/978-3-642-40846-5_23.

[7] D. Iakovakis *et al.*, “Early Parkinson’s Disease Detection via Touchscreen Typing Analysis using Convolutional Neural Networks,” *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, pp. 3535–3538, 2019, doi: 10.1109/EMBC.2019.8857211.

[8] D. Iakovakis, S. Hadjidimitriou, V. Charisis, S. Bostantzopoulou, Z. Katsarou, and L. J. Hadjileontiadis, “Touchscreen typing-pattern analysis for detecting fine motor skills decline in early-stage Parkinson’s disease,” *Sci. Rep.*, vol. 8, no. 1, pp. 1–13, 2018, doi: 10.1038/s41598-018-25999-0.

[9] K. Nakatsuka *et al.*, “Association between comprehensive workstation and neck and upper-limb pain among office worker,” *J. Occup. Health*, vol. 63, no. 1, pp. 1–7, 2021, doi: 10.1002/1348-9585.12194.

[10] K. Selvakumar and L. Y. Sheng, "The Prevalence and Risk Factors of Wrist Pain among Young Adults in UTAR during Movement Control Order: A Cross-Sectional Study," *Int. J. Heal. Sci. Res.*, vol. 11, no. 11, pp. 191–202, 2021, doi: 10.52403/ijhsr.20211125.

[11] W. Rauf, N. Anwar, and A. Ahmed, "Work Related Musculoskeletal Wrist Pain and Functional Disability in Of □ ce Workers Using Computer Abstract : Introduction : Objective : Results : Conclusions : Keywords : Results :," *Asian J. Allied Heal. Sci.*, vol. 02, no. 04, pp. 36–40, 2017.

[12] G. D. Logan and M. J. C. Crump, "The left hand doesn't know what the right hand is doing: The disruptive effects of attention to the hands in skilled typewriting: Research article," *Psychol. Sci.*, vol. 20, no. 10, pp. 1296–1300, 2009, doi: 10.1111/j.1467-9280.2009.02442.x.

[13] A. H. H. Onsorodi and O. Korhan, "Application of a genetic algorithm to the keyboard layout problem," *PLoS One*, vol. 15, no. 1, pp. 1–11, 2020, doi: 10.1371/journal.pone.0226611.

[14] A. K. Soutter, B. O'Steen, and A. Gilmore, "The student well-being model: A conceptual framework for the development of student well-being indicators," *Int. J. Adolesc. Youth*, vol. 19, no. 4, pp. 496–520, 2014, doi: 10.1080/02673843.2012.754362.

[15] R. L. Scheaffer, W. Mendenhall, L. Ott, and K. Gerow, *Elementary Survey Sampling*, Seventh Ed. Boston M: Brooks/Cole, 2012.

[16] J. Ooms, "pdftools: Text Extraction, Rendering and Converting of PDF Documents." rOpenSci, 2022, [Online]. Available: https://docs.ropensci.org/pdftools/.

[17] M. S. U. Miah *et al.*, "Sentence Boundary Extraction from Scientific Literature of Electric Double Layer Capacitor Domain: Tools and Techniques," *Appl. Sci.*, vol. 12, no. 3, pp. 1–19, 2022, doi: 10.3390/app12031352.

[18] A. Ross and V. L. Willson, "Paired Samples T-Test," in *Basic and Advanced Statistical Tests*, Rotterdam: SensePublishers, 2017, pp. 17–19.

[19] J. Noyes, "The QWERTY keyboard: a review," *Int. J. Man. Mach. Stud.*, vol. 18, no. 3, pp. 265–281, 1983, doi: 10.1016/S0020-7373(83)80010-8.

[20] A. Getis and J. Aldstadt, "Constructing the spatial weights matrix using a local statistic," *Adv. Spat. Sci.*, vol. 61, no. DECEMBER, pp. 147–163, 2010, doi: 10.1007/978-3-642-01976-0_11.

[21] S. Kumar and B. R. Parida, "Hydroponic farming hotspot analysis using the Getis–Ord Gi* statistic and high-resolution satellite data of Majuli Island, India," *Remote Sens. Lett.*, vol. 12, no. 4, pp. 408–418, 2021, doi: 10.1080/2150704X.2021.1895446.

[22] J. M. Sánchez-Martín, J. I. Rengifo-Gallego, and R. Blas-Morato, "Hot Spot Analysis versus Cluster and Outlier Analysis: An enquiry into the grouping of rural accommodation in Extremadura (Spain)," *ISPRS Int. J. Geo-Information*, vol. 8, no. 4, 2019, doi: 10.3390/ijgi8040176.

[23] J. K. Ord and A. Getis, "The Analysis of Spatial Association," *Geogr. Anal.*, vol. 24, no. 3, pp. 189–206, 1992, [Online]. Available: https://doi.org/10.1111/j.1538-4632.1992.tb00261.x.

[24] Z. Gu, L. Gu, R. Eils, M. Schlesner, and B. Brors, "Circlize implements and enhances circular visualization in R," *Bioinformatics*, vol. 30, no. 19, pp. 2811–2812, 2014, doi: 10.1093/bioinformatics/btu393.

[25] W. Qi, G. J. Abel, R. Muttarak, and S. Liu, "Circular visualization of China's internal migration flows 2010–2015," *Environ. Plan. A*, vol. 49, no. 11, pp. 2432–2436, 2017, doi: 10.1177/0308518X17718375.

[26] A. Shah, A. Z. Saidin, I. F. Taha, and A. M. Zeki, "Frequencies determination of characters for Bahasa Melayu: Results of preliminary investigations," *Procedia - Soc. Behav. Sci.*, vol. 27, no. Pacling, pp. 233–240, 2011, doi: 10.1016/j.sbspro.2011.10.603.